Table of Contents

[LANGCHAIN 3](#_Toc206938243)

[UNDERSTANDING LANGCHAIN COMPONENTS USING RAG EXAMPLE 3](#_Toc206938244)

[KEY FEATURES & MODULES 4](#_Toc206938245)

[CONCEPTS 5](#_Toc206938246)

[CHAINING 5](#_Toc206938247)

[COMPONENTS OF LANGCHAIN 6](#_Toc206938248)

[CHAT MODELS 6](#_Toc206938249)

[LANGCHAIN IN ACTION 6](#_Toc206938250)

[PROMPT TEMPLATE 7](#_Toc206938251)

[EXAMPLE 8](#_Toc206938252)

[Partial() 8](#_Toc206938253)

[CHAT PROMPT TEMPLATE 9](#_Toc206938254)

[OUTPUT PARSER 10](#_Toc206938255)

[COMMON OUTPUT PARSERS IN LANGCHAIN 10](#_Toc206938256)

[CREATING SIMPLE GENAI APPLICATION 11](#_Toc206938257)

[CODE 11](#_Toc206938258)

[CHAINS 13](#_Toc206938259)

[MESSAGE TYPES (langchain\_core.messages) 13](#_Toc206938260)

[LANGCHAIN EXPRESSION LANGUAGE (LCEL) – SIMPLE DEFINITION 13](#_Toc206938261)

[EXAMPLE 13](#_Toc206938262)

[SEQUENTIAL PROMPTS 14](#_Toc206938263)

[SIMPLE SEQUENTIAL CHAIN 15](#_Toc206938264)

[REGULAR SEQUENTIAL CHAIN 15](#_Toc206938265)

[CHAT HISTORY 15](#_Toc206938266)

[CHAT PROMPT TEMPLATE 15](#_Toc206938267)

[MAINTAINING HISTORY 16](#_Toc206938268)

[EMBEDDING MODELS 16](#_Toc206938269)

[EMBEDDINGS 16](#_Toc206938270)

[HOW EMBEDDINGS ARE GENERATED 16](#_Toc206938271)

[GENERATING EMBEDDINGS USING LANGCHAIN 17](#_Toc206938272)

[SIMILARITY FINDER 18](#_Toc206938273)

[HOW IT WORKS? 18](#_Toc206938274)

[VECTORS 19](#_Toc206938275)

[WHY ARE VECTORS USEFUL? 19](#_Toc206938276)

[VECTOR DATABASE 20](#_Toc206938277)

[RAG (RETRIEVER AUGUMENTED GENERATION) 24](#_Toc206938278)

[RAG WORKFLOW 24](#_Toc206938279)

[WHY RAG? 25](#_Toc206938280)

[BENEFITS 25](#_Toc206938281)

[RAG ARCHITECTURE 26](#_Toc206938282)

[RAG PROCESS- BUILDING RAG PIPELINE 26](#_Toc206938283)

[STEP 1: DATA INGESTION (USING DOCUMENT LOADERS) 27](#_Toc206938284)

[STEP 2- DATA CHUNKING 30](#_Toc206938285)

[CREATING EMBEDDINGS 35](#_Toc206938286)

[STORING AND QUERYING VECTOR DATA 39](#_Toc206938287)

[DATA RETRIEVAL 42](#_Toc206938288)

[EXAMPLES – RAG PIPELINES 45](#_Toc206938289)

[EXAMPLE – TEXT DOCUMENT DATA SOURCE 45](#_Toc206938290)

[EXAMPLE-2 – PDF DOCUMENT DATASOURCE 46](#_Toc206938291)

[EXAMPLE- 3 – WORD DOCUMENT DATASOURCE 48](#_Toc206938292)

[LANGSMITH 49](#_Toc206938293)

[STEP TO SET UP LANGSMITH 49](#_Toc206938294)

[CREATE API KEY 50](#_Toc206938295)

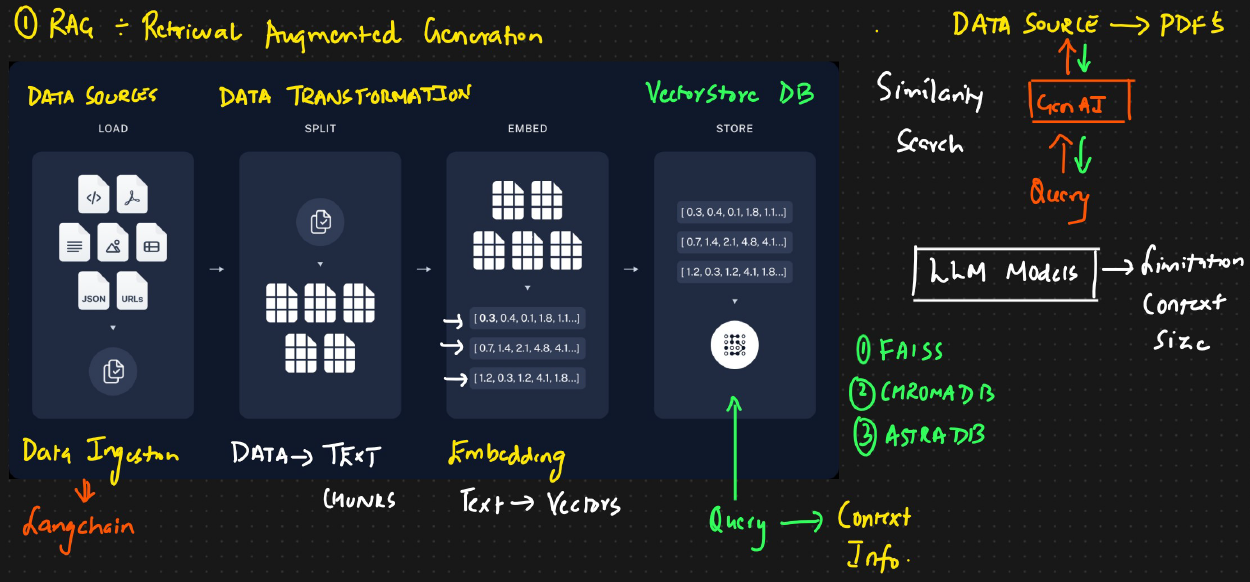
[CODE 50](#_Toc206938296)

[TRACING PROJECT 51](#_Toc206938297)

# LANGCHAIN

* **LangChain** is an open-source framework designed to help developers build applications powered by **large language models (LLMs)** like GPT-4.
* It provides tools to connect LLMs with external data sources, APIs, and memory, enabling more dynamic, context-aware, and interactive applications.

### UNDERSTANDING LANGCHAIN COMPONENTS USING RAG EXAMPLE

To build a RAG based application below are the steps we follow . In below examples we will see how langchain helps in performing those steps

Step 1 – Data Ingestion

1. Loading the data source

Step 2 – Data Chunking/Transformation

1. Splitting the data source into small text chunks

Step 3 –Creating Embeddings (Text Embeddings)

* The data chunks are broken into vectors
* These vectors are used to search using similarity search

Step 4 – Storing Embedding

* The generated embedding are used the stored in a DB called Vector Store DB
* Example – Chroma DB, FAISS, Astra DB

RETRIVING THE DATA

A screenshot of a computer

AI-generated content may be incorrect.

In this retrieval flow when user ask **any question we will see how** Lang Chain handles it behind the scenes.

Step 1: Designing the Prompt

* First, we create a **custom prompt**.  
  Think of it like giving instructions to your AI assistant.
* For example, we might say:  
  *“Hey, act like an AI researcher and answer all user questions.”*
* This prompt tells the language model **how to behave** when responding.

Step 2: Combining Prompt + User Question

* When a user asks a question, we **combine** it with our prompt.
* This combo becomes the input that we send to the system.

Step 3: Searching the Vector Store

* Before answering, the system checks the **Vector Store DB**.
* It looks for **similar content** or context related to the question.

Step 4: Understanding the Retrieval Chain

* The **Retrieval Chain** is the engine that queries the Vector Store.
* It helps us fetch the **right context** from our documents.

|  |
| --- |
| 1. The **Retrieval Chain** is a pipeline that:    * Takes user query.    * Searches a **Vector Store** (a database of embedded documents).    * Finds **relevant context** (like parts of PDFs, docs, or articles).   It’s like asking a librarian to find the best book pages that match your question.   1. The chain pulls out the most relevant chunks of information. |

Step 5: Getting the Final Answer

* Once we have the context from the vector store, we **combine it with the prompt**.
* This full package is sent to the **Language Model (LM)**.
* The LM then gives us a **final response** based on both the context and the prompt.

## KEY FEATURES & MODULES

1. LLM Integration

* Abstracts interaction with various LLMs (e.g., GPT, Claude, Mistral).
* Allows **easy switching between models** using a unified interface(Abstraction).
* Prevents **vendor lock-in** by supporting multiple providers.

2. Prompt Management

* Supports **prompt templates** and **dynamic prompt creation**.
* Helps with **prompt optimization** and **serialization**.

3. Document Loaders

* Load data from sources like PDFs, Notion, emails, etc.
* Converts them into a **standard LangChain document format** for easy processing.

4. Memory

* Maintains **conversation history** for context-aware interactions.

5. Tool & Agent Ecosystem

* Enables LLMs to use tools like:
  + Google Search
  + APIs
  + Databases
* Supports **agentic behavior** (reasoning + tool use).
* Includes components like **agent executors**, **tools**, and **LangGraph**.

## CONCEPTS

### CHAINING

* LangChain is not just another AI tool, it is a comprehensive framework designed to bridge the gap between large language models and real-world applications.
* It works by **chaining** together various components called **links** to provide a flow. Each link in the chain represents a step in the process, from input to output.
* This modular approach allows for flexibility and customization, making LangChain adaptable to a wide range of applications.

A diagram of a software development process

AI-generated content may be incorrect.

Example

Consider a chatbot application.

1. The first link in the chain might be a language model that understands the user’s input.
2. The next link could be a decision-making model that determines the best response.
3. The final link might be another language model that generates the chatbot’s reply.

Each link is independent but interconnected, creating a seamless user experience.

### COMPONENTS OF LANGCHAIN



### CHAT MODELS

## LANGCHAIN IN ACTION

|  |  |
| --- | --- |
| **.env** | |
| **AZURE\_OPENAI\_API\_KEY=**  **AZURE\_OPENAI\_ENDPOINT=https://ai-dojo-open-ai.openai.azure.com/**  **AZURE\_OPENAI\_API\_VERSION=2024-12-01-preview**  **AZURE\_OPENAI\_DEPLOYMENT=gpt-4o-mini** | |
| **from langchain\_openai import AzureChatOpenAI**  **import os**  **from dotenv import load\_dotenv**  **load\_dotenv()**  **# Set your Azure OpenAI credentials**  **subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]**  **endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]**  **api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]**  **deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]**  **# Initialize the Azure OpenAI LLM**  **llm = AzureChatOpenAI(**  **azure\_deployment=deployment,**  **api\_version=api\_version,**  **azure\_endpoint=endpoint,**  **api\_key=subscription\_key,**  **)**  **# Example usage**  **response = llm.invoke("Capital of India!")**  **print(response.content)** |  |

## PROMPT TEMPLATE

* A **Prompt Template** in LangChain is a way to **structure and reuse prompts** for LLMs.
* Instead of writing a full prompt every time, we create a **template** with placeholders, and then fill in those placeholders dynamically based on user input or other data.

A white background with black text

AI-generated content may be incorrect.

Example

Let’s say we want to ask the LLM to summarize a topic.

Step 1: Create a Prompt Template

|  |
| --- |
| **from langchain.prompts import PromptTemplate**  **template = PromptTemplate(**  **input\_variables=["topic"],**  **template="Summarize the following topic in simple terms: {topic}"**  **)** |

Step 2: Format the Prompt with User Input

|  |
| --- |
| **prompt = template.format(topic="Quantum Computing")**  **print(prompt)** |

Output:

**Summarize the following topic in simple terms: Quantum Computing**

Now – We can now send this prompt to an LLM like GPT-4 to get a response.

Real-World Use Case

Imagine a chatbot that:

1. Takes user questions
2. Formats them into a prompt like:\ "Answer the following question clearly and concisely: {question}"
3. Sends it to the LLM
4. This keeps the logic clean and lets us swap out prompts easily without changing the whole app.

