## Load OpenAI API Credentials

Here we load it from a file so we don't explore the credentials on the internet by mistake

# File QA RAG Chatbot App with ChatGPT, Llamaindex

Here we will implement an advanced RAG System with ChatGPT, LLamaindex to build a QA chatbot to analyze a LIC policy document.
Chatbot will have the following features:

- PDF Document Upload and Indexing
- RAG System for query analysis and response

## Environment Setup

```
import locale
locale.getpreferredencoding = lambda: "UTF-8"
```

```
import yaml

with open('api_credentials.yml', 'r') as file:
    api_creds = yaml.safe_load(file)
```

```
api_creds.keys()
```

```
dict_keys(['openai_key', 'ngrok_key', 'llamaindex_key', 'HF_Token', 'jina_key'])
```

```
import os

os.environ['OPENAI_API_KEY'] = api_creds['openai_key']
os.environ["LLAMA_API_KEY"] = api_creds['llamaindex_key']
os.environ["HF_TOKEN"] = api_creds['HF_Token']
os.environ["JINA_API_KEY"] = api_creds['jina_key']
```

```
Start coding or generate with AI.
```

```
print(api_creds['HF_Token'])
```

```
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
```

## Install Dependencies

```
!pip install --upgrade pip
!pip install llama-index
!pip install llama-index-llamaparse
!pip install llama-index PyMuPDF
!pip install chromadb
!pip install llama-index-vector-stores-chroma
!pip install openai
!pip install llama-index-llms-huggingface
!pip install sentence-transformers
!pip install jina
!pip install llama-index-embeddings-huggingface
#!pip install llama-index-embeddings-jinaai
```

```
Requirement already satisfied: pip in /usr/local/lib/python3.11/dist-packages (24.1.2)
Collecting pip
  Downloading pip-25.1.1-py3-none-any.whl.metadata (3.6 kB)
Downloading pip-25.1.1-py3-none-any.whl (1.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 26.9 MB/s eta 0:00:00
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 24.1.2
    Uninstalling pip-24.1.2:
      Successfully uninstalled pip-24.1.2
Successfully installed pip-25.1.1
Collecting llama-index
  Downloading llama_index-0.12.46-py3-none-any.whl.metadata (12 kB)
Collecting llama-index-agent-openai<0.5,>=0.4.0 (from llama-index)
  Downloading llama_index_agent_openai-0.4.12-py3-none-any.whl.metadata (439 bytes)
Collecting llama-index-cli<0.5,>=0.4.2 (from llama-index)
  Downloading llama_index_cli-0.4.3-py3-none-any.whl.metadata (1.4 kB)
Collecting llama-index-core<0.13,>=0.12.46 (from llama-index)
  Downloading llama_index_core-0.12.46-py3-none-any.whl.metadata (2.5 kB)
Collecting llama-index-embeddings-openai<0.4,>=0.3.0 (from llama-index)
  Downloading llama_index_embeddings_openai-0.3.1-py3-none-any.whl.metadata (684 bytes)
Collecting llama-index-indices-managed-llama-cloud>=0.4.0 (from llama-index)
  Downloading llama_index_indices_managed_llama_cloud-0.7.8-py3-none-any.whl.metadata (3.3 kB)
Collecting llama-index-llms-openai<0.5,>=0.4.0 (from llama-index)
  Downloading llama_index_llms_openai-0.4.7-py3-none-any.whl.metadata (3.0 kB)
Collecting llama-index-multi-modal-llms-openai<0.6,>=0.5.0 (from llama-index)
  Downloading llama_index_multi_modal_llms_openai-0.5.1-py3-none-any.whl.metadata (440 bytes)
Collecting llama-index-program-openai<0.4,>=0.3.0 (from llama-index)
  Downloading llama_index_program_openai-0.3.2-py3-none-any.whl.metadata (473 bytes)
Collecting llama-index-question-gen-openai<0.4,>=0.3.0 (from llama-index)
  Downloading llama_index_question_gen_openai-0.3.1-py3-none-any.whl.metadata (492 bytes)
Collecting llama-index-readers-file<0.5,>=0.4.0 (from llama-index)
  Downloading llama_index_readers_file-0.4.9-py3-none-any.whl.metadata (5.2 kB)
Collecting llama-index-readers-llama-parse>=0.4.0 (from llama-index)
  Downloading llama_index_readers_llama_parse-0.4.0-py3-none-any.whl.metadata (3.6 kB)
Requirement already satisfied: nltk>3.8.1 in /usr/local/lib/python3.11/dist-packages (from llama-index) (3.9.1)
Requirement already satisfied: openai>=1.14.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-agent-openai<0.5,>=0.4.0->llama-index) (1.93.0)
Requirement already satisfied: aiohttp<4,>=3.8.6 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (3.11.15)
Collecting aiosqlite (from llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading aiosqlite-0.21.0-py3-none-any.whl.metadata (4.3 kB)
Collecting banks<3,>=2.0.0 (from llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading banks-2.1.3-py3-none-any.whl.metadata (12 kB)
Collecting dataclasses-json (from llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading dataclasses_json-0.6.7-py3-none-any.whl.metadata (25 kB)
Collecting deprecated>=1.2.9.3 (from llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading Deprecated-1.2.18-py2.py3-none-any.whl.metadata (5.7 kB)
Collecting dirtyjson<2,>=1.0.8 (from llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading dirtyjson-1.0.8-py3-none-any.whl.metadata (11 kB)
Collecting filetype<2,>=1.2.0 (from llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading filetype-1.2.0-py2.py3-none-any.whl.metadata (6.5 kB)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (2025.3.2)
Requirement already satisfied: httpx in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (0.28.1)
Collecting llama-index-workflows<2,>=1.0.1 (from llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading llama_index_workflows-1.0.1-py3-none-any.whl.metadata (5.5 kB)
Requirement already satisfied: nest-asyncio<2,>=1.5.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (1.6.0)
Requirement already satisfied: networkx>=3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (3.5)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (2.0.2)
Requirement already satisfied: pillow>=9.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (11.2.1)
Requirement already satisfied: pydantic>=2.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (2.11.7)
Requirement already satisfied: pyyaml>=6.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (6.0.2)
Requirement already satisfied: requests>=2.31.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (2.32.3)
```

```
Collecting setuptools>=80.9.0 (from llama-index-core<0.13,>=0.12.46->llama-index)
  Using cached setuptools-80.9.0-py3-none-any.whl.metadata (6.6 kB)
Requirement already satisfied: sqlalchemy>=1.4.49 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13,>=0.12.46->llama-index)
Requirement already satisfied: tenacity!=8.4.0,<10.0.0,>=8.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (8.5.0)
Requirement already satisfied: tiktoken>=0.7.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (0.9.0)
Requirement already satisfied: tqdm<5,>=4.66.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (4.67.1)
Requirement already satisfied: typing-extensions>=4.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (4.14.0)
Requirement already satisfied: typing-inspect>=0.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (0.9.0)
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (1.17.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (2.6
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (1.7.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (6.6.3)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (0.3.2)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (1.20.1)
Collecting griffe (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading griffe-1.7.3-py3-none-any.whl.metadata (5.0 kB)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index) (3.1.6)
Requirement already satisfied: platformdirs in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index) (4.3.8)
Requirement already satisfied: beautifulsoup4<5,>=4.12.3 in /usr/local/lib/python3.11/dist-packages (from llama-index-readers-file<0.5,>=0.4.0->llama-index) (4.13.4)
Requirement already satisfied: pandas<2.3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-readers-file<0.5,>=0.4.0->llama-index) (2.2.2)
Collecting pypdf<6,>=5.1.0 (from llama-index-readers-file<0.5,>=0.4.0->llama-index)
  Downloading pypdf-5.7.0-py3-none-any.whl.metadata (7.2 kB)
Collecting striprtf<0.0.27,>=0.0.26 (from llama-index-readers-file<0.5,>=0.4.0->llama-index)
  Downloading striprtf-0.0.26-py3-none-any.whl.metadata (2.1 kB)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.11/dist-packages (from beautifulsoup4<5,>=4.12.3->llama-index-readers-file<0.5,>=0.4.0->llama-index) (
Collecting llama-index-instrumentation>=0.1.0 (from llama-index-workflows<2,>=1.0.1->llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading llama_index_instrumentation-0.2.0-py3-none-any.whl.metadata (252 bytes)
Requirement already satisfied: anyio<5,>=3.5.0 in /usr/local/lib/python3.11/dist-packages (from openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index) (4.9.0)
Requirement already satisfied: distro<2,>=1.7.0 in /usr/local/lib/python3.11/dist-packages (from openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index) (1.9.0)
Requirement already satisfied: jiter<1,>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index) (0.10.0)
Requirement already satisfied: sniffio in /usr/local/lib/python3.11/dist-packages (from openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index) (1.3.1)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.11/dist-packages (from anyio<5,>=3.5.0->openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index)
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13,>=0.12.46->llama-index) (2025.6.15)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13,>=0.12.46->llama-index) (1.0.9)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx->llama-index-core<0.13,>=0.12.46->llama-index) (0.16.0)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas<2.3.0->llama-index-readers-file<0.5,>=0.4.0->llama-index) (2.9.
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas<2.3.0->llama-index-readers-file<0.5,>=0.4.0->llama-index) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas<2.3.0->llama-index-readers-file<0.5,>=0.4.0->llama-index) (2025.2)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.46->llama-index) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.46->llama-index) (2.33.2)
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.46->llama-index) (0.4.
Collecting llama-cloud==0.1.30 (from llama-index-indices-managed-llama-cloud>=0.4.0->llama-index)
  Downloading llama_cloud-0.1.30-py3-none-any.whl.metadata (1.2 kB)
Collecting llama-parse>=0.5.0 (from llama-index-readers-llama-parse>=0.4.0->llama-index)
  Downloading llama_parse-0.6.41-py3-none-any.whl.metadata (6.9 kB)
Collecting llama-cloud-services>=0.6.41 (from llama-parse>=0.5.0->llama-index-readers-llama-parse>=0.4.0->llama-index)
  Downloading llama_cloud_services-0.6.41-py3-none-any.whl.metadata (3.5 kB)
Requirement already satisfied: click<9.0.0,>=8.1.7 in /usr/local/lib/python3.11/dist-packages (from llama-cloud-services>=0.6.41->llama-parse>=0.5.0->llama-index-readers-lla
Requirement already satisfied: python-dotenv<2.0.0,>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-cloud-services>=0.6.41->llama-parse>=0.5.0->llama-index-rea
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index) (1.5.1)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index) (2024.11.6)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas<2.3.0->llama-index-readers-file<0.5,>=0.4.0->llama-in
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.46->llama-index) (3.4
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.46->llama-index) (1.26.20)
Requirement already satisfied: greenlet>=1 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy>=1.4.49->sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13,>=0.12.46-
Requirement already satisfied: mypy-extensions>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from typing-inspect>=0.8.0->llama-index-core<0.13,>=0.12.46->llama-index) (
Collecting marshmallow<4.0.0,>=3.18.0 (from dataclasses-json->llama-index-core<0.13,>=0.12.46->llama-index)
  Downloading marshmallow-3.26.1-py3-none-any.whl.metadata (7.3 kB)
Requirement already satisfied: packaging>=17.0 in /usr/local/lib/python3.11/dist-packages (from marshmallow<4.0.0,>=3.18.0->dataclasses-json->llama-index-core<0.13,>=0.12.46
Collecting colorama>=0.4 (from griffe->banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index)
```

```
  Downloading colorama-0.4.6-py2.py3-none-any.whl.metadata (17 kB)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index) (3.0.2
Downloading llama_index-0.12.46-py3-none-any.whl (7.1 kB)
Downloading llama_index_agent_openai-0.4.12-py3-none-any.whl (14 kB)
Downloading llama_index_cli-0.4.3-py3-none-any.whl (28 kB)
Downloading llama_index_core-0.12.46-py3-none-any.whl (7.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.6/7.6 MB 115.1 MB/s eta 0:00:00
Downloading banks-2.1.3-py3-none-any.whl (28 kB)
Downloading dirtyjson-1.0.8-py3-none-any.whl (25 kB)
Downloading filetype-1.2.0-py2.py3-none-any.whl (19 kB)
Downloading llama_index_embeddings_openai-0.3.1-py3-none-any.whl (6.2 kB)
Downloading llama_index_llms_openai-0.4.7-py3-none-any.whl (25 kB)
Downloading llama_index_multi_modal_llms_openai-0.5.1-py3-none-any.whl (3.4 kB)
Downloading llama_index_program_openai-0.3.2-py3-none-any.whl (6.1 kB)
Downloading llama_index_question_gen_openai-0.3.1-py3-none-any.whl (3.7 kB)
Downloading llama_index_readers_file-0.4.9-py3-none-any.whl (40 kB)
Downloading llama_index_workflows-1.0.1-py3-none-any.whl (36 kB)
Downloading pypdf-5.7.0-py3-none-any.whl (305 kB)
Downloading striprtf-0.0.26-py3-none-any.whl (6.9 kB)
Downloading Deprecated-1.2.18-py2.py3-none-any.whl (10.0 kB)
Downloading llama_index_indices_managed_llama_cloud-0.7.8-py3-none-any.whl (16 kB)
Downloading llama_cloud-0.1.30-py3-none-any.whl (282 kB)
Downloading llama_index_instrumentation-0.2.0-py3-none-any.whl (14 kB)
Downloading llama_index_readers_llama_parse-0.4.0-py3-none-any.whl (2.5 kB)
Downloading llama_parse-0.6.41-py3-none-any.whl (4.9 kB)
Downloading llama_cloud_services-0.6.41-py3-none-any.whl (40 kB)
Using cached setuptools-80.9.0-py3-none-any.whl (1.2 MB)
Downloading aiosqlite-0.21.0-py3-none-any.whl (15 kB)
Downloading dataclasses_json-0.6.7-py3-none-any.whl (28 kB)
Downloading marshmallow-3.26.1-py3-none-any.whl (50 kB)
Downloading griffe-1.7.3-py3-none-any.whl (129 kB)
Downloading colorama-0.4.6-py2.py3-none-any.whl (25 kB)
Installing collected packages: striprtf, filetype, dirtyjson, setuptools, pypdf, marshmallow, deprecated, colorama, aiosqlite, griffe, dataclasses-json, llama-index-instrume
  Attempting uninstall: setuptools
    Found existing installation: setuptools 75.2.0
    Uninstalling setuptools-75.2.0:
      Successfully uninstalled setuptools-75.2.0
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 29/29 [llama-index]
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflicts
ipython 7.34.0 requires jedi>=0.16, which is not installed.
Successfully installed aiosqlite-0.21.0 banks-2.1.3 colorama-0.4.6 dataclasses-json-0.6.7 deprecated-1.2.18 dirtyjson-1.0.8 filetype-1.2.0 griffe-1.7.3 llama-cloud-0.1.30 ll
WARNING: The following packages were previously imported in this runtime:
  [_distutils_hack,pkg_resources,setuptools]
You must restart the runtime in order to use newly installed versions.
```

RESTART SESSION

```
ERROR: Could not find a version that satisfies the requirement llama-index-llamaparse (from versions: none)
ERROR: No matching distribution found for llama-index-llamaparse
Requirement already satisfied: llama-index in /usr/local/lib/python3.11/dist-packages (0.12.46)
Collecting PyMuPDF
  Downloading pymupdf-1.26.3-cp39-abi3-manylinux_2_28_x86_64.whl.metadata (3.4 kB)
Requirement already satisfied: llama-index-agent-openai<0.5,>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.4.12)
Requirement already satisfied: llama-index-cli<0.5,>=0.4.2 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.4.3)
Requirement already satisfied: llama-index-core<0.13,>=0.12.46 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.12.46)
Requirement already satisfied: llama-index-embeddings-openai<0.4,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.3.1)
Requirement already satisfied: llama-index-indices-managed-llama-cloud>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.7.8)
Requirement already satisfied: llama-index-llms-openai<0.5,>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.4.7)
Requirement already satisfied: llama-index-multi-modal-llms-openai<0.6,>=0.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.5.1)
Requirement already satisfied: llama-index-program-openai<0.4,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.3.2)
Requirement already satisfied: llama-index-question-gen-openai<0.4,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.3.1)
Requirement already satisfied: llama-index-readers-file<0.5,>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.4.9)
```

```
Requirement already satisfied: llama-index-readers-llama-parse>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from llama-index) (0.4.0)
Requirement already satisfied: nltk>3.8.1 in /usr/local/lib/python3.11/dist-packages (from llama-index) (3.9.1)
Requirement already satisfied: openai>=1.14.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-agent-openai<0.5,>=0.4.0->llama-index) (1.93.0)
Requirement already satisfied: aiohttp<4,>=3.8.6 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (3.11.15)
Requirement already satisfied: aiosqlite in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (0.21.0)
Requirement already satisfied: banks<3,>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (2.1.3)
Requirement already satisfied: dataclasses-json in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (0.6.7)
Requirement already satisfied: deprecated>=1.2.9.3 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (1.2.18)
Requirement already satisfied: dirtyjson<2,>=1.0.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (1.0.8)
Requirement already satisfied: filetype<2,>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (1.2.0)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (2025.3.2)
Requirement already satisfied: httpx in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (0.28.1)
Requirement already satisfied: llama-index-workflows<2,>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (1.0.1)
Requirement already satisfied: nest-asyncio<2,>=1.5.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (1.6.0)
Requirement already satisfied: networkx>=3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (3.5)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (2.0.2)
Requirement already satisfied: pillow>=9.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (11.2.1)
Requirement already satisfied: pydantic>=2.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (2.11.7)
Requirement already satisfied: pyyaml>=6.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (6.0.2)
Requirement already satisfied: requests>=2.31.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (2.32.3)
Requirement already satisfied: setuptools>=80.9.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (80.9.0)
Requirement already satisfied: sqlalchemy>=1.4.49 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13,>=0.12.46->llama-index)
Requirement already satisfied: tenacity!=8.4.0,<10.0.0,>=8.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (8.5.0)
Requirement already satisfied: tiktoken>=0.7.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (0.9.0)
Requirement already satisfied: tqdm<5,>=4.66.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (4.67.1)
Requirement already satisfied: typing-extensions>=4.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (4.14.0)
Requirement already satisfied: typing-inspect>=0.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (0.9.0)
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.46->llama-index) (1.17.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (2.6
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (1.7.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (6.6.3)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (0.3.2)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.46->llama-index) (1.20.1)
Requirement already satisfied: griffe in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index) (1.7.3)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index) (3.1.6)
Requirement already satisfied: platformdirs in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index) (4.3.8)
Requirement already satisfied: beautifulsoup4<5,>=4.12.3 in /usr/local/lib/python3.11/dist-packages (from llama-index-readers-file<0.5,>=0.4.0->llama-index) (4.13.4)
Requirement already satisfied: pandas<2.3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-readers-file<0.5,>=0.4.0->llama-index) (2.2.2)
Requirement already satisfied: pypdf<6,>=5.1.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-readers-file<0.5,>=0.4.0->llama-index) (5.7.0)
Requirement already satisfied: striprtf<0.0.27,>=0.0.26 in /usr/local/lib/python3.11/dist-packages (from llama-index-readers-file<0.5,>=0.4.0->llama-index) (0.0.26)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.11/dist-packages (from beautifulsoup4<5,>=4.12.3->llama-index-readers-file<0.5,>=0.4.0->llama-index) (
Requirement already satisfied: llama-index-instrumentation>=0.1.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-workflows<2,>=1.0.1->llama-index-core<0.13,>=0
Requirement already satisfied: anyio<5,>=3.5.0 in /usr/local/lib/python3.11/dist-packages (from openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index) (4.9.0)
Requirement already satisfied: distro<2,>=1.7.0 in /usr/local/lib/python3.11/dist-packages (from openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index) (1.9.0)
Requirement already satisfied: jiter<1,>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index) (0.10.0)
Requirement already satisfied: sniffio in /usr/local/lib/python3.11/dist-packages (from openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index) (1.3.1)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.11/dist-packages (from anyio<5,>=3.5.0->openai>=1.14.0->llama-index-agent-openai<0.5,>=0.4.0->llama-index)
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13,>=0.12.46->llama-index) (2025.6.15)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13,>=0.12.46->llama-index) (1.0.9)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx->llama-index-core<0.13,>=0.12.46->llama-index) (0.16.0)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas<2.3.0->llama-index-readers-file<0.5,>=0.4.0->llama-index) (2.9.