### EXAMPLE

|  |
| --- |
| import os  from langchain\_openai import AzureChatOpenAI  from langchain.prompts import PromptTemplate  from dotenv import load\_dotenv  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  **api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]**  **deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]**  **# Initialize the Azure OpenAI LLM**  **llm = AzureChatOpenAI(**  **azure\_deployment=deployment,**  **api\_version=api\_version,**  **azure\_endpoint=endpoint,**  **api\_key=subscription\_key,**  **)**  **prompt\_template = PromptTemplate(**  **input\_variables=["country", "no\_of\_paras", "language"],**  **template="""You are an expert in traditional cuisines.**  **You provide information about a specific dish from a specific country.**  **Avoid giving information about fictional places. If the country is fictional**  **or non-existent answer: I don't know.**  **Answer the question: What is the traditional cuisine of {country}?**  **Answer in {no\_of\_paras} short paras in {language}**  **""",**  **)**  **country = input("Enter the country:")**  **no\_of\_paras = int(input("Enter the number of paras"))**  **language = input("Enter the language:")**  **question = prompt\_template.format(**  **country=country, no\_of\_paras=no\_of\_paras, language=language**  **)**  **# Example usage**  **response = llm.invoke(question)**  **print(response.content)** |

### Partial()

In **LangChain**, the partial() method of a PromptTemplate is used to **pre-fill some variables** in the template, allowing you to reuse the prompt with fewer arguments later.

Example

* Imagine we have a prompt with multiple variables, but some of them stay constant across many uses. Instead of passing all variables every time, you can "lock in" some values using partial().

|  |  |
| --- | --- |
| from langchain.prompts import PromptTemplate  # Define a prompt with two variables  prompt = PromptTemplate.from\_template("Translate the following {language} sentence: {sentence}")  # Partially fill in the language  partial\_prompt **= prompt.partial(language="French")**  # Now you only need to provide 'sentence'  final\_prompt = partial\_prompt.format(sentence="Bonjour, comment ça va?")  print(final\_prompt) | **Output:**  Translate the following French sentence: Bonjour, comment ça va? |

## CHAT PROMPT TEMPLATE

|  |
| --- |
| from langchain\_core.prompts import ChatPromptTemplate  template = ChatPromptTemplate([  **("system", "You are a helpful AI assistant."), 🡨 Tuples**  **("human", "{user\_input}")**  ])  prompt\_value = template.invoke({"user\_input": "What is the weather today?"}) |

The main purpose of ChatPromptTemplate is to:

* **Structure conversations**: We can define how the AI should behave (e.g., “You are a helpful assistant”) and how it should respond to user inputs.
* **Reuse and customize prompts**: Templates can be reused across different applications and customized with variables like {user\_input} or {name}.
* **Support multi-turn interactions**: It enables chaining multiple messages (system, human, AI) to simulate realistic dialogue flows

Example

|  |
| --- |
| from langchain\_openai import AzureChatOpenAI  **from langchain\_core.prompts import ChatPromptTemplate**  import os  from dotenv import load\_dotenv  from langchain\_core.output\_parsers import StrOutputParser  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  os.environ["LANGSMITH\_TRACING"] = os.getenv("LANGSMITH\_TRACING")  os.environ["LANGSMITH\_API\_KEY"] = os.getenv("LANGSMITH\_API\_KEY")  os.environ["LANGSMITH\_PROJECT"] = os.getenv("LANGSMITH\_PROJECT")  # Initialize the Azure OpenAI LLM  llm = AzureChatOpenAI(      azure\_deployment=deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,      temperature=0)  prompt = ChatPromptTemplate.from\_messages(      [          (              "system", "You are an expert AI Engineer. Provide me answers based on my questions."),          ("user", "{input}"),      ]  )  chain = prompt | llm | StrOutputParser()  res = chain.invoke({"input": "Explain the LangChain in detail."})  print(res) |

## OUTPUT PARSER

* An **output parser** is a component responsible for **interpreting and structuring the raw output** from a language model (like GPT) into a more usable format.
* This is especially useful when we are expecting structured data (like JSON, lists, or specific fields) from a model that typically returns plain text.

|  |
| --- |
| Why It's Important?  Language models like GPT often return **free-form text**, which can be inconsistent or hard to process programmatically. Output parsers help in :   1. Extracting specific fields or values 2. Enforceing a schema (like JSON or Pydantic models) 3. Handle errors or malformed outputs 4. Make chaining between components more robust |
| Why Use Output Parsers?   * Language models often return unstructured text. If you're building a pipeline or application that needs structured responses (e.g., for chaining tasks, storing results, or triggering actions), output parsers help enforce consistency and reliability. |
| What It Does?  An output parser typically performs these steps:   * **Receives the raw output** from the model (usually a string). * **Validates and parses** it into a structured format (like a dictionary, list, or custom object). * **Returns** the parsed result for further use in your LangChain pipeline. |

### COMMON OUTPUT PARSERS IN LANGCHAIN

|  |  |
| --- | --- |
| Parser | Description |
| **StrOutputParser** | * Returns the raw string output. * Useful when we just want the plain text. |
| **CommaSeparatedListOutputParser** | * Parses output into a list of strings separated by commas. * Example: "apple, banana, cherry" → ["apple", "banana", "cherry"] |
| **JsonOutputParser** | * Parses output that is expected to be in JSON format. * Ensures the output is valid JSON and converts it to a Python dictionary. |
| **PydanticOutputParser** | * Parses output into a Pydantic model. * Great for enforcing schema and validation. |
| **StructuredOutputParser** | * Uses a schema (like JSON) to parse structured outputs. * Often used with function calling or OpenAI tools. |
| **OutputFixingParser** | * Wraps another parser and tries to fix malformed outputs using the model itself. * Useful when the model output is close to correct but needs minor adjustments. |

EXAMPLE

|  |  |
| --- | --- |
| from langchain\_openai import AzureChatOpenAI  **from langchain\_core.prompts import ChatPromptTemplate**  import os  from dotenv import load\_dotenv  from langchain\_core.output\_parsers import StrOutputParser  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  os.environ["LANGSMITH\_TRACING"] = os.getenv("LANGSMITH\_TRACING")  os.environ["LANGSMITH\_API\_KEY"] = os.getenv("LANGSMITH\_API\_KEY")  os.environ["LANGSMITH\_PROJECT"] = os.getenv("LANGSMITH\_PROJECT")  # Initialize the Azure OpenAI LLM  llm = AzureChatOpenAI(      azure\_deployment=deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,      temperature=0)  prompt = ChatPromptTemplate.from\_messages(      [          ( "system", "You are an expert AI Engineer. Provide me answers based on my questions."),      ("user", "{input}"),      ]  )  chain = prompt | llm | StrOutputParser()  res = chain.invoke({"input": "Explain the LangChain in detail."})  print(res) |  |

## CREATING SIMPLE GENAI APPLICATION

Application Objective

1. Scraping content from a website (e.g. <https://docs.smith.langchain.com/administration/tutorials/manage_spend>
   1. We need to beautifulsoup4 library to scrap the webpage
   2. INSTALL the Beauitifulsoup Library: **pip install beautifulsoup4**
2. Chunking the scraped text
3. Converting chunks into vector embeddings
4. Using a Language Model (LM) with prompt engineering to query or analyze the content

### CODE

* The details of the below is explained in RAG section

|  |
| --- |
| from langchain\_openai import AzureChatOpenAI  from langchain\_core.prompts import ChatPromptTemplate  import os  from dotenv import load\_dotenv  from langchain\_core.output\_parsers import StrOutputParser  from langchain\_community.document\_loaders import WebBaseLoader  from langchain\_text\_splitters import RecursiveCharacterTextSplitter  from langchain\_openai import AzureOpenAIEmbeddings  from langchain\_community.vectorstores import FAISS  from langchain.chains.combine\_documents import create\_stuff\_documents\_chain  from langchain.chains import create\_retrieval\_chain  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  os.environ["LANGSMITH\_TRACING"] = os.getenv("LANGSMITH\_TRACING")  os.environ["LANGSMITH\_API\_KEY"] = os.getenv("LANGSMITH\_API\_KEY")  os.environ["LANGSMITH\_PROJECT"] = os.getenv("LANGSMITH\_PROJECT")  # Initialize the Azure OpenAI LLM  llm = AzureChatOpenAI(      azure\_deployment=deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,      temperature=0)  # Initialize the Azure OpenAI Embeddings  embeddings\_client = AzureOpenAIEmbeddings(      azure\_deployment=embedding\_deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,  )  loader = WebBaseLoader(      web\_paths=(          "https://docs.smith.langchain.com/administration/tutorials/manage\_spend",      )  )  docs = loader.load()  text\_splitter = RecursiveCharacterTextSplitter(chunk\_size=1000, chunk\_overlap=200)  text\_chunks = text\_splitter.split\_documents(docs)  vector\_db = FAISS.from\_documents(text\_chunks, embeddings\_client)  query = "How to set a good total traces limit"  prompt = ChatPromptTemplate.from\_template(      """Answer the following question based only on the provided context:  <context>  {context}  </context>  Question: {input}"""  )  document\_chain = create\_stuff\_documents\_chain(llm, prompt=prompt)  retriever = vector\_db.as\_retriever()  retrieval\_chain = create\_retrieval\_chain(retriever, document\_chain)  response = retrieval\_chain.invoke({"input": query})  print(f"Response: {response['answer']}") |

## CHAINS

### MESSAGE TYPES (langchain\_core.messages)

### LANGCHAIN EXPRESSION LANGUAGE (LCEL) – SIMPLE DEFINITION

* **LCEL** is a way to build AI workflows in LangChain using a clean and modular style.
* It lets us connect components like prompts, models, and output parsers using a simple syntax, making the code easier to read and maintain.

### EXAMPLE

|  |  |
| --- | --- |
| **import os**  **from langchain\_openai import AzureChatOpenAI**  **from langchain.prompts import PromptTemplate**  **from dotenv import load\_dotenv**  **from langchain\_core.output\_parsers import StrOutputParser**  **load\_dotenv()**  **# Set your Azure OpenAI credentials**  **subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]**  **endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]**  **api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]**  **deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]**  **# Initialize the Azure OpenAI LLM**  **llm = AzureChatOpenAI(**  **azure\_deployment=deployment,**  **api\_version=api\_version,**  **azure\_endpoint=endpoint,**  **api\_key=subscription\_key,**  **)**  **prompt\_template = PromptTemplate(**  **input\_variables=["language", "text"],**  **template="Translate the following text to {language}: {text}",**  **)**  **text = input("Enter the text to translate: ")**  **language = input("Enter the target language: ")**  **parser = StrOutputParser()**  **# Example usage**  **chain = prompt\_template | llm | parser**  **response = chain.invoke({"text": text, "language": language})**  **print(response)** | **OUTPUT**     * **Note : The chain.invoke() takes dictionary of prompt parameters** |

Explanation

* prompt formats the input.
* llm generates the response.
* parser extracts the final output.
* The | operator connects them in sequence.

### SEQUENTIAL PROMPTS

* A **sequential prompt** in LangChain refers to a **workflow where multiple prompts are executed in a specific order**, **with the output of one prompt serving as the input to the next**.
* This is implemented using **Sequential Chains**, which are ideal for tasks that require **step-by-step reasoning or multi-stage processing**.

Example

|  |
| --- |
| **import os**  **from langchain\_openai import AzureChatOpenAI**  **from langchain.prompts import PromptTemplate**  **from dotenv import load\_dotenv**  **from langchain\_core.output\_parsers import StrOutputParser**  **load\_dotenv()**  **# Set your Azure OpenAI credentials**  **subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]**  **endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]**  **api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]**  **deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]**  **# Initialize the Azure OpenAI LLM**  **llm = AzureChatOpenAI(**  **azure\_deployment=deployment,**  **api\_version=api\_version,**  **azure\_endpoint=endpoint,**  **api\_key=subscription\_key,**  **)**  **topic\_prompt = PromptTemplate(**  **input\_variables=["topic"], template="Create a speech on following topic: {topic}"**  **)**  **speech\_prompt = PromptTemplate(**  **input\_variables=["no\_of\_words"],**  **template="keep the speech under {no\_of\_words} words. "**  **"Make it engaging and informative. ",**  **)**  **first\_chain = topic\_prompt | llm | StrOutputParser()**  **second\_chain = speech\_prompt | llm | StrOutputParser()**  **final\_chain = first\_chain | second\_chain**  **response = final\_chain.invoke({"topic": "Artificial Intelligence", "no\_of\_words": 1000})**  **print(response)** |

### SIMPLE SEQUENTIAL CHAIN

1. **Linear flow**: Each step takes the output of the previous step as its input.
2. **No variable mapping**: You don’t specify input/output keys.
3. **Best for quick prototyping** with simple chains.

### REGULAR SEQUENTIAL CHAIN

## CHAT HISTORY

### CHAT PROMPT TEMPLATE

* ChatPromptTemplate is a specialized prompt class in LangChain designed for **chat-based models** like OpenAI's GPT-4 or Anthropic's Claude.
* It helps us structure prompts in a way that mimics a **conversation**, using roles like system, user, and assistant.