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas<2.3.0->llama-index-readers-file<0.5,>=0.4.0->llama-index) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas<2.3.0->llama-index-readers-file<0.5,>=0.4.0->llama-index) (2025.2)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.46->llama-index) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.46->llama-index) (2.33.2)
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.46->llama-index) (0.4.
Requirement already satisfied: llama-cloud==0.1.30 in /usr/local/lib/python3.11/dist-packages (from llama-index-indices-managed-llama-cloud>=0.4.0->llama-index) (0.1.30)
Requirement already satisfied: llama-parse>=0.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-readers-llama-parse>=0.4.0->llama-index) (0.6.41)
Requirement already satisfied: llama-cloud-services>=0.6.41 in /usr/local/lib/python3.11/dist-packages (from llama-parse>=0.5.0->llama-index-readers-llama-parse>=0.4.0->llam
```

```
Requirement already satisfied: click<9.0.0,>=8.1.7 in /usr/local/lib/python3.11/dist-packages (from llama-cloud-services>=0.6.41->llama-parse>=0.5.0->llama-index-readers-lla
Requirement already satisfied: python-dotenv<2.0.0,>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-cloud-services>=0.6.41->llama-parse>=0.5.0->llama-index-rea
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index) (1.5.1)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index) (2024.11.6)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas<2.3.0->llama-index-readers-file<0.5,>=0.4.0->llama-in
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.46->llama-index) (3.4
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.46->llama-index) (1.26.20)
Requirement already satisfied: greenlet>=1 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy>=1.4.49->sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13,>=0.12.46-
Requirement already satisfied: mypy-extensions>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from typing-inspect>=0.8.0->llama-index-core<0.13,>=0.12.46->llama-index) (
Requirement already satisfied: marshmallow<4.0.0,>=3.18.0 in /usr/local/lib/python3.11/dist-packages (from dataclasses-json->llama-index-core<0.13,>=0.12.46->llama-index) (3
Requirement already satisfied: packaging>=17.0 in /usr/local/lib/python3.11/dist-packages (from marshmallow<4.0.0,>=3.18.0->dataclasses-json->llama-index-core<0.13,>=0.12.46
Requirement already satisfied: colorama>=0.4 in /usr/local/lib/python3.11/dist-packages (from griffe->banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index) (0.4.6)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.46->llama-index) (3.0.2
Downloading pymupdf-1.26.3-cp39-abi3-manylinux_2_28_x86_64.whl (24.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.1/24.1 MB 65.4 MB/s eta 0:00:00
Installing collected packages: PyMuPDF
Successfully installed PyMuPDF-1.26.3
Collecting chromadb
  Downloading chromadb-1.0.15-cp39-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (7.0 kB)
Requirement already satisfied: build>=1.0.3 in /usr/local/lib/python3.11/dist-packages (from chromadb) (1.2.2.post1)
Requirement already satisfied: pydantic>=1.9 in /usr/local/lib/python3.11/dist-packages (from chromadb) (2.11.7)
Collecting pybase64>=1.4.1 (from chromadb)
  Downloading pybase64-1.4.1-cp311-cp311-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (8.4 kB)
Requirement already satisfied: uvicorn>=0.18.3 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb) (0.23.1)
Requirement already satisfied: numpy>=1.22.5 in /usr/local/lib/python3.11/dist-packages (from chromadb) (2.0.2)
Collecting posthog<6.0.0,>=2.4.0 (from chromadb)
  Downloading posthog-5.4.0-py3-none-any.whl.metadata (5.7 kB)
Requirement already satisfied: typing-extensions>=4.5.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (4.14.0)
Collecting onnxruntime>=1.14.1 (from chromadb)
  Downloading onnxruntime-1.22.0-cp311-cp311-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl.metadata (4.5 kB)
Requirement already satisfied: opentelemetry-api>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (1.34.1)
Requirement already satisfied: opentelemetry-exporter-otlp-proto-grpc>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (1.34.1)
Requirement already satisfied: opentelemetry-sdk>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (1.34.1)
Requirement already satisfied: tokenizers>=0.13.2 in /usr/local/lib/python3.11/dist-packages (from chromadb) (0.21.2)
Collecting pypika>=0.48.9 (from chromadb)
  Downloading PyPika-0.48.9.tar.gz (67 kB)
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
Requirement already satisfied: tqdm>=4.65.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (4.67.1)
Collecting overrides>=7.3.1 (from chromadb)
  Downloading overrides-7.7.0-py3-none-any.whl.metadata (5.8 kB)
Requirement already satisfied: importlib-resources in /usr/local/lib/python3.11/dist-packages (from chromadb) (6.5.2)
Requirement already satisfied: grpcio>=1.58.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (1.68.0)
Collecting bcrypt>=4.0.1 (from chromadb)
  Downloading bcrypt-4.3.0-cp39-abi3-manylinux_2_34_x86_64.whl.metadata (10 kB)
Requirement already satisfied: typer>=0.9.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (0.16.0)
Collecting kubernetes>=28.1.0 (from chromadb)
  Downloading kubernetes-33.1.0-py2.py3-none-any.whl.metadata (1.7 kB)
Requirement already satisfied: tenacity>=8.2.3 in /usr/local/lib/python3.11/dist-packages (from chromadb) (8.5.0)
Requirement already satisfied: pyyaml>=6.0.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (6.0.2)
Collecting mmh3>=4.0.1 (from chromadb)
  Downloading mmh3-5.1.0-cp311-cp311-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (16 kB)
Requirement already satisfied: orjson>=3.9.12 in /usr/local/lib/python3.11/dist-packages (from chromadb) (3.10.18)
Requirement already satisfied: httpx>=0.27.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (0.28.1)
Requirement already satisfied: rich>=10.11.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (13.9.4)
Requirement already satisfied: jsonschema>=4.19.0 in /usr/local/lib/python3.11/dist-packages (from chromadb) (4.24.0)
Requirement already satisfied: requests<3.0,>=2.7 in /usr/local/lib/python3.11/dist-packages (from posthog<6.0.0,>=2.4.0->chromadb) (2.32.3)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from posthog<6.0.0,>=2.4.0->chromadb) (1.17.0)
Requirement already satisfied: python-dateutil>=2.2 in /usr/local/lib/python3.11/dist-packages (from posthog<6.0.0,>=2.4.0->chromadb) (2.9.0.post0)
Collecting backoff>=1.10.0 (from posthog<6.0.0,>=2.4.0->chromadb)
  Downloading backoff-2.2.1-py3-none-any.whl.metadata (14 kB)
```

```
Requirement already satisfied: distro>=1.5.0 in /usr/local/lib/python3.11/dist-packages (from posthog<6.0.0,>=2.4.0->chromadb) (1.9.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests<3.0,>=2.7->posthog<6.0.0,>=2.4.0->chromadb) (3.4.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests<3.0,>=2.7->posthog<6.0.0,>=2.4.0->chromadb) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests<3.0,>=2.7->posthog<6.0.0,>=2.4.0->chromadb) (1.26.20)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests<3.0,>=2.7->posthog<6.0.0,>=2.4.0->chromadb) (2025.6.15)
Requirement already satisfied: packaging>=19.1 in /usr/local/lib/python3.11/dist-packages (from build>=1.0.3->chromadb) (24.2)
Requirement already satisfied: pyproject_hooks in /usr/local/lib/python3.11/dist-packages (from build>=1.0.3->chromadb) (1.2.0)
Requirement already satisfied: anyio in /usr/local/lib/python3.11/dist-packages (from httpx>=0.27.0->chromadb) (4.9.0)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx>=0.27.0->chromadb) (1.0.9)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx>=0.27.0->chromadb) (0.16.0)
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.19.0->chromadb) (25.3.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.19.0->chromadb) (2025.4.1)
Requirement already satisfied: referencing>=0.28.4 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.19.0->chromadb) (0.36.2)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.19.0->chromadb) (0.26.0)
Requirement already satisfied: google-auth>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from kubernetes>=28.1.0->chromadb) (2.38.0)
Requirement already satisfied: websocket-client!=0.40.0,!=0.41.*,!=0.42.*,>=0.32.0 in /usr/local/lib/python3.11/dist-packages (from kubernetes>=28.1.0->chromadb) (1.8.0)
Requirement already satisfied: requests-oauthlib in /usr/local/lib/python3.11/dist-packages (from kubernetes>=28.1.0->chromadb) (2.0.0)
Requirement already satisfied: oauthlib>=3.2.2 in /usr/local/lib/python3.11/dist-packages (from kubernetes>=28.1.0->chromadb) (3.3.1)
Collecting durationpy>=0.7 (from kubernetes>=28.1.0->chromadb)
  Downloading durationpy-0.10-py3-none-any.whl.metadata (340 bytes)
Requirement already satisfied: cachetools<6.0,>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from google-auth>=1.0.1->kubernetes>=28.1.0->chromadb) (5.5.2)
Requirement already satisfied: pyasn1-modules>=0.2.1 in /usr/local/lib/python3.11/dist-packages (from google-auth>=1.0.1->kubernetes>=28.1.0->chromadb) (0.4.2)
Requirement already satisfied: rsa<5,>=3.1.4 in /usr/local/lib/python3.11/dist-packages (from google-auth>=1.0.1->kubernetes>=28.1.0->chromadb) (4.9.1)
Requirement already satisfied: pyasn1>=0.1.3 in /usr/local/lib/python3.11/dist-packages (from rsa<5,>=3.1.4->google-auth>=1.0.1->kubernetes>=28.1.0->chromadb) (0.6.1)
Collecting coloredlogs (from onnxruntime>=1.14.1->chromadb)
  Downloading coloredlogs-15.0.1-py2.py3-none-any.whl.metadata (12 kB)
Requirement already satisfied: flatbuffers in /usr/local/lib/python3.11/dist-packages (from onnxruntime>=1.14.1->chromadb) (25.2.10)
Requirement already satisfied: protobuf in /usr/local/lib/python3.11/dist-packages (from onnxruntime>=1.14.1->chromadb) (5.29.5)
Requirement already satisfied: sympy in /usr/local/lib/python3.11/dist-packages (from onnxruntime>=1.14.1->chromadb) (1.13.1)
Requirement already satisfied: importlib-metadata<8.8.0,>=6.0 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-api>=1.2.0->chromadb) (8.7.0)
Requirement already satisfied: zipp>=3.20 in /usr/local/lib/python3.11/dist-packages (from importlib-metadata<8.8.0,>=6.0->opentelemetry-api>=1.2.0->chromadb) (3.23.0)
Requirement already satisfied: googleapis-common-protos~=1.52 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp-proto-grpc>=1.2.0->chromadb) (1.70
Requirement already satisfied: opentelemetry-exporter-otlp-proto-common==1.34.1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp-proto-grpc>=1.2.
Requirement already satisfied: opentelemetry-proto==1.34.1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp-proto-grpc>=1.2.0->chromadb) (1.34.1)
Requirement already satisfied: opentelemetry-semantic-conventions==0.55b1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-sdk>=1.2.0->chromadb) (0.55b1)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=1.9->chromadb) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=1.9->chromadb) (2.33.2)
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=1.9->chromadb) (0.4.1)
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.11/dist-packages (from rich>=10.11.0->chromadb) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.11/dist-packages (from rich>=10.11.0->chromadb) (2.19.2)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0->rich>=10.11.0->chromadb) (0.1.2)
Requirement already satisfied: huggingface-hub<1.0,>=0.16.4 in /usr/local/lib/python3.11/dist-packages (from tokenizers>=0.13.2->chromadb) (0.33.1)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.16.4->tokenizers>=0.13.2->chromadb) (3.18.0)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.16.4->tokenizers>=0.13.2->chromadb) (2025.3.2)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.16.4->tokenizers>=0.13.2->chromadb) (1.1.5)
Requirement already satisfied: click>=8.0.0 in /usr/local/lib/python3.11/dist-packages (from typer>=0.9.0->chromadb) (8.2.1)
Requirement already satisfied: shellingham>=1.3.0 in /usr/local/lib/python3.11/dist-packages (from typer>=0.9.0->chromadb) (1.5.4)
Collecting httptools>=0.5.0 (from uvicorn[standard]>=0.18.3->chromadb)
  Downloading httptools-0.6.4-cp311-cp311-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (3.6 kB)
Requirement already satisfied: python-dotenv>=0.13 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb) (1.1.1)
Requirement already satisfied: uvloop!=0.15.0,!=0.15.1,>=0.14.0 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb) (0.21.0)
Collecting watchfiles>=0.13 (from uvicorn[standard]>=0.18.3->chromadb)
  Downloading watchfiles-1.1.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.9 kB)
Requirement already satisfied: websockets>=10.4 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb) (15.0.1)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.11/dist-packages (from anyio->httpx>=0.27.0->chromadb) (1.3.1)
Collecting humanfriendly>=9.1 (from coloredlogs->onnxruntime>=1.14.1->chromadb)
  Downloading humanfriendly-10.0-py2.py3-none-any.whl.metadata (9.2 kB)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy->onnxruntime>=1.14.1->chromadb) (1.3.0)
Downloading chromadb-1.0.15-cp39-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (19.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 19.5/19.5 MB 162.5 MB/s eta 0:00:00
Downloading posthog-5.4.0-py3-none-any.whl (105 kB)
Downloading backoff-2.2.1-py3-none-any.whl (15 kB)
```

```
Downloading backoff-2.2.1-py3-none-any.whl (15 kB)
Downloading bcrypt-4.3.0-cp39-abi3-manylinux_2_34_x86_64.whl (284 kB)
Downloading kubernetes-33.1.0-py2.py3-none-any.whl (1.9 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.9/1.9 MB 101.9 MB/s eta 0:00:00
Downloading durationpy-0.10-py3-none-any.whl (3.9 kB)
Downloading mmh3-5.1.0-cp311-cp311-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (101 kB)
Downloading onnxruntime-1.22.0-cp311-cp311-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl (16.4 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.4/16.4 MB 160.9 MB/s eta 0:00:00
Downloading overrides-7.7.0-py3-none-any.whl (17 kB)
Downloading pybase64-1.4.1-cp311-cp311-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (71 kB)
Downloading httptools-0.6.4-cp311-cp311-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (459 kB)
Downloading watchfiles-1.1.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (453 kB)
Downloading coloredlogs-15.0.1-py2.py3-none-any.whl (46 kB)
Downloading humanfriendly-10.0-py2.py3-none-any.whl (86 kB)
Building wheels for collected packages: pypika
  Building wheel for pypika (pyproject.toml) ... done
  Created wheel for pypika: filename=pypika-0.48.9-py2.py3-none-any.whl size=53803 sha256=115c5bbf754e084301279d02275e9978578edc96966ec135c99f014e92450c62
  Stored in directory: /root/.cache/pip/wheels/a3/01/bd/4c40ceb9d5354160cb186dcc153360f4ab7eb23e2b24daf96d
Successfully built pypika
Installing collected packages: pypika, durationpy, pybase64, overrides, mmh3, humanfriendly, httptools, bcrypt, backoff, watchfiles, posthog, coloredlogs, onnxruntime, kuber
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 15/15 [chromadb]
Successfully installed backoff-2.2.1 bcrypt-4.3.0 chromadb-1.0.15 coloredlogs-15.0.1 durationpy-0.10 httptools-0.6.4 humanfriendly-10.0 kubernetes-33.1.0 mmh3-5.1.0 onnxrunt
Collecting llama-index-vector-stores-chroma
  Downloading llama_index_vector_stores_chroma-0.4.2-py3-none-any.whl.metadata (413 bytes)
Requirement already satisfied: chromadb>=0.5.17 in /usr/local/lib/python3.11/dist-packages (from llama-index-vector-stores-chroma) (1.0.15)
Requirement already satisfied: llama-index-core<0.13,>=0.12.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-vector-stores-chroma) (0.12.46)
Requirement already satisfied: aiohttp<4,>=3.8.6 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (3.11.15)
Requirement already satisfied: aiosqlite in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (0.21.0)
Requirement already satisfied: banks<3,>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (2.1.3)
Requirement already satisfied: dataclasses-json in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (0.6.7)
Requirement already satisfied: deprecated>=1.2.9.3 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (1.2.18
Requirement already satisfied: dirtyjson<2,>=1.0.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (1.0.8)
Requirement already satisfied: filetype<2,>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (1.2.0)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (2025.3.2)
Requirement already satisfied: httpx in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (0.28.1)
Requirement already satisfied: llama-index-workflows<2,>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chr
Requirement already satisfied: nest-asyncio<2,>=1.5.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (1.6
Requirement already satisfied: networkx>=3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (3.5)
Requirement already satisfied: nltk>3.8.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (3.9.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (2.0.2)
Requirement already satisfied: pillow>=9.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (11.2.1)
Requirement already satisfied: pydantic>=2.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (2.11.7)
Requirement already satisfied: pyyaml>=6.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (6.0.2)
Requirement already satisfied: requests>=2.31.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (2.32.3)
Requirement already satisfied: setuptools>=80.9.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (80.9.0)
Requirement already satisfied: sqlalchemy>=1.4.49 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13,>=0.12.0->llama-index-v
Requirement already satisfied: tenacity!=8.4.0,<10.0.0,>=8.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chr
Requirement already satisfied: tiktoken>=0.7.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (0.9.0)
Requirement already satisfied: tqdm<5,>=4.66.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (4.67.1)
Requirement already satisfied: typing-extensions>=4.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (4
Requirement already satisfied: typing-inspect>=0.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (0.9.