Example

|  |
| --- |
| **from langchain\_openai import AzureChatOpenAI**  **from langchain\_core.prompts import ChatPromptTemplate**  **import os**  **from dotenv import load\_dotenv**  **load\_dotenv()**  **# Set your Azure OpenAI credentials**  **subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]**  **endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]**  **api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]**  **deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]**  **# Initialize the Azure OpenAI LLM**  **llm = AzureChatOpenAI(**  **azure\_deployment=deployment,**  **api\_version=api\_version,**  **azure\_endpoint=endpoint,**  **api\_key=subscription\_key,**  **)**  **prompt = ChatPromptTemplate.from\_messages(**  **[("system", "You are a Sales Man."),**  **("human", "{input}")]**  **)**  **chain = prompt | llm**  **response = chain.invoke({"input": "Sell me the pen."})**  **print(response.content)** |

### MAINTAINING HISTORY

## EMBEDDING MODELS

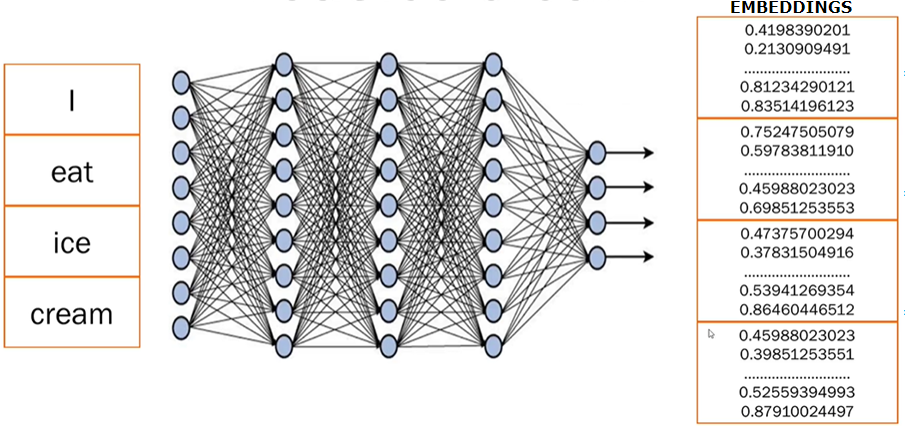
### EMBEDDINGS

* Embeddings are numerical representations of text(called vectors)—a way to convert words into numbers so machines can understand and process them.
* These vectors **capture meaning and relationships** between words or sentences.

How Embeddings Work?

* Embeddings encode:
  + Frequency of word usage in different contexts.
  + Relationships between words and sentences.
  + These vectors allow machines to compare meanings effectively.

### HOW EMBEDDINGS ARE GENERATED



1. **STEP 1:** **BREAKING TEXT INTO TOKENS**

* Sentences are broken into smaller pieces or tokens (e.g., splitting "I eat ice cream" into 4 tokens—"I", "eat", "ice", "cream").

2. **STEP 2: NEURAL NETWORK PROCESSING**:

* Trained **transformer models** analyze the text, generate embeddings, and capture meaning, context, and relations between tokens.

1. **STEP 3: NUMERICAL EMBEDDINGS:**

* Each token is converted into numerical data (random numbers).
* These numbers represent embeddings, storing all learned information about the word or sentence.
* Only the transformer model understands what these embeddings mean based on its training.

Understanding Semantic Relationships

|  |  |
| --- | --- |
|  | Considering a simple example   * Here is an example of three words **happy, joyful, and glad**, which are usually used in a similar context. * If we assign numeric codes (like 1, 2, 3) to these words, any algorithm or machine, by looking at these numeric codes, can easily figure out the relationship between these words. * 1 and 2 are much closer than 1 and 3, so it will know that happy and joyful are much more frequently used, are closely related than happy and glad. * *But in real time, these numeric codes or embeddings are much more complex.* |

Example of Sentence Embeddings

Sentences like:

***"The sun is shining brightly in clear blue sky."***

***"It's a beautiful day with clear sunny sky."***

Have similar meanings, and their embeddings will be close together in vector space.

Applications of Embeddings

* 1. Document similarity: Find texts with similar meanings.
  2. Search engines: Enable semantic search beyond keyword matching.
  3. Recommendation systems: Suggest similar items based on meaning.
  4. Language translation: Match meaning across languages.

### GENERATING EMBEDDINGS USING LANGCHAIN

|  |
| --- |
| **from langchain\_openai import AzureOpenAIEmbeddings**  **import os**  **from dotenv import load\_dotenv**  **load\_dotenv()**  **# Set your Azure OpenAI credentials**  **subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]**  **endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]**  **api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]**  **deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]**  **# Initialize the Azure OpenAI Embeddings**  **embeddings = AzureOpenAIEmbeddings(**  **azure\_deployment=deployment,**  **api\_version=api\_version,**  **azure\_endpoint=endpoint,**  **api\_key=subscription\_key,**  **)**  **# Example usage**  **try:**  **text = input("Enter the text to embed: ")**  **response = embeddings.embed\_query(text)**  **print(f"Embedding vector length: {len(response)}")**  **print(f"First 5 dimensions: {response[:5]}")**  **except Exception as e:**  **print(f"Error generating embeddings: {e}")** |
| **.env**  **AZURE\_OPENAI\_API\_KEY=2H8z7i6cZxgDx8X7WksANiaTZQveZBoqYOsHBjzsXB1ObMv7MyqKJQQJ99BHACYeBjFXJ3w3AAABACOGs6Zg**  **AZURE\_OPENAI\_ENDPOINT=https://ai-dojo-open-ai.openai.azure.com/**  **AZURE\_OPENAI\_API\_VERSION=2024-12-01-preview**  **AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING=text-embedding-ada-002** |
| **OUTPUT** |

### SIMILARITY FINDER

* A similarity finder is a tool that compares two pieces of text—words, phrases, or sentences—and tells us how similar they are in meaning.
* It uses embeddings, which are numerical representations of text, to do this.

### HOW IT WORKS?

1. Convert Text to Embeddings

* + Each input text is passed through an embedding model (like OpenAI’s text-embedding-3-large).
  + The model returns a vector (a list of numbers) that captures the semantic meaning of the text.

2. Compare the Vectors

* The two vectors are compared using cosine similarity, which measures the angle between them.
* This is done using a mathematical function like np.dot() from the NumPy library.

3. Interpret the Similarity Score

The result is a number between 0 and 1:

* 1 means the texts are very similar.
* 0 means they are completely different.

Example

|  |  |  |
| --- | --- | --- |
| Text 1 | Text 2 | Similarity Score |
| "dog" | "cat" | 0.76 (76%) |
| "cat" | "kitten" | 0.88 (88%) |
| "You are great" | "You are bad" | 0.87 (87%) |
| "You are awesome" | "You are good" | 0.91 (91%) |

Why It’s Useful

* Semantic Search: Find documents or answers that match a query in meaning.
* Recommendation Systems: Suggest similar items based on user input.
* Plagiarism Detection: Identify texts that are similar meanings.
* Chatbots & Assistants: Retrieve relevant responses or context.

#### EXAMPLE

|  |  |
| --- | --- |
| INSTALL NUMPY | **pip install numpy** |

* Similarity finder that will figure out how similar given two words or sentences are using their embedding values.

|  |  |  |  |
| --- | --- | --- | --- |
| from langchain\_openai import AzureOpenAIEmbeddings  import os  from dotenv import load\_dotenv  import numpy as np  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  # Initialize the Azure OpenAI Embeddings  embeddings = AzureOpenAIEmbeddings(  azure\_deployment=deployment,  api\_version=api\_version,  azure\_endpoint=endpoint,  api\_key=subscription\_key,  )  # Example usage  try:  **text1 = input("Enter the text1: ")**  **text2 = input("Enter the text2: ")**  **# Get the embedding of the text**  **response1 = embeddings.embed\_query(text1)**  **response2 = embeddings.embed\_query(text2)**  **# Calculate the cosine similarity**  **similarity = np.dot(response1, response2)**  **print(similarity)**  except Exception as e:  print(f"Error generating embeddings: {e}") | | | |
| **OUTPUT: The output will be the similarity score in between 0 and 1. The higher the score, the closer those texts are.** | | | |
|  |  |  |  |

## VECTORS

A **vector** is a list of numbers that represents something—like a word, an image, or even a sentence—in a way that a computer can understand and work with.

|  |  |
| --- | --- |
| Simple Analogy | |
| Imagine we want to describe a fruit (say, an **apple**) using numbers:   * Sweetness: 8 * Crunchiness: 7 * Juiciness: 6   We could represent the apple as a vector: [8, 7, 6]  Now, a **banana** might be:[9, 3, 7]  These vectors help a computer compare fruits based on their features. | A diagram of a diagram  AI-generated content may be incorrect. |

In Language Models

When we talk about **words**, we turn them into vectors using **embeddings**. For example:

* “cat” → [0.12, -0.45, 0.88, ..., 0.03]
* “dog” → [0.10, -0.40, 0.85, ..., 0.05]

These vectors are **high-dimensional** (often 300 to 1,000+ numbers long) and capture the **meaning** of the word based on how it’s used in language.

### WHY ARE VECTORS USEFUL?

They allow computers to:

* **Compare** things (e.g., how similar two words or images are)
* **Search** by meaning (semantic search)
* **Cluster** similar items together
* **Feed data into machine learning models**

### VECTOR DATABASE

A diagram of a data processing process

AI-generated content may be incorrect.

* A **vector database** is a special kind of database designed to store and search **vectors (***which are just lists of numbers that represent things like text, images, or audio in a way that computers can understand*.)
* Primarily used for storing embeddings that represent complex data like images, text and audio in a form that machine can understand and process

Why Vectors?

* When you use **embeddings** (like we discussed earlier), we turn data (like the word *“cat”*) into a vector, such as:

**[0.12, -0.45, 0.88, ..., 0.03]**

* These vectors capture **meaning** and **context**. But once we have millions of them, we need a smart way to **store** and **search for** them efficiently. That’s where vector databases come in.

What Does a Vector Database Do?

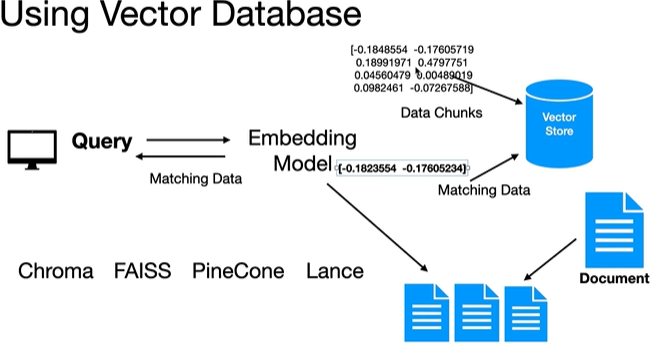
It helps to:

* **Store** millions or billions of embeddings
* **Search** for the most similar vectors (e.g., “find texts similar to this one”)
* **Rank** results by similarity (using distance metrics like cosine similarity)

Real-World Example

Let’s say we run a **document search engine**:

1. We convert all the documents into vectors using an LLM.
2. We store those vectors in a vector database (like **Pinecone**, **Weaviate**, **FAISS**, or **Milvus**).
3. When a user asks a question, we:
   * Convert the question into a vector
   * Search the database for the **most similar document vectors**
   * Return the most relevant documents



* A vector store allows us to store and retrieve data using high dimensional vectors.
* This is very useful to build similarity search applications.

#### STEPS

* Step 1: Chunking
  + The process to store and retrieve data in a vector store involves taking that data and breaking it down into small chunks.
* Step 2: Calculate Embeddings
  + Once we have the chunks, we'll calculate the embeddings for each of these chunks using an embedding large language model.
* Step 3: Store in Vector Database
  + Once we have the embedding, **we'll use this embedding as metadata will take the chunk of the data and then send it to the vector store.**
  + The vector store will create an index internally using these embeddings, and it will store the data chunk under that index.(So the vector store is aware of the high dimensional vector for that data chunk.)

When Queried By User

* + When the user later comes up with a query or a search will take that search query, calculate the embedding for it, and will hand over that embedding to the vector store.
  + The vector store will use that embedding, find all the similar embeddings, and return back the data chunks for those embeddings, so all the matching data will be returned back by the vector store, which we can then send back to the user.

#### TYPES OF VECTOR STORE

* **FAISS**
* **Chroma**
* **Pinecone**
* **Weaviate**
* **Milvus**
* **Qdrant**
* **Redis**
* **ElasticSearch**

#### EXAMPLE - LANGCHAIN

**DATA- joblist.txt**

|  |
| --- |
| 1. **Software Engineer at TechCorp - Responsibilities include developing and maintaining software applications, collaborating with cross-functional teams, and ensuring code quality. Requires proficiency in Java, Python, and SQL.**  **2. Data Scientist at DataMinds - Duties involve analyzing large datasets, building predictive models, and presenting insights to stakeholders. Requires expertise in Python, R, and machine learning.**  **3. Digital Marketing Specialist at MarketGurus - Role includes creating and managing online marketing campaigns, analyzing web traffic, and optimizing SEO. Requires experience with Google Analytics, SEM, and content creation.**  **4. Project Manager at BuildIt - Responsibilities include overseeing construction projects, managing budgets, and coordinating with contractors. Requires strong leadership skills and knowledge of project management software.**  **5. Graphic Designer at CreativeWorks - Role involves designing marketing materials, collaborating with the creative team, and adhering to brand guidelines. Requires proficiency in Adobe Creative Suite and a strong portfolio.**  **6. Financial Analyst at FinExperts - Duties include analyzing financial data, preparing reports, and advising on investment decisions. Requires strong analytical skills and experience with financial modeling.**  **7. Human Resources Manager at PeopleFirst - Responsibilities include recruiting, onboarding, and managing employee relations. Requires excellent communication skills and knowledge of HR software.**  **8. Cybersecurity Specialist at SecureNet - Role involves protecting the company's IT infrastructure, monitoring for security breaches, and implementing security protocols. Requires expertise in network security and experience with security tools.**  **9. Sales Manager at RetailStars - Duties include managing the sales team, developing sales strategies, and achieving sales targets. Requires strong leadership skills and experience in retail sales.**  **10. Content Writer at WordSmiths - Responsibilities include creating engaging content for blogs, social media, and websites. Requires excellent writing skills and a creative mindset.** |

In this example

* We will use job\_listings.txt file which has several job positions like software engineer, data scientist, digital marketing project, manager, graphic designer, cyber security specialist, content writer and more.
* We will use this data(**job\_listing.txt**)
  1. Break in into chunks
  2. Calculate the embeddings
  3. Load it up into the vector store

|  |  |
| --- | --- |
| **INSTALL LANGCHAIN CHROMA(Vector Store)** | **pip install langchain\_chroma** |

|  |
| --- |
| from langchain\_openai import AzureOpenAIEmbeddings  **from langchain\_community.document\_loaders import TextLoader**  **from langchain\_text\_splitters import RecursiveCharacterTextSplitter**  **from langchain\_chroma import Chroma**  import os  from dotenv import load\_dotenv  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  # Initialize the Azure OpenAI Embeddings  llm = AzureOpenAIEmbeddings(  azure\_deployment=deployment,  api\_version=api\_version,  azure\_endpoint=endpoint,  api\_key=subscription\_key,  )  **# Load the job listings document from a text file**  **document = TextLoader("job\_listings.txt").load()**  **# Create a text splitter to break the document into smaller chunks**  **text\_spiltter = RecursiveCharacterTextSplitter(chunk\_size=200, chunk\_overlap=10)**  **# Split the loaded document into manageable chunks for processing**  **chunks = text\_spiltter.split\_documents(document)**  **# Create a Chroma vector database from the document chunks using embeddings**  **vector\_db = Chroma.from\_documents(chunks, llm)**  **# Get user input for the search query**  **query = input("Enter the query: ")**  **# Generate embedding vector for the user's query**  **embedding\_of\_query = llm.embed\_query(query)**  **# Search the vector database for documents similar to the query embedding**  **docs = vector\_db.similarity\_search\_by\_vector(embedding\_of\_query)**  **# Iterate through the retrieved similar documents**  **for doc in docs:**  **# Print the content of each similar document**  **print(doc.page\_content)** |
| **OUTPUT** |
|  |

#### QUERY THE VECTOR DB USING RETRIEVER

* Another easy way to query the vector database is to use a retriever instance.

**retriever = vector\_db.as\_retriever()**

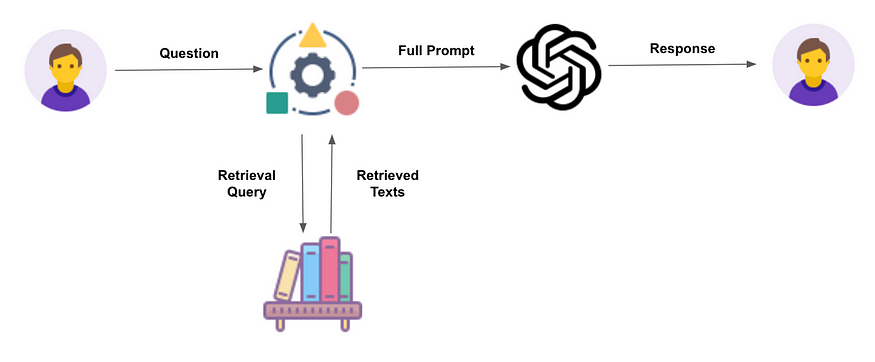
**The retriever instance is capable of taking plain text, calculate the embedding for it, then query the vector database with that embedded text and retrieve the data as well.**

|  |
| --- |
| from langchain\_openai import AzureOpenAIEmbeddings  **from langchain\_community.document\_loaders import TextLoader**  **from langchain\_text\_splitters import RecursiveCharacterTextSplitter**  **from langchain\_chroma import Chroma**  import os  from dotenv import load\_dotenv  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  # Initialize the Azure OpenAI Embeddings  llm = AzureOpenAIEmbeddings(  azure\_deployment=deployment,  api\_version=api\_version,  azure\_endpoint=endpoint,  api\_key=subscription\_key,  )  **# Load the job listings document from a text file**  **document = TextLoader("job\_listings.txt").load()**  **# Create a text splitter to break the document into smaller chunks**  **text\_spiltter = RecursiveCharacterTextSplitter(chunk\_size=200, chunk\_overlap=10)**  **# Split the loaded document into manageable chunks for processing**  **chunks = text\_spiltter.split\_documents(document)**  **# Create a Chroma vector database from the document chunks using embeddings**  **vector\_db = Chroma.from\_documents(chunks, llm)**  **# Convert the vector database to a retriever interface for easier querying**  **retriever = vector\_db.as\_retriever()**  **# Get user input for the search query**  **query = input("Enter the query: ")**  **# Use the retriever to find relevant documents based on the query**  **docs = retriever.invoke(query)**  **# Iterate through the retrieved relevant documents**  **for doc in docs:**  **# Print the content of each relevant document**  **print(doc.page\_content)** |

# RAG (RETRIEVER AUGUMENTED GENERATION)

A white background with blue text

AI-generated content may be incorrect.



* **Retriever**: Fetches relevant information from a knowledge base.
* **Augmented**: Adds value by combining retrieved data with generative capabilities.
* **Generation**: Produces a response using a language model (e.g., ChatGPT, Gemini).

## RAG WORKFLOW

|  |
| --- |
| Step 1: Ask a Question –   * When we ask a question. The question goes into a knowledge base (like a smart library). * The knowledge source can be Databases, Articles or Websites * This knowledge base is often a vector database, which stores information in a format that helps find similar meanings.   Step 2: It Finds Relevant Text   * The system retrieves the most relevant documents or pieces of information that match the question. In the vector Database. This is called the retrieval step.   Step 3: It Builds a Full Input   * The retrieved information is combined with the original question. * This combination becomes a full prompt or input for the next step.   Step 4: AI Generates the Answer   * **Full prompt is sent to a Language Model (like ChatGPT or Gemini).** * **Note**: Retrieved data + original query = **prompt** for the **Language Model (LM)**. * The model reads both questions and the retrieved info and then generates a smart answer. * The response is based on internal data (from the knowledge base) and the intelligence of the AI model. |

## WHY RAG?

**RAG is important due to the following limitations of LLMs**

1. KNOWLEDGE CUT-OFF DATE
   1. The LLM model will have information up to their training cut-off date and lack of information beyond that point.
   2. Hence LLMs **can’t access real-time updates** or dynamic data, and they may miss recent changes in the organization or product.
2. LACK OF ACCESS OF ENTERPRISE DATA
   1. They lack access of enterprise specific data unless they are fine-tuned for customized for that enterprise

|  |
| --- |
| * Let’s say in banks or an enterprise company, the data will be sitting privately within companies private infrastructure or within companies’ network. * In this use-case - We don't want ChatGPT or any other LLMs to have information about private data. So that is where vector databases come into picture where the questions retrievals will happen from the company’s internal source. * Then – we are just using the capabilities of LLM to prepare a nice prompt to give a proper reply. |

Enterprise Use Cases

* **Customer support**: Pulls product details from internal DBs.
* **Educational tools**: Provides precise, sourced answers.
* **Banking/Enterprise**: Keeps sensitive data internal while leveraging LMs.

## BENEFITS

Access to Private/Internal Data

* RAG allows LLMs to use your organization’s internal documents (e.g., product manuals, policies, reports) that are not part of public training data.
* This makes responses more relevant and accurate for enterprise use.

Reduces Hallucinations

* LLMs sometimes make up answers when they don’t have enough information.
* RAG reduces this by grounding responses in real, retrieved documents.

Improves Contextual Accuracy

* RAG retrieves context-specific information before generating a response.
* This ensures the answer is tailored to the user’s query and environment.

Keeps Data Secure

* Sensitive data stays within private infrastructure.
* The LLM only sees the retrieved content, not the entire database—helping with data privacy and compliance.

Dynamic and Up-to-Date Responses

* Instead of relying on static training data, RAG can pull real-time or recently updated documents.
* This makes it ideal for fast-changing domains like tech support or policy updates.

## RAG ARCHITECTURE

A diagram of a computer system

AI-generated content may be incorrect.Step 1: Prepare Your Data

* Collect documents, images, videos, etc.
* Send them to an **Embedder**, which converts them into a format (vectors) that computers can search easily.

Step 2: Store the Data

* The Embedder sends these vectors to a **Vector Storage and Retrieval Engine** (like a smart library).

Step 3: User Asks a Question

* A person types a question into a **chat interface**.

Step 4: Process the Question

* The question goes to a **User Query module**, then to the **Embedder** to be converted into a vector (just like the data was).

Step 5: Search for Matching Info

* The query vector is sent to the **Vector Storage and Retrieval Engine**.
* It searches for the most relevant information based on meaning (semantic search).

Step 6: Generate the Answer

* The retrieved information is sent to a **Large Language Model** (like ChatGPT).
* The model uses both the question and the retrieved info to create a smart, accurate response.

Step 7: Show the Answer

* The response is sent back to the **chat interface**, where the user sees the final answer.

## RAG PROCESS- BUILDING RAG PIPELINE

A diagram of a step 3

AI-generated content may be incorrect.

A diagram of a computer

AI-generated content may be incorrect.

### STEP 1: DATA INGESTION (USING DOCUMENT LOADERS)

#### TEXT LOADER

|  |
| --- |
| **from langchain\_community.document\_loaders import TextLoader**  **import os**  **# Get the directory of the current script**  **current\_dir = os.path.dirname(os.path.abspath(\_\_file\_\_))**  **speech\_file\_path = os.path.join(current\_dir, "speech.txt")**  **text\_document = TextLoader(speech\_file\_path).load()**  **print(text\_document)** |

#### PDF LOADER

|  |  |
| --- | --- |
| INSTALL pypdf | **pip install pypdf** |

|  |
| --- |
| **from langchain\_community.document\_loaders import PyPDFLoader**  **import os**  **# Get the directory of the current script**  **current\_dir = os.path.dirname(os.path.abspath(\_\_file\_\_))**  **pdf\_file\_path = os.path.join(current\_dir, "attention.pdf")**  **pdf\_document = PyPDFLoader(pdf\_file\_path).load()**  **print(pdf\_document) 🡨 List of document with some meta data** |

#### WEB BASED LOADER

* Read the content of the web pages(Web Scraping)
* The “web\_paths” parameter will allow as to configure multiple URLs
* We can use beautiful Soup package to parse the document further

|  |
| --- |
| **web\_loader = WebBaseLoader(**  **web\_paths=("https://www.aarpmedicareplans.com/",),**  **header\_template={**  **"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36"**  **},**  **)**  **web\_document = web\_loader.load()**  **print(web\_document)** |

#### WIKIPEDIA LOADER

|  |
| --- |
|  |

1. **Step 3: Data Retrieval – Embedding And Vector Store**

* Once the documents are chunked, the next step is **embedding** and storing them in a **vector database**. Embedding transforms text into high-dimensional vectors that capture meaning and context. Each vector serves as a **numerical fingerprint** of its corresponding chunk.
* These embeddings are then stored in a vector database, which enables **fast similarity search** when a user submits a query. (Note : The vector store will have the embeddings as the metadata and the chunk itself as the data in them)
* This process ensures that relevant information can be retrieved efficiently, even if the query doesn’t use the exact wording found in the original text.