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (1.17.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-vector
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-ch
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-vector-store
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-vector-stc
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-vector-store
Requirement already satisfied: griffe in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (1.
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (3.
Requirement already satisfied: platformdirs in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chrom
```

```
Requirement already satisfied: llama-index-instrumentation>=0.1.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-workflows<2,>=1.0.1->llama-index-core<0.13,>=0
Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.11/dist-packages (from yarl<2.0,>=1.17.0->aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-v
Requirement already satisfied: build>=1.0.3 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.2.2.post1)
Requirement already satisfied: pybase64>=1.4.1 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.4.1)
Requirement already satisfied: uvicorn>=0.18.3 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb>=0.5.17->llama-index-vector-stores-chroma
Requirement already satisfied: posthog<6.0.0,>=2.4.0 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (5.4.0)
Requirement already satisfied: onnxruntime>=1.14.1 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.22.0)
Requirement already satisfied: opentelemetry-api>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.34.1)
Requirement already satisfied: opentelemetry-exporter-otlp-proto-grpc>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chr
Requirement already satisfied: opentelemetry-sdk>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.34.1)
Requirement already satisfied: tokenizers>=0.13.2 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (0.21.2)
Requirement already satisfied: pypika>=0.48.9 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (0.48.9)
Requirement already satisfied: overrides>=7.3.1 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (7.7.0)
Requirement already satisfied: importlib-resources in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (6.5.2)
Requirement already satisfied: grpcio>=1.58.0 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.68.0)
Requirement already satisfied: bcrypt>=4.0.1 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (4.3.0)
Requirement already satisfied: typer>=0.9.0 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (0.16.0)
Requirement already satisfied: kubernetes>=28.1.0 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (33.1.0)
Requirement already satisfied: mmh3>=4.0.1 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (5.1.0)
Requirement already satisfied: orjson>=3.9.12 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (3.10.18)
Requirement already satisfied: rich>=10.11.0 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (13.9.4)
Requirement already satisfied: jsonschema>=4.19.0 in /usr/local/lib/python3.11/dist-packages (from chromadb>=0.5.17->llama-index-vector-stores-chroma) (4.24.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from posthog<6.0.0,>=2.4.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.17.0)
Requirement already satisfied: python-dateutil>=2.2 in /usr/local/lib/python3.11/dist-packages (from posthog<6.0.0,>=2.4.0->chromadb>=0.5.17->llama-index-vector-stores-chrom
Requirement already satisfied: backoff>=1.10.0 in /usr/local/lib/python3.11/dist-packages (from posthog<6.0.0,>=2.4.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (2
Requirement already satisfied: distro>=1.5.0 in /usr/local/lib/python3.11/dist-packages (from posthog<6.0.0,>=2.4.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.9
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.0->llama-index-vector
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-store
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-store
Requirement already satisfied: packaging>=19.1 in /usr/local/lib/python3.11/dist-packages (from build>=1.0.3->chromadb>=0.5.17->llama-index-vector-stores-chroma) (24.2)
Requirement already satisfied: pyproject_hooks in /usr/local/lib/python3.11/dist-packages (from build>=1.0.3->chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.2.0)
Requirement already satisfied: anyio in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (4.9.0)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (1.0.9
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chr
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.19.0->chromadb>=0.5.17->llama-index-vector
Requirement already satisfied: referencing>=0.28.4 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.19.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.19.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (0.26.
Requirement already satisfied: google-auth>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from kubernetes>=28.1.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (2
Requirement already satisfied: websocket-client!=0.40.0,!=0.41.*,!=0.42.*,>=0.32.0 in /usr/local/lib/python3.11/dist-packages (from kubernetes>=28.1.0->chromadb>=0.5.17->lla
Requirement already satisfied: requests-oauthlib in /usr/local/lib/python3.11/dist-packages (from kubernetes>=28.1.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (2.
Requirement already satisfied: oauthlib>=3.2.2 in /usr/local/lib/python3.11/dist-packages (from kubernetes>=28.1.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (3.3.
Requirement already satisfied: durationpy>=0.7 in /usr/local/lib/python3.11/dist-packages (from kubernetes>=28.1.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (0.10
Requirement already satisfied: cachetools<6.0,>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from google-auth>=1.0.1->kubernetes>=28.1.0->chromadb>=0.5.17->llama-index-
Requirement already satisfied: pyasn1-modules>=0.2.1 in /usr/local/lib/python3.11/dist-packages (from google-auth>=1.0.1->kubernetes>=28.1.0->chromadb>=0.5.17->llama-index-v
Requirement already satisfied: rsa<5,>=3.1.4 in /usr/local/lib/python3.11/dist-packages (from google-auth>=1.0.1->kubernetes>=28.1.0->chromadb>=0.5.17->llama-index-vector-st
Requirement already satisfied: pyasn1>=0.1.3 in /usr/local/lib/python3.11/dist-packages (from rsa<5,>=3.1.4->google-auth>=1.0.1->kubernetes>=28.1.0->chromadb>=0.5.17->llama-
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (8.2.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma) (1.5.1)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma)
Requirement already satisfied: coloredlogs in /usr/local/lib/python3.11/dist-packages (from onnxruntime>=1.14.1->chromadb>=0.5.17->llama-index-vector-stores-chroma) (15.0.1)
Requirement already satisfied: flatbuffers in /usr/local/lib/python3.11/dist-packages (from onnxruntime>=1.14.1->chromadb>=0.5.17->llama-index-vector-stores-chroma) (25.2.10
Requirement already satisfied: protobuf in /usr/local/lib/python3.11/dist-packages (from onnxruntime>=1.14.1->chromadb>=0.5.17->llama-index-vector-stores-chroma) (5.29.5)
Requirement already satisfied: sympy in /usr/local/lib/python3.11/dist-packages (from onnxruntime>=1.14.1->chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.13.1)
Requirement already satisfied: importlib-metadata<8.8.0,>=6.0 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-api>=1.2.0->chromadb>=0.5.17->llama-index-vector
Requirement already satisfied: zipp>=3.20 in /usr/local/lib/python3.11/dist-packages (from importlib-metadata<8.8.0,>=6.0->opentelemetry-api>=1.2.0->chromadb>=0.5.17->llama-
Requirement already satisfied: googleapis-common-protos~=1.52 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp-proto-grpc>=1.2.0->chromadb>=0.5.1
Requirement already satisfied: opentelemetry-exporter-otlp-proto-common==1.34.1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp-proto-grpc>=1.2.
Requirement already satisfied: opentelemetry-proto==1.34.1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp-proto-grpc>=1.2.0->chromadb>=0.5.17->
Requirement already satisfied: opentelemetry-semantic-conventions==0.55b1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-sdk>=1.2.0->chromadb>=0.5.17->llama-
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-st
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-stc
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-
```

```
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.11/dist-packages (from rich>=10.11.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (3.0
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.11/dist-packages (from rich>=10.11.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (2
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0->rich>=10.11.0->chromadb>=0.5.17->llama-index-vector-stores-
Requirement already satisfied: greenlet>=1 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy>=1.4.49->sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13,>=0.12.0->
Requirement already satisfied: huggingface-hub<1.0,>=0.16.4 in /usr/local/lib/python3.11/dist-packages (from tokenizers>=0.13.2->chromadb>=0.5.17->llama-index-vector-stores-
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.16.4->tokenizers>=0.13.2->chromadb>=0.5.17->llama-index-vect
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.16.4->tokenizers>=0.13.2->chromadb>=0.5.17->llam
Requirement already satisfied: shellingham>=1.3.0 in /usr/local/lib/python3.11/dist-packages (from typer>=0.9.0->chromadb>=0.5.17->llama-index-vector-stores-chroma) (1.5.4)
Requirement already satisfied: mypy-extensions>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from typing-inspect>=0.8.0->llama-index-core<0.13,>=0.12.0->llama-index-vec
Requirement already satisfied: httptools>=0.5.0 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb>=0.5.17->llama-index-vector-stores-chrom
Requirement already satisfied: python-dotenv>=0.13 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb>=0.5.17->llama-index-vector-stores-ch
Requirement already satisfied: uvloop!=0.15.0,!=0.15.1,>=0.14.0 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb>=0.5.17->llama-index-vec
Requirement already satisfied: watchfiles>=0.13 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb>=0.5.17->llama-index-vector-stores-chrom
Requirement already satisfied: websockets>=10.4 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.18.3->chromadb>=0.5.17->llama-index-vector-stores-chrom
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.11/dist-packages (from anyio->httpx->llama-index-core<0.13,>=0.12.0->llama-index-vector-stores-chroma)
Requirement already satisfied: humanfriendly>=9.1 in /usr/local/lib/python3.11/dist-packages (from coloredlogs->onnxruntime>=1.14.1->chromadb>=0.5.17->llama-index-vector-stc
Requirement already satisfied: marshmallow<4.0.0,>=3.18.0 in /usr/local/lib/python3.11/dist-packages (from dataclasses-json->llama-index-core<0.13,>=0.12.0->llama-index-vect
Requirement already satisfied: colorama>=0.4 in /usr/local/lib/python3.11/dist-packages (from griffe->banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-stc
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-vector-s
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy->onnxruntime>=1.14.1->chromadb>=0.5.17->llama-index-vector-stores-ch
Downloading llama_index_vector_stores_chroma-0.4.2-py3-none-any.whl (6.0 kB)
Installing collected packages: llama-index-vector-stores-chroma
Successfully installed llama-index-vector-stores-chroma-0.4.2
Requirement already satisfied: openai in /usr/local/lib/python3.11/dist-packages (1.93.0)
Requirement already satisfied: anyio<5,>=3.5.0 in /usr/local/lib/python3.11/dist-packages (from openai) (4.9.0)
Requirement already satisfied: distro<2,>=1.7.0 in /usr/local/lib/python3.11/dist-packages (from openai) (1.9.0)
Requirement already satisfied: httpx<1,>=0.23.0 in /usr/local/lib/python3.11/dist-packages (from openai) (0.28.1)
Requirement already satisfied: jiter<1,>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from openai) (0.10.0)
Requirement already satisfied: pydantic<3,>=1.9.0 in /usr/local/lib/python3.11/dist-packages (from openai) (2.11.7)
Requirement already satisfied: sniffio in /usr/local/lib/python3.11/dist-packages (from openai) (1.3.1)
Requirement already satisfied: tqdm>4 in /usr/local/lib/python3.11/dist-packages (from openai) (4.67.1)
Requirement already satisfied: typing-extensions<5,>=4.11 in /usr/local/lib/python3.11/dist-packages (from openai) (4.14.0)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.11/dist-packages (from anyio<5,>=3.5.0->openai) (3.10)
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-packages (from httpx<1,>=0.23.0->openai) (2025.6.15)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx<1,>=0.23.0->openai) (1.0.9)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx<1,>=0.23.0->openai) (0.16.0)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<3,>=1.9.0->openai) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic<3,>=1.9.0->openai) (2.33.2)
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<3,>=1.9.0->openai) (0.4.1)
Requirement already satisfied: openai in /usr/local/lib/python3.11/dist-packages (1.93.0)
Requirement already satisfied: anyio<5,>=3.5.0 in /usr/local/lib/python3.11/dist-packages (from openai) (4.9.0)
Requirement already satisfied: distro<2,>=1.7.0 in /usr/local/lib/python3.11/dist-packages (from openai) (1.9.0)
Requirement already satisfied: httpx<1,>=0.23.0 in /usr/local/lib/python3.11/dist-packages (from openai) (0.28.1)
Requirement already satisfied: jiter<1,>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from openai) (0.10.0)
Requirement already satisfied: pydantic<3,>=1.9.0 in /usr/local/lib/python3.11/dist-packages (from openai) (2.11.7)
Requirement already satisfied: sniffio in /usr/local/lib/python3.11/dist-packages (from openai) (1.3.1)
Requirement already satisfied: tqdm>4 in /usr/local/lib/python3.11/dist-packages (from openai) (4.67.1)
Requirement already satisfied: typing-extensions<5,>=4.11 in /usr/local/lib/python3.11/dist-packages (from openai) (4.14.0)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.11/dist-packages (from anyio<5,>=3.5.0->openai) (3.10)
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-packages (from httpx<1,>=0.23.0->openai) (2025.6.15)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx<1,>=0.23.0->openai) (1.0.9)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx<1,>=0.23.0->openai) (0.16.0)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<3,>=1.9.0->openai) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic<3,>=1.9.0->openai) (2.33.2)
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<3,>=1.9.0->openai) (0.4.1)
Requirement already satisfied: sentence-transformers in /usr/local/lib/python3.11/dist-packages (4.1.0)
Requirement already satisfied: transformers<5.0.0,>=4.41.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (4.53.0)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (4.67.1)
Requirement already satisfied: torch>=1.11.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (2.6.0+cu124)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (1.6.1)
Requirement already satisfied: scipy in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (1.15.3)
```

```
Requirement already satisfied: huggingface-hub>=0.20.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (0.33.1)
Requirement already satisfied: Pillow in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (11.2.1)
Requirement already satisfied: typing_extensions>=4.5.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (4.14.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers) (3.18.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers) (0.21.2)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers) (0.5.3)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (2025.3.2)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (1.1.5)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (3.5)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (3.1.6)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch>=1.11.0->sentence-transformers) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch>=1.11.0->sentence-transformers) (3.0.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers<5.0.0,>=4.41.0->sentence-transformers) (3.4.2
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers<5.0.0,>=4.41.0->sentence-transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers<5.0.0,>=4.41.0->sentence-transformers) (1.26.20)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers<5.0.0,>=4.41.0->sentence-transformers) (2025.6.15)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->sentence-transformers) (1.5.1)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->sentence-transformers) (3.6.0)
Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl (363.4 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 56.6 MB/s eta 0:00:00
Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (13.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 60.0 MB/s eta 0:00:00
Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (24.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 65.5 MB/s eta 0:00:00
Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 31.5 MB/s eta 0:00:00
Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl (664.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 22.9 MB/s eta 0:00:00
Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl (211.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 33.4 MB/s eta 0:00:00
Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl (56.3 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 23.6 MB/s eta 0:00:00
```

```
Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl (127.9 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 31.2 MB/s eta 0:00:00
Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl (207.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 46.7 MB/s eta 0:00:00
Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (21.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 63.0 MB/s eta 0:00:00
Installing collected packages: nvidia-nvjitlink-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.3.61
    Uninstalling nvidia-cufft-cu12-11.2.3.61:
      Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
  Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
    Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
    Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
    Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
  Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.5.3.2
    Uninstalling nvidia-cublas-cu12-12.5.3.2:
      Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.5.1.3
    Uninstalling nvidia-cusparse-cu12-12.5.1.3:
      Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.3.0.75
    Uninstalling nvidia-cudnn-cu12-9.3.0.75:
      Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.3.83
    Uninstalling nvidia-cusolver-cu12-11.6.3.83:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10/10 [nvidia-cusolver-cu12]
Successfully installed nvidia-cublas-cu12-12.4.5.8 nvidia-cuda-cupti-cu12-12.4.127 nvidia-cuda-nvrtc-cu12-12.4.127 nvidia-cuda-runtime-cu12-12.4.127 nvidia-cudnn-cu12-9.1.0.