1. **Step 3: Reranking**
   1. Sometimes, the system retrieves **multiple chunks of text**, but the **best one isn’t at the top**. That’s where **reranking** helps:
      1. It takes the top few chunks (say, top 10).
      2. It uses a smarter model (called a **cross-encoder**) to **re-evaluate** how well each chunk matches the question.
      3. Then it **reorders** them so the most relevant chunk is first.
   2. Benefits
      1. It improves **accuracy**.
      2. It helps in **complex topics** like medical or legal documents.
      3. It ensures the AI gets the **best possible context** before answering.
2. **Step 4: Response Generation**
   1. When we receive a prompt from the user, we will use the embedding model to calculate the embedding for that prompt.
   2. We then use that embedding to retrieve the relevant data(chunk – it can be reranked chunks) from the vector store
   3. The prompt + the data retrieved from Vector database is send it to the LLM so that it can respond back correctly.
3. **Step 5: Evaluation**
   1. RAG pipelineis important to check whether the answers are:
      1. **Accurate**
      2. **Based on the right context**
      3. **Helpful to clinical teams**
   2. LangChain doesn’t force you to use a specific evaluation tool, but it supports:
      1. **Manual reviews**
      2. **Automated metrics**
   3. Key areas to evaluate include:
      1. **Precision and recall** of retrieved chunks
      2. **Faithfulness and factuality** of LLM-generated responses
      3. **Human feedback**
      4. **Groundedness scores** to ensure the model uses only retrieved data

#### EXAMPLE -LOAD DATA – PDF

|  |
| --- |
| **from langchain\_openai import AzureOpenAIEmbeddings, AzureChatOpenAI**  **from langchain\_community.document\_loaders import PyMuPDFLoader**  **import os**  **from dotenv import load\_dotenv**  **load\_dotenv()**  **# Set your Azure OpenAI credentials (single resource for both models)**  **subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]**  **endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]**  **api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]**  **embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]**  **gpt\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  # Your GPT model deployment**  **# Initialize the Azure OpenAI Chat model for text generation**  **llm = AzureChatOpenAI(**  **azure\_deployment=gpt\_deployment,**  **api\_version=api\_version,**  **azure\_endpoint=endpoint,**  **api\_key=subscription\_key,**  **)**  **# Initialize the Azure OpenAI Embeddings for vector creation**  **embeddings = AzureOpenAIEmbeddings(**  **azure\_deployment=embedding\_deployment,**  **api\_version=api\_version,**  **azure\_endpoint=endpoint,**  **api\_key=subscription\_key,**  **)**  **def load\_pdf\_with\_langchain(pdf\_path):**  **loader = PyMuPDFLoader(pdf\_path)**  **document = loader.load()**  **return document**  **docs = load\_pdf\_with\_langchain("rag/academic\_research\_data.pdf")**  **print("\n Sample Extracted Content:")**  **for i, doc in enumerate(docs[:2]):**  **print(f"\n--- Chunk {i + 1} ---")**  **print(doc.page\_content[:500])  # Show first 500 characters**  **print("Metadata:", doc.metadata)** |
| OUTPUT: The Pdf data is read by the loaded page by page |
| * The loader extracted the text **page by page**, and wrapped each into a Document object with metadata (like page number and file name). * This gives us a **basic form of chunking** — each page is its own chunk . For better performance in retrieval(from Vector DB), we’ll need to apply **smarter chunking strategies** next — like:   + Splitting by characters or sentences   + Ensuring chunks have overlap   + Preserving context between chunks |

### STEP 2- DATA CHUNKING

* The first step is to break down the documents into smaller, meaningful pieces – a process called **chunking**.
* Data chunking is important when dealing with huge data like large clinical trial PDFs, which can often exceed 100 pages. Large Language Models (LLMs) are not able to process such extensive content in one go, and even if they could, retrieving relevant information would be inefficient.
* By splitting the documents into bite-sized chunks, such as paragraphs or sections, we can retain sufficient context without overwhelming the system.
* In LangChain, utilities like `**RecursiveCharacterTextSplitter**` or `**MarkdownHeaderTextSplitter**` can be used depending on the structure of the data.
* These tools allow us to split text based on logical separators like headings or paragraphs, define chunk sizes (for example, 500 characters), and maintain overlaps (such as 100 characters) to ensure context is preserved across chunk boundaries.

|  |
| --- |
| **Example** |
| **document = TextLoader("rag/product-data.txt").load()**  **text\_spiltter = RecursiveCharacterTextSplitter(chunk\_size=1000, chunk\_overlap=200)**  **chunks = text\_spiltter.split\_documents(document)** |

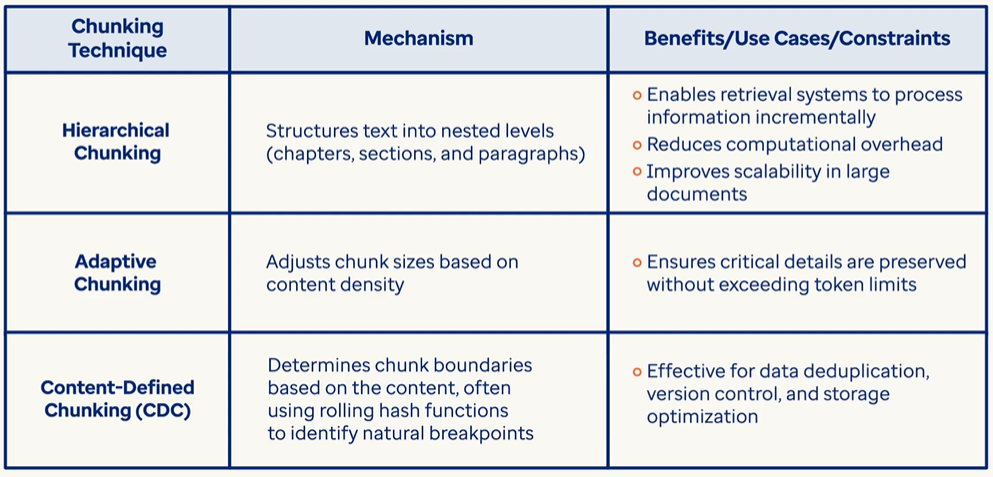
* Once we extract text from a document like a PDF, we usually don’t feed that entire text to a language model at once. It’s too large, and most of it might not be relevant to a specific question. That’s where **data chunking** comes in.
* **Chunking** is the process of breaking down a large block of text into smaller, manageable parts (called "chunks") — so they:
* **Fit within a model’s token limits**
* **Make sense on their own**
* **Can be searched and retrieved efficiently**

#### DIFFERENT CHUNKING METHODS

A diagram of a mechanism

AI-generated content may be incorrect.

CHUNKING TECHNIQUES FOR COMPLEX AND VARAIBLE DATA STRUCTURES



##### FIXED-SIZE CHUNKING

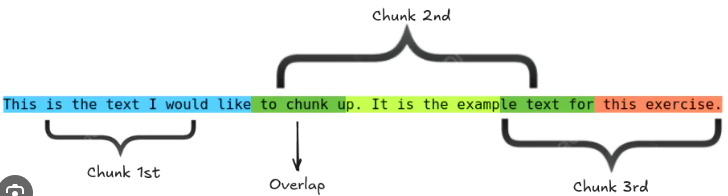
* Splits text into chunks of a fixed number of characters or tokens (e.g., 500 characters).
* It doesn’t care about sentence or paragraph boundaries.
* Though it is Simple and fast but may break sentences or lose context.

**Example**

A paragraph of 1,200 characters might be split into 3 chunks of 500, 500, and 200 characters—even if that cuts a sentence in half.

What is chunk\_size and chunk\_overlap?

* **chunk\_size**:
  + The number of characters in each chunk.
  + Example: chunk\_size=500 creates chunks of 500 characters.
* **chunk\_overlap**:
  + The number of characters repeated between chunks to keep context.
  + Example: chunk\_overlap=50 means the last 50 characters of one chunk appear at the start of the next.
  + Use overlap to avoid cutting off sentences or breaking flow across chunks.



###### TYPES OF FIXED CHUNKING

There are two types of **fixed chunking** in LangChain:

1. **CharacterTextSplitter**
2. **RecursiveCharacterTextSplitter**

Both are used to split long documents into smaller parts (chunks), but they work differently.

CharacterTextSplitter

* Splits text into chunks based on a **fixed number of characters**.
* Uses a **single separator** (e.g. space " " or newline "\n").
* If a sentence or paragraph is too long, it may split mid-sentence.
* Simple and fast, but may not always preserve context cleanly.

|  |
| --- |
| **# Define a function to split text into fixed-size character chunks using LangChain's CharacterTextSplitter.**  **from langchain\_text\_splitters import CharacterTextSplitter**  **def fixed\_size\_chunking(docs, chunk\_size=500, chunk\_overlap=50):**    **splitter = CharacterTextSplitter(**  **separator=" ",**  **chunk\_size=chunk\_size,**  **chunk\_overlap=chunk\_overlap**  **)**  **return splitter.split\_documents(docs)**  **fixed\_chunks = fixed\_size\_chunking(docs)**  **print(f" Total fixed-size chunks: {len(fixed\_chunks)}\n")**  **print(f" Example:First Chunk \n{fixed\_chunks[0].page\_content[:]}")** |

CODE : [Langchain/buildiing\_rag\_pipeline/ai\_dojo\_fixed\_size\_data\_chunking\_CharacterTextSplitter.py at main · avishekhsinhaRepo/Langchain](https://github.com/avishekhsinhaRepo/Langchain/blob/main/buildiing_rag_pipeline/ai_dojo_fixed_size_data_chunking_CharacterTextSplitter.py)

RecursiveCharacterTextSplitter

* Tries to split text at **natural boundaries**: paragraphs → sentences → words → characters.
* Uses a **list of separators**, and recursively falls back if cleaner splits aren't possible.
* Produces better structured chunks with less context loss.
* More advanced and preserves semantic meaning better.
* We use RecursiveCharacterTextSplitter when we want **smart chunking**, especially for documents with mixed formatting or longer sentences.

|  |
| --- |
| **# Splits documents using RecursiveCharacterTextSplitter which preserves context better**  **from langchain\_text\_splitters import RecursiveCharacterTextSplitter**  **def recursive\_chunking(docs, chunk\_size=500, chunk\_overlap=50):**  **splitter = RecursiveCharacterTextSplitter(**  **chunk\_size=chunk\_size,**  **chunk\_overlap=chunk\_overlap**  **)**  **return splitter.split\_documents(docs)**  **recursive\_chunks = recursive\_chunking(docs)**  **print(f" Total recursive chunks: {len(recursive\_chunks)}\n")**  **print(f" Example: First Chunk \n{recursive\_chunks[0].page\_content[:]}")** |

Example

|  |
| --- |
| Let’s say we have this text:  ***"Healthcare personalization is key. It improves outcomes.\n\nAI tools help tailor experiences."***  And we want to split it into chunks of max 40 characters.  **Step-by-step breakdown:**  Try splitting by paragraphs:   * "Healthcare personalization is key. It improves outcomes." * "AI tools help tailor experiences."   **Each paragraph is ~50 characters, so still too long.**  Try splitting by sentences:   * "Healthcare personalization is key." * "It improves outcomes." * "AI tools help tailor experiences."   **These are ~30–35 characters each — perfect!**  So the final chunks would be:  **[**  **"Healthcare personalization is key.",**  **"It improves outcomes.",**  **"AI tools help tailor experiences."**  **]** |

CODE : [Langchain/buildiing\_rag\_pipeline/ai\_dojo\_fixed\_size\_data\_chunking\_RecursiveCharacterTextSplitter.py at main · avishekhsinhaRepo/Langchain](https://github.com/avishekhsinhaRepo/Langchain/blob/main/buildiing_rag_pipeline/ai_dojo_fixed_size_data_chunking_RecursiveCharacterTextSplitter.py)

HTMLTextSplitter

* Designed to handle HTML content.
* It strips tags and splits the text into chunks of fixed size.
* Useful when are working with web pages or scraped HTML data.

JSONTextSplitter

* Works with **JSON files**.
* It extracts values from JSON keys and splits them into fixed-size chunks.
* Ideal for structured data like logs, configs, or API responses.

|  |
| --- |
| import json  import requests  from langchain\_text\_splitters import **RecursiveJsonSplitter**  json\_data = requests.get('https://api.smith.langchain.com/openapi.json').json()  json\_splitter = RecursiveJsonSplitter(max\_chunk\_size=300)  json\_chunks = json\_splitter.split\_json(json\_data)  print(json\_chunks) |

##### SENTENCE BASED CHUNKING

* **What it does**: Splits text based on complete sentences.
* **How it works**: Groups sentences together until a size limit is reached (e.g., 3–5 sentences per chunk).
* **Pros**: Preserves natural language flow.
* **Cons**: Chunk sizes can vary, and some chunks may be too short or too long.