Requirement already satisfied: jina in /usr/local/lib/python3.11/dist-packages (3.34.0)
Requirement already satisfied: opentelemetry-instrumentation-grpc>=0.35b0 in /usr/local/lib/python3.11/dist-packages (from jina) (0.55b1)
Requirement already satisfied: docarray>=0.16.4 in /usr/local/lib/python3.11/dist-packages (from jina) (0.41.0)
Requirement already satisfied: pathspec in /usr/local/lib/python3.11/dist-packages (from jina) (0.12.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from jina) (2.0.2)
Requirement already satisfied: opentelemetry-exporter-prometheus>=0.33b0 in /usr/local/lib/python3.11/dist-packages (from jina) (0.55b1)
Requirement already satisfied: websockets in /usr/local/lib/python3.11/dist-packages (from jina) (15.0.1)
Requirement already satisfied: aiofiles in /usr/local/lib/python3.11/dist-packages (from jina) (24.1.0)
Requirement already satisfied: jcloud>=0.0.35 in /usr/local/lib/python3.11/dist-packages (from jina) (0.3)
Requirement already satisfied: opentelemetry-instrumentation-fastapi>=0.33b0 in /usr/local/lib/python3.11/dist-packages (from jina) (0.55b1)
Requirement already satisfied: grpcio<=1.68.0,>=1.46.0 in /usr/local/lib/python3.11/dist-packages (from jina) (1.68.0)
Requirement already satisfied: opentelemetry-exporter-otlp-proto-grpc>=1.13.0 in /usr/local/lib/python3.11/dist-packages (from jina) (1.34.1)
Requirement already satisfied: grpcio-reflection<1.68.0,>=1.46.0 in /usr/local/lib/python3.11/dist-packages (from jina) (1.68.0)
```

```
Requirement already satisfied: grpcio-reflection<=1.68.0,>=1.46.0 in /usr/local/lib/python3.11/dist-packages (from jina) (1.68.0)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from jina) (2.32.3)
Requirement already satisfied: pydantic<3.0.0 in /usr/local/lib/python3.11/dist-packages (from jina) (2.11.7)
Requirement already satisfied: opentelemetry-exporter-otlp>=1.12.0 in /usr/local/lib/python3.11/dist-packages (from jina) (1.34.1)
Requirement already satisfied: fastapi>=0.76.0 in /usr/local/lib/python3.11/dist-packages (from jina) (0.115.14)
Requirement already satisfied: opentelemetry-instrumentation-aiohttp-client>=0.33b0 in /usr/local/lib/python3.11/dist-packages (from jina) (0.55b1)
Requirement already satisfied: docker in /usr/local/lib/python3.11/dist-packages (from jina) (7.1.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from jina) (3.18.0)
Requirement already satisfied: protobuf>=3.19.0 in /usr/local/lib/python3.11/dist-packages (from jina) (5.29.5)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from jina) (24.2)
Requirement already satisfied: prometheus_client>=0.12.0 in /usr/local/lib/python3.11/dist-packages (from jina) (0.22.1)
Requirement already satisfied: jina-hubble-sdk>=0.30.4 in /usr/local/lib/python3.11/dist-packages (from jina) (0.39.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (from jina) (3.11.15)
Requirement already satisfied: opentelemetry-api>=1.12.0 in /usr/local/lib/python3.11/dist-packages (from jina) (1.34.1)
Requirement already satisfied: uvloop in /usr/local/lib/python3.11/dist-packages (from jina) (0.21.0)
Requirement already satisfied: python-multipart in /usr/local/lib/python3.11/dist-packages (from jina) (0.0.20)
Requirement already satisfied: opentelemetry-sdk>=1.14.0 in /usr/local/lib/python3.11/dist-packages (from jina) (1.34.1)
Requirement already satisfied: grpcio-health-checking<=1.68.0,>=1.46.0 in /usr/local/lib/python3.11/dist-packages (from jina) (1.68.0)
Requirement already satisfied: pyyaml>=5.3.1 in /usr/local/lib/python3.11/dist-packages (from jina) (6.0.2)
Requirement already satisfied: urllib3<2.0.0,>=1.25.9 in /usr/local/lib/python3.11/dist-packages (from jina) (1.26.20)
Requirement already satisfied: uvicorn<=0.23.1 in /usr/local/lib/python3.11/dist-packages (from jina) (0.23.1)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0->jina) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0->jina) (2.33.2)
Requirement already satisfied: typing-extensions>=4.12.2 in /usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0->jina) (4.14.0)
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0->jina) (0.4.1)
Requirement already satisfied: click>=7.0 in /usr/local/lib/python3.11/dist-packages (from uvicorn<=0.23.1->jina) (8.2.1)
Requirement already satisfied: h11>=0.8 in /usr/local/lib/python3.11/dist-packages (from uvicorn<=0.23.1->jina) (0.16.0)
Requirement already satisfied: orjson>=3.8.2 in /usr/local/lib/python3.11/dist-packages (from docarray>=0.16.4->jina) (3.10.18)
Requirement already satisfied: rich>=13.1.0 in /usr/local/lib/python3.11/dist-packages (from docarray>=0.16.4->jina) (13.9.4)
Requirement already satisfied: types-requests>=2.28.11.6 in /usr/local/lib/python3.11/dist-packages (from docarray>=0.16.4->jina) (2.31.0.6)
Requirement already satisfied: typing-inspect>=0.8.0 in /usr/local/lib/python3.11/dist-packages (from docarray>=0.16.4->jina) (0.9.0)
Requirement already satisfied: starlette<0.47.0,>=0.40.0 in /usr/local/lib/python3.11/dist-packages (from fastapi>=0.76.0->jina) (0.46.2)
Requirement already satisfied: anyio<5,>=3.6.2 in /usr/local/lib/python3.11/dist-packages (from starlette<0.47.0,>=0.40.0->fastapi>=0.76.0->jina) (4.9.0)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.11/dist-packages (from anyio<5,>=3.6.2->starlette<0.47.0,>=0.40.0->fastapi>=0.76.0->jina) (3.10)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.11/dist-packages (from anyio<5,>=3.6.2->starlette<0.47.0,>=0.40.0->fastapi>=0.76.0->jina) (1.3.1)
Requirement already satisfied: python-dotenv in /usr/local/lib/python3.11/dist-packages (from jcloud>=0.0.35->jina) (1.1.1)
Requirement already satisfied: python-dateutil in /usr/local/lib/python3.11/dist-packages (from jcloud>=0.0.35->jina) (2.9.0.post0)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->jina) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp->jina) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->jina) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp->jina) (1.7.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp->jina) (6.6.3)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->jina) (0.3.2)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->jina) (1.20.1)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.11/dist-packages (from jina-hubble-sdk>=0.30.4->jina) (8.7.0)
Requirement already satisfied: python-jose in /usr/local/lib/python3.11/dist-packages (from jina-hubble-sdk>=0.30.4->jina) (3.5.0)
Requirement already satisfied: zipp>=3.20 in /usr/local/lib/python3.11/dist-packages (from importlib-metadata->jina-hubble-sdk>=0.30.4->jina) (3.23.0)
Requirement already satisfied: opentelemetry-exporter-otlp-proto-http==1.34.1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp>=1.12.0->jina) (1.
Requirement already satisfied: googleapis-common-protos~=1.52 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp-proto-grpc>=1.13.0->jina) (1.70.0)
Requirement already satisfied: opentelemetry-exporter-otlp-proto-common==1.34.1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp-proto-grpc>=1.13
Requirement already satisfied: opentelemetry-proto==1.34.1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-exporter-otlp-proto-grpc>=1.13.0->jina) (1.34.1)
Requirement already satisfied: opentelemetry-semantic-conventions==0.55b1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-sdk>=1.14.0->jina) (0.55b1)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->jina) (3.4.2)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->jina) (2025.6.15)
Requirement already satisfied: opentelemetry-instrumentation==0.55b1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-instrumentation-aiohttp-client>=0.33b0->j
Requirement already satisfied: opentelemetry-util-http==0.55b1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-instrumentation-aiohttp-client>=0.33b0->jina) (
Requirement already satisfied: wrapt<2.0.0,>=1.0.0 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-instrumentation-aiohttp-client>=0.33b0->jina) (1.17.2)
Requirement already satisfied: opentelemetry-instrumentation-asgi==0.55b1 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-instrumentation-fastapi>=0.33b0->jin
Requirement already satisfied: asgiref~=3.0 in /usr/local/lib/python3.11/dist-packages (from opentelemetry-instrumentation-asgi==0.55b1->opentelemetry-instrumentation-fastap
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.11/dist-packages (from rich>=13.1.0->docarray>=0.16.4->jina) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.11/dist-packages (from rich>=13.1.0->docarray>=0.16.4->jina) (2.19.2)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0->rich>=13.1.0->docarray>=0.16.4->jina) (0.1.2)
```

```
Requirement already satisfied: types-urllib3 in /usr/local/lib/python3.11/dist-packages (from types-requests>=2.28.11.6->docarray>=0.16.4->jina) (1.26.25.14)
Requirement already satisfied: mypy-extensions>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from typing-inspect>=0.8.0->docarray>=0.16.4->jina) (1.1.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil->jcloud>=0.0.35->jina) (1.17.0)
Requirement already satisfied: ecdsa!=0.15 in /usr/local/lib/python3.11/dist-packages (from python-jose->jina-hubble-sdk>=0.30.4->jina) (0.19.1)
Requirement already satisfied: rsa!=4.1.1,!=4.4,<5.0,>=4.0 in /usr/local/lib/python3.11/dist-packages (from python-jose->jina-hubble-sdk>=0.30.4->jina) (4.9.1)
Requirement already satisfied: pyasn1>=0.5.0 in /usr/local/lib/python3.11/dist-packages (from python-jose->jina-hubble-sdk>=0.30.4->jina) (0.6.1)
Collecting llama-index-embeddings-huggingface
  Downloading llama_index_embeddings_huggingface-0.5.5-py3-none-any.whl.metadata (458 bytes)
Requirement already satisfied: huggingface-hub>=0.19.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub[inference]>=0.19.0->llama-index-embeddings-huggingfac
Requirement already satisfied: llama-index-core<0.13,>=0.12.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-embeddings-huggingface) (0.12.46)
Requirement already satisfied: sentence-transformers>=2.6.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-embeddings-huggingface) (4.1.0)
Requirement already satisfied: aiohttp<4,>=3.8.6 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (3.11.1
Requirement already satisfied: aiosqlite in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (0.21.0)
Requirement already satisfied: banks<3,>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (2.1.3)
Requirement already satisfied: dataclasses-json in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (0.6.7)
Requirement already satisfied: deprecated>=1.2.9.3 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (1.2.
Requirement already satisfied: dirtyjson<2,>=1.0.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (1.0.
Requirement already satisfied: filetype<2,>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (1.2.0
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (2025.3.
Requirement already satisfied: httpx in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (0.28.1)
Requirement already satisfied: llama-index-workflows<2,>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggin
Requirement already satisfied: nest-asyncio<2,>=1.5.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (1
Requirement already satisfied: networkx>=3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (3.5)
Requirement already satisfied: nltk>3.8.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (3.9.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (2.0.2)
Requirement already satisfied: pillow>=9.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (11.2.1)
Requirement already satisfied: pydantic>=2.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (2.11.7)
Requirement already satisfied: pyyaml>=6.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (6.0.2)
Requirement already satisfied: requests>=2.31.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (2.32.3)
Requirement already satisfied: setuptools>=80.9.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (80.9.