**What is sentences\_per\_chunk?**

* **sentences\_per\_chunk** defines **how many sentences** to include in each chunk.
* Example: sentences\_per\_chunk=3 will group 3 sentences together as one chunk.
* It keeps the chunks **meaningful and readable** by not cutting through sentences.
* Great for preserving logical flow and making sure each chunk conveys a complete thought or mini-topic.

|  |  |
| --- | --- |
| INSTALL **Natural Language Toolkit (NLTK)** | pip install **nltk** |

|  |
| --- |
| **from nltk.tokenize import sent\_tokenize**  **import nltk**  **# Download required NLTK data**  **try:**  **nltk.data.find("tokenizers/punkt\_tab")**  **except LookupError:**  **print("Downloading NLTK punkt tokenizer...")**  **nltk.download("punkt\_tab")**  **# Define a function to split each page into chunks of N sentences.**  **def sentence\_based\_chunking(docs, sentences\_per\_chunk=3):**  **chunks = []**  **for doc in docs:**  **sentences = sent\_tokenize(doc.page\_content)**  **for i in range(0, len(sentences), sentences\_per\_chunk):**  **chunk\_text = " ".join(sentences[i:i + sentences\_per\_chunk])**  **chunks.append(chunk\_text)**  **return chunks**  **sentence\_chunks = sentence\_based\_chunking(docs)**  **print(f" Total sentence-based chunks: {len(sentence\_chunks)}\n")**  **print(f" Example:\n{sentence\_chunks[0][:]}")** |
| * The sentence\_based\_chunking() function breaks the text into chunks of a fixed number of sentences (e.g. 3). * This method **respects natural language flow**, making each chunk easier for the model to understand. * It’s especially helpful when documents have clean sentence structures. |
| NOTE ON **Natural Language Toolkit (NLTK)**   * **Natural Language Toolkit (NLTK)** is Python library for working with human language data. * sent\_tokenize() is used to **split a paragraph or block of text into individual sentences**   **Example Code**  **from nltk.tokenize import sent\_tokenize**  **text = "Healthcare personalization is key. It improves outcomes. AI tools help tailor experiences."**  **sentences = sent\_tokenize(text)**  **print(sentences)**  **Output:**  **[**  **"Healthcare personalization is key.",**  **"It improves outcomes.",**  **"AI tools help tailor experiences."**  **]**  Each sentence is cleanly separated, even though the original text was a single string. |

CODE : [Langchain/buildiing\_rag\_pipeline/ai-dojo-sentence-based-data-chunking.py at main · avishekhsinhaRepo/Langchain](https://github.com/avishekhsinhaRepo/Langchain/blob/main/buildiing_rag_pipeline/ai-dojo-sentence-based-data-chunking.py)

##### SEMANTIC CHUNKING

* **Semantic chunking splits content based on paragraph breaks or logical boundaries. This helps preserve the meaning of each idea.**

|  |
| --- |
| **# Define a function to split based on paragraph breaks (using two newlines)**  **def semantic\_chunking(docs):**  **chunks = []**  **for doc in docs:**  **paragraphs = doc.page\_content.split("\n\n")**  **for para in paragraphs:**  **cleaned = para.strip()**  **if cleaned:**  **chunks.append(cleaned)**  **return chunks**  **semantic\_chunks = semantic\_chunking(docs)**  **print(f" Total semantic chunks: {len(semantic\_chunks)}")**  **print(f" Example:\n{semantic\_chunks[0][:]}")** |
| * **The semantic\_chunking() function splits text by paragraph breaks, using \n\n as the divider.** * **This keeps ideas grouped by their topic or theme, preserving meaning.** * **It's great when the original text is well-formatted with clear paragraph structure.** * **Best for: structured PDFs or reports with proper formatting (like research papers).** |

CODE: [Langchain/buildiing\_rag\_pipeline/ai-dojo-semantic-data-chunking.py at main · avishekhsinhaRepo/Langchain](https://github.com/avishekhsinhaRepo/Langchain/blob/main/buildiing_rag_pipeline/ai-dojo-semantic-data-chunking.py)

### CREATING EMBEDDINGS

#### COMMON EMBEDDING TECHNIQUES

|  |  |  |
| --- | --- | --- |
| **TECHNIQUE** | **DESCRIPTION** | **COMMON USE CASES** |
| **One-Hot Encoding** | Represents words or categories as binary vectors with a single high bit | Simple categorical data, small vocabularies |
| **TF-IDF** | Weighs word frequency by inverse document frequency | Text classification, information retrieval |
| **Word2Vec** | Learns word embeddings using skip-gram or CBOW models | Semantic similarity, NLP tasks |
| **GloVe** | Global Vectors for word representation using co-occurrence statistics | NLP tasks, semantic analysis |
| **FastText** | Extension of Word2Vec that includes subword information | Morphologically rich languages |
| **ELMo** | Contextual embeddings using deep bi-directional LSTMs | Named entity recognition, sentiment analysis |
| **BERT Embeddings** | Contextual embeddings from transformer-based BERT model | Question answering, text classification |
| **Sentence Embeddings** | Embeddings for entire sentences using models like Sentence-BERT | Semantic search, clustering |
| **Doc2Vec** | Embeddings for entire documents | Document classification, retrieval |
| **Autoencoders** | Neural networks that learn compressed representations | Dimensionality reduction, anomaly detection |
| **PCA (Principal Component Analysis)** | Linear dimensionality reduction technique | Preprocessing, visualization |
| **t-SNE / UMAP** | Non-linear dimensionality reduction for visualization | Embedding visualization |
| **Graph Embeddings** | Represent nodes/edges in graphs as vectors | Social networks, recommendation systems |
| **Image Embeddings** | Feature vectors from CNNs or vision transformers | Image classification, similarity search |

#### EMBEDDING PROVIDERS/PLATFORMS

**These embedding techniques are provided by different platforms or providers. For example,**

1. OpenAI
2. Hugging Face
3. Ollama

These are **platforms or providers** that offer access to models that **implement embedding techniques**. The below tab;e explains how these platform r**elate to Embedding Techniques**

|  |  |  |
| --- | --- | --- |
| **Platform** | **Uses Which Techniques?** | **How You Use It** |
| **OpenAI** | Transformer-based embeddings (e.g., GPT) | Call API → Get embeddings |
| **Hugging Face** | Word2Vec, BERT, RoBERTa, etc. | Download model → Run locally |
| **Ollama** | Transformer-based (e.g., LLaMA2, Mistral) | Install model → Run locally via CLI |

##### OpenAI

* Think of OpenAI as a **cloud-based service**.
* You send your text to their API, and they use a **pretrained model** (like text-embedding-3-small) to generate embeddings.
* The technique behind it is based on **transformer models**, similar to BERT or GPT.
* You don’t need to worry about math or training—just use the API and get high-quality embeddings.

###### CREATING EMBEDDING USING OpenAI PLATFORM

|  |
| --- |
| **from langchain\_openai import AzureOpenAIEmbeddings, AzureChatOpenAI**  **import os**  **from dotenv import load\_dotenv**  **load\_dotenv()**  **# Set your Azure OpenAI credentials**  **subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]**  **endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]**  **api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]**  **deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]**  **embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]**  **# Model used is text-embedding-ada-002**  **# Initialize the Azure OpenAI Embeddings**  **embeddings\_client = AzureOpenAIEmbeddings(**  **azure\_deployment=embedding\_deployment,**  **api\_version=api\_version,**  **azure\_endpoint=endpoint,**  **api\_key=subscription\_key,**  **)**  **query = "This is open AI embeddings"**  **query\_embedding = embeddings\_client.embed\_query(query)**  **print(f"Embedding for query: {query\_embedding}")**  **print(f"Length of the embedding: {len(query\_embedding)}") 🡪 *Length of the embedding: 1536*** |

* In the above example – it has created vectors (embeddings) of dimension 1536.
* We can customize the dimension of embeddings too (supported by few embedding models)

|  |
| --- |
| embeddings\_client = AzureOpenAIEmbeddings(      azure\_deployment=embedding\_deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,      dimensions=1024 # Specify the dimensions for the embedding (default is 1536, but can be set to 1024 for text-embedding-ada-002  } |

Example

In the below example- we will use text document

1. To create chunks
2. Create vectors of the chunks using embedding model
3. Store in Vector DB (Chroma)
4. Retrieve the data using query from vector DB

|  |
| --- |
| **Chroma.from\_documents()** **method** :   1. Takes a list of documents (usually text chunks). 2. Convert them into **embeddings** using a specified embedding model. 3. Store those embeddings in a Chroma vector store. |

|  |
| --- |
| from langchain\_openai import AzureOpenAIEmbeddings, AzureChatOpenAI  import os  from langchain\_community.document\_loaders import TextLoader  from dotenv import load\_dotenv  from langchain\_text\_splitters import RecursiveCharacterTextSplitter  from langchain\_community.vectorstores import Chroma  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  # Model used is text-embedding-ada-002  # Initialize the Azure OpenAI Embeddings  embeddings\_client = AzureOpenAIEmbeddings(  azure\_deployment=embedding\_deployment,  api\_version=api\_version,  azure\_endpoint=endpoint,  api\_key=subscription\_key,    )  # Get the directory of the current script  current\_dir = os.path.dirname(os.path.abspath(\_\_file\_\_))  speech\_file\_path = os.path.join(current\_dir, "speech.txt")  text\_document = TextLoader(speech\_file\_path).load()  text\_splitter = RecursiveCharacterTextSplitter(chunk\_size=50, chunk\_overlap=10)  chunks = text\_splitter.split\_documents(text\_document)  **vector\_db = Chroma.from\_documents(chunks, embeddings\_client)**  query = "The world must be made safe for democracy."  result = vector\_db.similarity\_search(query, k=2)  print(f"Query: {query}")  print(f"Result: {result}") |

##### Hugging Face

* Hugging Face is like a **library of models**.
* You can download models like BERT, RoBERTa, or Sentence Transformers and run them locally.
* These models use techniques like:
  + **Word embeddings** (e.g., Word2Vec, GloVe)
  + **Contextual embeddings** (e.g., BERT, DistilBERT)
* You have full control—you can fine-tune, customize, or even train your own embedding models.

###### CREATING EMBEDDING USING HUGGINGFACE PLATFORM

##### Ollama

* Ollama is a **local-first platform**.
* It lets you run large language models (like LLaMA2 or Mistral) on your own machine.
* Some of these models support generating embeddings using transformer-based techniques.
* It’s great for privacy and offline use, but you’re limited to the models Ollama supports.

###### CREATING EMBEDDING USING OLLAMA PLATFORM

#### EMBEDDING TECHNIQUES

There are several ways to generate embeddings. Two widely used methods:

**a) Word2Vec**

* A traditional method using shallow neural networks.
* Captures the relationship between words based on **how often they appear together**.
* Learns embeddings like: “king - man + woman = queen”.

**b) AzureOpenAI using Langchain Embeddings**

OpenAI provides powerful pre-trained models for creating embeddings using their API.

**Most Common Model:**

* **text-embedding-ada-002** — small, fast, and very good.
* It produces a **1536-dimensional vector** for each chunk.
* Ideal for search, retrieval, and similarity tasks.

**How It Works:**

* You send a chunk of text to AzureOpenAI Endpoint.
* The model returns a fixed-length embedding (vector) that captures the **semantic meaning** of the text.
* AzureOpenAI using Langchain embeddings are **widely used in production RAG pipelines**, especially when quality and ease-of-use matter.

Embedding Dimensions

|  |  |  |
| --- | --- | --- |
| Method | Dimension Size | Notes |
| **Word2Vec** | Custom (e.g., 100) | Based on training setup; basic context |
| **OpenAI** | 1536 | Very high quality, best for production use |

* All methods turn text into **vectors**, but the **quality and context-awareness** improve as you go from Word2Vec → OpenAI.

#### CREATING WORD EMBEDDING USING WORD2VEC

#### AZUREOPENAI USING LANGCHAIN EMBEDDINGS

### STORING AND QUERYING VECTOR DATA

#### VECTOR DB – FAISS

* Facebook AI Similarity Search (**Faiss**) is a library for efficient similarity search and clustering of dense vectors.
* It contains algorithms that search in sets of vectors of any size, up to ones that possibly do not fit in RAM.
* It also contains supporting code for evaluation and parameter tuning.

|  |  |
| --- | --- |
| **INSTALL FAISS** | **pip install faiss-cpu** |

|  |
| --- |
| from langchain\_openai import AzureOpenAIEmbeddings, AzureChatOpenAI  import os  from langchain\_community.document\_loaders import TextLoader  from dotenv import load\_dotenv  from langchain\_text\_splitters import RecursiveCharacterTextSplitter  from langchain\_community.vectorstores import FAISS  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  # Model used is text-embedding-ada-002  # Initialize the Azure OpenAI Embeddings  embeddings\_client = AzureOpenAIEmbeddings(      azure\_deployment=embedding\_deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,)  # Get the directory of the current script  current\_dir = os.path.dirname(os.path.abspath(\_\_file\_\_))  speech\_file\_path = os.path.join(current\_dir, "speech.txt")  text\_document = TextLoader(speech\_file\_path).load()  text\_splitter = RecursiveCharacterTextSplitter(chunk\_size=50, chunk\_overlap=10)  chunks = text\_splitter.split\_documents(text\_document)  **vector\_db = FAISS.from\_documents(chunks, embeddings\_client)**  query="How does the speaker describe the desired outcome of the war?"  result = vector\_db.similarity\_search(query)  print(f"Query: {query}")  print(f"Result: {result}") |

#### VECTOR DB - CHROMA

|  |  |
| --- | --- |
| **INSTALL Chroma** | pip install Chroma |

|  |
| --- |
| from langchain\_openai import AzureOpenAIEmbeddings, AzureChatOpenAI  import os  from langchain\_community.document\_loaders import TextLoader  from dotenv import load\_dotenv  from langchain\_text\_splitters import RecursiveCharacterTextSplitter  **from langchain\_community.vectorstores import Chroma**  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  # Model used is text-embedding-ada-002  # Initialize the Azure OpenAI Embeddings  embeddings\_client = AzureOpenAIEmbeddings(      azure\_deployment=embedding\_deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,)  # Get the directory of the current script  current\_dir = os.path.dirname(os.path.abspath(\_\_file\_\_))  speech\_file\_path = os.path.join(current\_dir, "speech.txt")  text\_document = TextLoader(speech\_file\_path).load()  text\_splitter = RecursiveCharacterTextSplitter(chunk\_size=50, chunk\_overlap=10)  chunks = text\_splitter.split\_documents(text\_document)  **vector\_db = Chroma.from\_documents(chunks, embeddings\_client)**  query="How does the speaker describe the desired outcome of the war?"  result = vector\_db.similarity\_search(query)  print(f"Query: {query}")  print(f"Result: {result}") |

#### SAVE VECTOR INDEX IN LOCAL DISK

* In LangChain, **save\_local()** is a method available on certain vector store implementations (like FAISS) that allows us to persist the vector index to disk.
* This is useful when we want to avoid recomputing embeddings every time the app starts.

|  |
| --- |
| What is vector Index?   * A vector index is a data structure that stores embeddings and allows for efficient similarity search. * It enables fast retrieval of documents or chunks that are semantically similar to a query. * FAISS, Chroma, and Annoy are popular vector index libraries used in LangChain. * Process   + Generate embeddings from your documents.   + We store those embeddings in a vector index.   + Then we query the index with a new embedding (from a user query) to find similar documents. |

|  |  |
| --- | --- |
| from langchain\_openai import AzureOpenAIEmbeddings, AzureChatOpenAI  import os  from langchain\_community.document\_loaders import TextLoader  from dotenv import load\_dotenv  from langchain\_text\_splitters import RecursiveCharacterTextSplitter  from langchain\_community.vectorstores import FAISS  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  # Model used is text-embedding-ada-002  # Initialize the Azure OpenAI Embeddings  embeddings\_client = AzureOpenAIEmbeddings(      azure\_deployment=embedding\_deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key)  # Get the directory of the current script  current\_dir = os.path.dirname(os.path.abspath(\_\_file\_\_))  speech\_file\_path = os.path.join(current\_dir, "speech.txt")  text\_document = TextLoader(speech\_file\_path).load()  text\_splitter = RecursiveCharacterTextSplitter(chunk\_size=50, chunk\_overlap=10)  chunks = text\_splitter.split\_documents(text\_document)   * **Converts the chunks into embeddings and stores them in a FAISS index.** * **Saves the index locally for reuse.**   vector\_db = FAISS.from\_documents(chunks, embeddings\_client)  vector\_db.save\_local("faiss\_index")   * **Loads the saved FAISS index.** * **Performs a similarity search using the query to find the most relevant chunk(s) from the speech.**   new\_vector\_db = vector\_db.load\_local("faiss\_index",embeddings\_client,allow\_dangerous\_deserialization=True)  query="How does the speaker describe the desired outcome of the war?"  result = new\_vector\_db.similarity\_search(query)  print(f"Query: {query}")  print(f"Result: {result}") | |
| Note   * FAISS will create an adjacent folder with the name of “**index name**” * It create .pkl and .faiss file to store the index |  |

### DATA RETRIEVAL

#### RETRIEVAL CHAIN

* Retrieval Chain is a powerful pattern used to retrieve relevant information from a knowledge base (like a vector store) and then use a Language Model (LM) to generate a response based on that information.
* A Retrieval Chain combines two main components:
  + **Retriever**: Finds relevant documents or chunks based on a query.
  + **LLM** (Language Model): Uses those documents to answer the query, summarize, or perform other tasks.

**NOTE : This is commonly used in RAG (Retrieval-Augmented Generation) systems.**

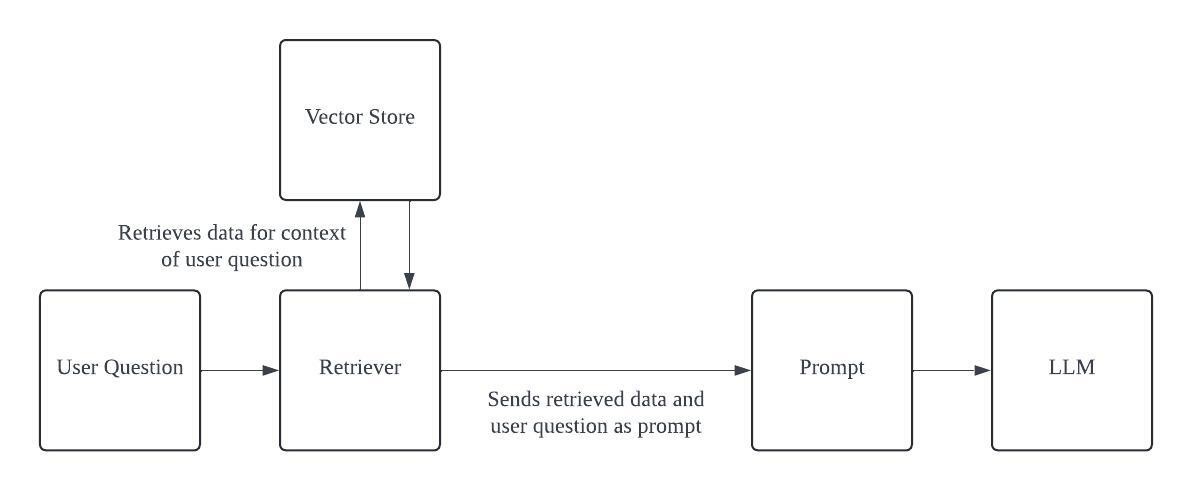
##### CORE COMPONENTS

|  |  |
| --- | --- |
| **Component** | **Role** |
| **Vector Store** | Stores embedded chunks of text (e.g., FAISS, Chroma, Pinecone). |
| **Retriever** | Queries the vector store to find relevant chunks. |
| **Prompt Template** | Defines how the retrieved content and user query are passed to the LLM. |
| **LLM Chain** | Uses the prompt and retrieved documents to generate a response. |
| **Output Parser *(optional)*** | Structures the output from the LLM. |

* In the above steps – so far, we have the data in the vector store, let’s create a retrieval chain.
* For the retrieval chain, we need a prompt. The prompt will have the retrieved data and the user question. To do this, we will use **ChatPromptTemplate**.

|  |  |
| --- | --- |
| **prompt = ChatPromptTemplate.from\_template("""Answer the following question based only on the provided context:**  **<context>**  **{context}**  **</context>**  **Question: {input}""")** | |
| **ChatPromptTemplate.from\_template(...)** | This creates a prompt that can be reused with different inputs. |
| **{context}** | This is where LangChain will insert the retrieved documents or chunks (your background info). |
| **{input}** | This is where LangChain will insert the user's question |
| **<context> ... </context>** | These tags are just for formatting — they help the model clearly see what part is the context. |
| **Example**  If the context is:  "LangChain helps build apps with language models."  And the question is: "What is LangChain?" | Answer the following question based only on the provided context:   <context>  LangChain helps build apps with language models.  </context>  Question: What is LangChain? |
| Why This Is Useful | This prompt ensures the model:   * Focuses only on the provided context * Doesn’t guess or hallucinate * Gives accurate answers based on your data |

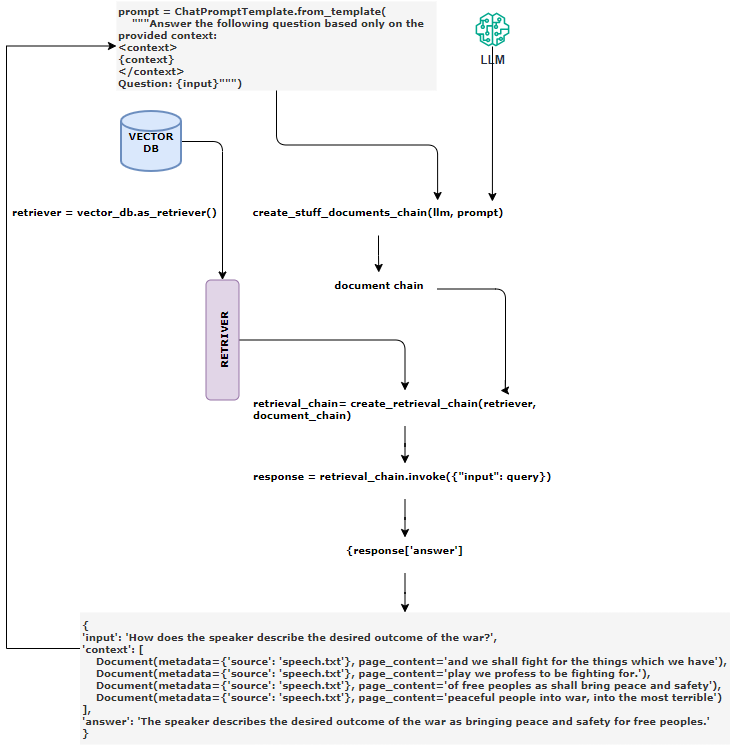
**The next step is to retrieve the data that is closely related to the user question from the vector store and provide it to the LLM as part of the prompt**



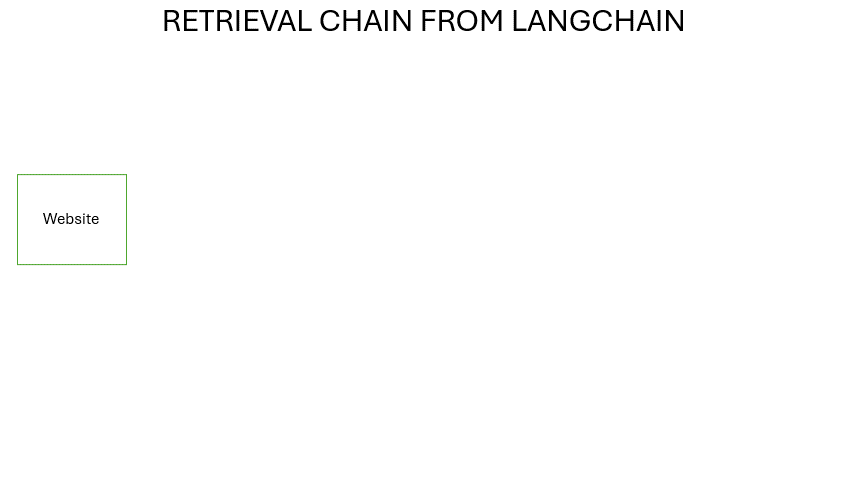
This is done using a retriever and a retrieval chain.

A diagram of a process flow

AI-generated content may be incorrect.



##### SUMMARY OF STEPS

****

##### EXAMPLE

## EXAMPLES – RAG PIPELINES

### EXAMPLE – TEXT DOCUMENT DATA SOURCE

* Data source: [Docs/Machine Learning/resources/product-data.txt at master · avishekhsinhaRepo/Docs](https://github.com/avishekhsinhaRepo/Docs/blob/master/Machine%20Learning/resources/product-data.txt)

|  |
| --- |
| from langchain\_openai import AzureOpenAIEmbeddings, AzureChatOpenAI  from langchain\_community.document\_loaders import TextLoader  from langchain\_text\_splitters import RecursiveCharacterTextSplitter  from langchain\_chroma import Chroma  from langchain.prompts import ChatPromptTemplate  import os  from dotenv import load\_dotenv  from langchain.chains import create\_retrieval\_chain  from langchain.chains.combine\_documents import create\_stuff\_documents\_chain  load\_dotenv()  # Set your Azure OpenAI credentials (single resource for both models)  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  gpt\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"] # Your GPT model deployment  # Initialize the Azure OpenAI Chat model for text generation  llm = AzureChatOpenAI(  azure\_deployment=gpt\_deployment,  api\_version=api\_version,  azure\_endpoint=endpoint,  api\_key=subscription\_key,  )  # Initialize the Azure OpenAI Embeddings for vector creation  embeddings = AzureOpenAIEmbeddings(  azure\_deployment=embedding\_deployment,  api\_version=api\_version,  azure\_endpoint=endpoint,  api\_key=subscription\_key,)  # Initialize the Azure OpenAI LLM  prompt\_template = ChatPromptTemplate.from\_messages(  [  (  "system",  """You are an assistant for answering questions.  Use the provided context to respond.If the answer  isn't clear, acknowledge that you don't know.  Limit your response to three concise sentences.  {context}  """,  ),  ("human", "{input}"),  ]  )  document = TextLoader("rag/product-data.txt").load()  text\_spiltter = RecursiveCharacterTextSplitter(chunk\_size=1000, chunk\_overlap=200)  chunks = text\_spiltter.split\_documents(document)  vector\_store = Chroma.from\_documents(chunks, embeddings)  retriever = vector\_store.as\_retriever()  qa\_chain = create\_stuff\_documents\_chain(llm, prompt\_template)  rag\_chain = create\_retrieval\_chain(retriever, qa\_chain)  question = input("Enter Your Question")  if question:  response = rag\_chain.invoke({"input": question})  print(response["answer"]) |

### EXAMPLE-2 – PDF DOCUMENT DATASOURCE

|  |  |
| --- | --- |
| INSTALL pypdf | pip install pypdf |

|  |
| --- |
| from langchain\_openai import AzureOpenAIEmbeddings, AzureChatOpenAI  from langchain\_community.document\_loaders import PyPDFLoader  from langchain\_text\_splitters import RecursiveCharacterTextSplitter  from langchain\_chroma import Chroma  from langchain.prompts import ChatPromptTemplate  import os  from dotenv import load\_dotenv  from langchain.chains import create\_retrieval\_chain  from langchain.chains.combine\_documents import create\_stuff\_documents\_chain  load\_dotenv()  # Set your Azure OpenAI credentials (single resource for both models)  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  gpt\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  # Your GPT model deployment  # Initialize the Azure OpenAI Chat model for text generation  llm = AzureChatOpenAI(      azure\_deployment=gpt\_deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,)  # Initialize the Azure OpenAI Embeddings for vector creation  embeddings = AzureOpenAIEmbeddings(      azure\_deployment=embedding\_deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,  )  # Initialize the Azure OpenAI LLM  prompt\_template = ChatPromptTemplate.from\_messages(      [          (              "system",              """You are an assistant for answering questions.                  Use the provided context to respond.If the answer                  isn't clear, acknowledge that you don't know.                  Limit your response to three concise sentences.                  {context}        """,          ),          ("human", "{input}"),      ]  )  document = PyPDFLoader("rag/academic\_research\_data.pdf").load()  text\_spiltter = RecursiveCharacterTextSplitter(chunk\_size=1000, chunk\_overlap=200)  chunks = text\_spiltter.split\_documents(document)  vector\_store = Chroma.from\_documents(chunks, embeddings)  retriever = vector\_store.as\_retriever()  qa\_chain = create\_stuff\_documents\_chain(llm, prompt\_template)  rag\_chain = create\_retrieval\_chain(retriever, qa\_chain)  question = input("Enter Your Question=")  if question:      response = rag\_chain.invoke({"input": question})      print(response["answer"]) |

### EXAMPLE- 3 – WORD DOCUMENT DATASOURCE

|  |  |
| --- | --- |
| INSTALL docx2txt | pip install docx2txt |

|  |
| --- |
| from langchain\_openai import AzureOpenAIEmbeddings, AzureChatOpenAI  from langchain\_community.document\_loaders import Docx2txtLoader  from langchain\_text\_splitters import RecursiveCharacterTextSplitter  from langchain\_chroma import Chroma  from langchain.prompts import ChatPromptTemplate  import os  from dotenv import load\_dotenv  from langchain.chains import create\_retrieval\_chain  from langchain.chains.combine\_documents import create\_stuff\_documents\_chain  load\_dotenv()  # Set your Azure OpenAI credentials (single resource for both models)  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  embedding\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT\_FOR\_EMBEDDING"]  gpt\_deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  # Your GPT model deployment  # Initialize the Azure OpenAI Chat model for text generation  llm = AzureChatOpenAI(      azure\_deployment=gpt\_deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,  )  # Initialize the Azure OpenAI Embeddings for vector creation  embeddings = AzureOpenAIEmbeddings(      azure\_deployment=embedding\_deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,)  # Initialize the Azure OpenAI LLM  prompt\_template = ChatPromptTemplate.from\_messages(      [          (              "system",              """You are an assistant for answering questions.                  Use the provided context to respond.If the answer                  isn't clear, acknowledge that you don't know.                  Limit your response to three concise sentences.                  {context}        """,          ),          ("human", "{input}"),      ]  )  document = Docx2txtLoader("rag/Expenses\_policy.docx").load()  text\_spiltter = RecursiveCharacterTextSplitter(chunk\_size=1000, chunk\_overlap=200)  chunks = text\_spiltter.split\_documents(document)  vector\_store = Chroma.from\_documents(chunks, embeddings)  retriever = vector\_store.as\_retriever()  qa\_chain = create\_stuff\_documents\_chain(llm, prompt\_template)  rag\_chain = create\_retrieval\_chain(retriever, qa\_chain)  question = input("Enter Your Question=")  if question:      response = rag\_chain.invoke({"input": question})      print(response["answer"]) |

# LANGSMITH

**LangSmith** is a platform designed to help developers build, debug, and evaluate applications powered by large language models (LLMs). It provides tools for:

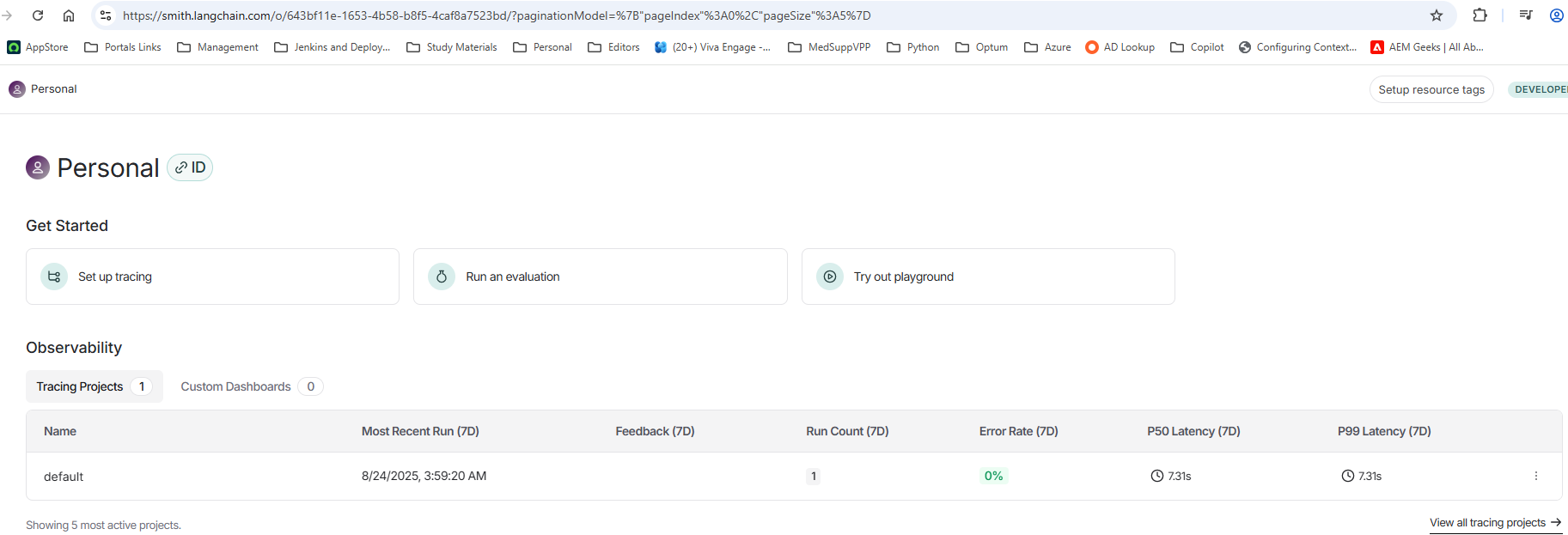
* + **Observability**: Tracking and analyzing how LLMs respond to prompts.
  + **Evaluation**: Measuring the quality and reliability of model outputs.
  + **Prompt Engineering**: Iterating and refining prompts to improve performance.

It works independently or alongside frameworks like LangChain and LangGraph, making it useful for both prototyping and production-grade AI systems.

## STEP TO SET UP LANGSMITH

* Step 1: Login /Signup to <https://smith.langchain.com/>

LangSmith Dashboard



### CREATE API KEY

1. Go to Settings 🡪 Create API Key

A screenshot of a computer

AI-generated content may be incorrect.

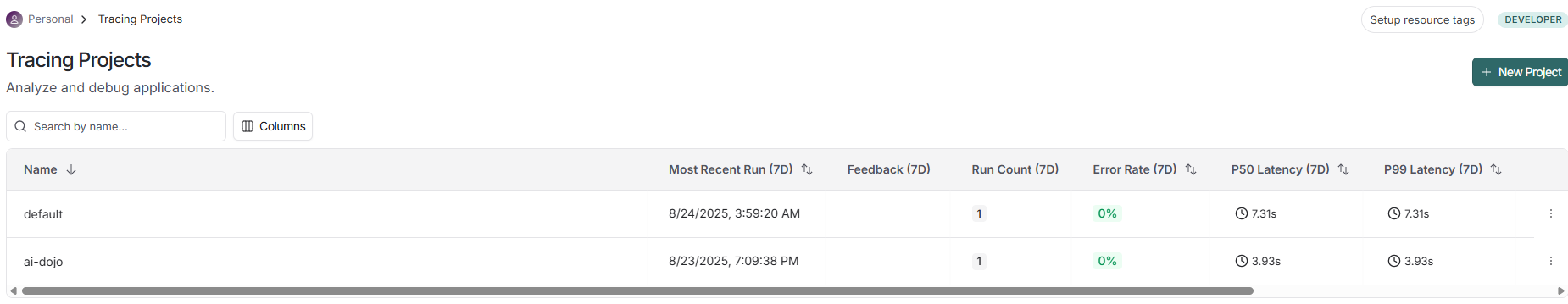
### CODE

|  |
| --- |
| from langchain\_openai import AzureChatOpenAI  import os  from dotenv import load\_dotenv  load\_dotenv()  # Set your Azure OpenAI credentials  subscription\_key = os.environ["AZURE\_OPENAI\_API\_KEY"]  endpoint = os.environ["AZURE\_OPENAI\_ENDPOINT"]  api\_version = os.environ["AZURE\_OPENAI\_API\_VERSION"]  deployment = os.environ["AZURE\_OPENAI\_DEPLOYMENT"]  os.environ["LANGSMITH\_TRACING"] = os.getenv("LANGSMITH\_TRACING")  os.environ["LANGSMITH\_API\_KEY"] = os.getenv("LANGSMITH\_API\_KEY")  os.environ["LANGSMITH\_PROJECT"] = os.getenv("LANGSMITH\_PROJECT")  # Initialize the Azure OpenAI LLM  llm = AzureChatOpenAI(      azure\_deployment=deployment,      api\_version=api\_version,      azure\_endpoint=endpoint,      api\_key=subscription\_key,      temperature=0)  response = llm.invoke("What is Generative AI?")  print(response) |
| **LANGSMITH\_TRACING=true**  **LANGSMITH\_API\_KEY=<lang\_chain\_api\_key>**  **LANGSMITH\_PROJECT=ai-dojo** |

### TRACING PROJECT

A screenshot of a computer

AI-generated content may be incorrect.



* New Tracing project will be created “ai-dojo”
* It will show the trace of all the inputs & outputs.

A screenshot of a computer

AI-generated content may be incorrect.