Requirement already satisfied: sqlalchemy>=1.4.49 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13,>=0.12.0->llama-index-e
Requirement already satisfied: tenacity!=8.4.0,<10.0.0,>=8.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggin
Requirement already satisfied: tiktoken>=0.7.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (0.9.0)
Requirement already satisfied: tqdm<5,>=4.66.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (4.67.1)
Requirement already satisfied: typing-extensions>=4.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface)
Requirement already satisfied: typing-inspect>=0.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (0.
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (1.17.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-embedd
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-hu
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggi
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-h
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-embeddings
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-hu
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-h
Requirement already satisfied: griffe in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (
Requirement already satisfied: platformdirs in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingf
Requirement already satisfied: llama-index-instrumentation>=0.1.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-workflows<2,>=1.0.1->llama-index-core<0.13,>=0
Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.11/dist-packages (from yarl<2.0,>=1.17.0->aiohttp<4,>=3.8.6->llama-index-core<0.13,>=0.12.0->llama-index-e
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.19.0->huggingface-hub[inference]>=0.19.0->llama-index-embeddings-
Requirement already satisfied: packaging>=20.9 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.19.0->huggingface-hub[inference]>=0.19.0->llama-index-embe
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.19.0->huggingface-hub[inference]>=0.19.0->llama-index
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (8.2.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (1.5.1
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingfac
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.0->llama-index-embedding
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.0->llama-index-embeddings
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13,>=0.12.0->llama-index-embeddi
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.0->llama-index-embedd
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-h
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-h
Requirement already satisfied: transformers<5.0.0,>=4.41.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=2.6.1->llama-index-embeddings-huggingface)
```

```
Requirement already satisfied: transformers<5.0.0,>=4.41.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=2.6.1->llama-index-embeddings-huggingface) (4.51.3)
Requirement already satisfied: torch>=1.11.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=2.6.1->llama-index-embeddings-huggingface) (2.6.0+cu124)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=2.6.1->llama-index-embeddings-huggingface) (1.6.1)
Requirement already satisfied: scipy in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=2.6.1->llama-index-embeddings-huggingface) (1.15.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers>=2.6.1->llama-index
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->sentence-transformers>=2.6.1->llama-index-emb
Requirement already satisfied: greenlet>=1 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy>=1.4.49->sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13,>=0.12.0->
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-emb
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-e
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-emb
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embeddin
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embeddi
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embeddin
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embed
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embed
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-emb
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embedd
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embeddings-
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embedding
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embe
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embeddings-huggingface
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-embeddings-huggingface
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch>=1.11.0->sentence-transformers>=2.6.1->llama-index-em
Requirement already satisfied: mypy-extensions>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from typing-inspect>=0.8.0->llama-index-core<0.13,>=0.12.0->llama-index-emb
Requirement already satisfied: marshmallow<4.0.0,>=3.18.0 in /usr/local/lib/python3.11/dist-packages (from dataclasses-json->llama-index-core<0.13,>=0.12.0->llama-index-embe
Requirement already satisfied: colorama>=0.4 in /usr/local/lib/python3.11/dist-packages (from griffe->banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-embeddings
Requirement already satisfied: anyio in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (4.9.0)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface) (1.0
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggin
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.11/dist-packages (from anyio->httpx->llama-index-core<0.13,>=0.12.0->llama-index-embeddings-huggingface
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->banks<3,>=2.0.0->llama-index-core<0.13,>=0.12.0->llama-index-embeddin
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->sentence-transformers>=2.6.1->llama-index-embeddings-huggi
Downloading llama_index_embeddings_huggingface-0.5.5-py3-none-any.whl (8.9 kB)
Installing collected packages: llama-index-embeddings-huggingface
Successfully installed llama-index-embeddings-huggingface-0.5.5
Collecting llama-index-embeddings-jinaai
  Downloading llama_index_embeddings_jinaai-0.4.0-py3-none-any.whl.metadata (652 bytes)
Requirement already satisfied: llama-index-core<0.13.0,>=0.12.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-embeddings-jinaai) (0.12.46)
Requirement already satisfied: aiohttp<4,>=3.8.6 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (3.11.15)
Requirement already satisfied: aiosqlite in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (0.21.0)
Requirement already satisfied: banks<3,>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (2.1.3)
Requirement already satisfied: dataclasses-json in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (0.6.7)
Requirement already satisfied: deprecated>=1.2.9.3 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (1.2.18)
Requirement already satisfied: dirtyjson<2,>=1.0.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (1.0.8)
Requirement already satisfied: filetype<2,>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (1.2.0)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (2025.3.2)
Requirement already satisfied: httpx in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (0.28.1)
Requirement already satisfied: llama-index-workflows<2,>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jina
Requirement already satisfied: nest-asyncio<2,>=1.5.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (1.6.
Requirement already satisfied: networkx>=3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (3.5)
Requirement already satisfied: nltk>3.8.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (3.9.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (2.0.2)
Requirement already satisfied: pillow>=9.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (11.2.1)
Requirement already satisfied: pydantic>=2.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (2.11.7)
Requirement already satisfied: pyyaml>=6.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (6.0.2)
Requirement already satisfied: requests>=2.31.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (2.32.3)
Requirement already satisfied: setuptools>=80.9.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (80.9.0)
Requirement already satisfied: sqlalchemy>=1.4.49 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13.0,>=0.12.0->llama-index
Requirement already satisfied: tenacity!=8.4.0,<10.0.0,>=8.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jina
Requirement already satisfied: tiktoken>=0.7.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (0.9.0)
Requirement already satisfied: tqdm<5,>=4.66.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (4.67.1)
Requirement already satisfied: typing-extensions>=4.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (4.
```

```
Requirement already satisfied: typing-inspect>=0.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (0.9.0
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (1.17.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-embe
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jin
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddin
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings
Requirement already satisfied: griffe in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (1.7
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (3.1
Requirement already satisfied: platformdirs in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai
Requirement already satisfied: llama-index-instrumentation>=0.1.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-workflows<2,>=1.0.1->llama-index-core<0.13.0,>
Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.11/dist-packages (from yarl<2.0,>=1.17.0->aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (8.2.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (1.5.1)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddi
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddin
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embed
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embe
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings
Requirement already satisfied: greenlet>=1 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy>=1.4.49->sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13.0,>=0.12.0
Requirement already satisfied: mypy-extensions>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from typing-inspect>=0.8.0->llama-index-core<0.13.0,>=0.12.0->llama-index-e
Requirement already satisfied: marshmallow<4.0.0,>=3.18.0 in /usr/local/lib/python3.11/dist-packages (from dataclasses-json->llama-index-core<0.13.0,>=0.12.0->llama-index-em
Requirement already satisfied: packaging>=17.0 in /usr/local/lib/python3.11/dist-packages (from marshmallow<4.0.0,>=3.18.0->dataclasses-json->llama-index-core<0.13.0,>=0.12.
Requirement already satisfied: colorama>=0.4 in /usr/local/lib/python3.11/dist-packages (from griffe->banks<3,>=2.0.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddin
Requirement already satisfied: anyio in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (4.9.0)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (1.0.9)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jina
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.11/dist-packages (from anyio->httpx->llama-index-core<0.13.0,>=0.12.0->llama-index-embeddings-jinaai) (
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->banks<3,>=2.0.0->llama-index-core<0.13.0,>=0.12.0->llama-index-embedd
Downloading llama_index_embeddings_jinaai-0.4.0-py3-none-any.whl (3.9 kB)
Installing collected packages: llama-index-embeddings-jinaai
Successfully installed llama-index-embeddings-jinaai-0.4.0
```

```
Start coding or generate with AI.
```

```
!ls
```

```
 api_credentials.yml                                sample_data
'Final Policy document_LICs New Jeevan Shanti.pdf'
```

## Move Your PDF into a Folder

```
import shutil

# Move the PDF into a temp folder
!mkdir -p /content/lic_pdfs
shutil.copy("/content/Final Policy document_LICs New Jeevan Shanti.pdf", "/content/lic_pdfs/")
```

```
'/content/lic_pdfs/Final Policy document_LICs New Jeevan Shanti.pdf'
```

## Load with SimpleDirectoryReader

```
from llama_index.core import SimpleDirectoryReader

# Load all PDFs in the folder
documents = SimpleDirectoryReader(input_dir="/content/lic_pdfs").load_data()

print(f"Loaded {len(documents)} document(s).")
print("Sample content:\n", documents[0].text[:1000])
```

```
Loaded 21 document(s).
Sample content:
  LIC’s New Jeevan Shanti (UIN: 512N338V05)      Page 1 of 21
                                    LIFE INSURANCE CORPORATION OF INDIA
(Established by the Life Insurance Corporation Act, 1956)
Registration Number: 512

                                            LIC’S NEW JEEVAN SHANTI (UIN: 512N338V05)
(A Non-Linked, Non-Participating, Individual, Single Premium, Deferred Annuity Plan)


PART – A


Ref: NB       (Address and e-mail id of Branch Office):



Dear Policyholder,         Date:
Re: Your Policy No. _____

We have pleasure in forwarding herewith the above policy document comprising of Part A to Part G
which please find in order.
```

```
We would also like to draw your kind attention to the information mentioned in the Schedule of the
Policy and the benefits available under the Policy.

Some of our Plans have certain options available under them. It is important that the o
```

```
Start coding or generate with AI.
```

Double-click (or enter) to edit

## Vector Store Creation

```
import chromadb
from llama_index.vector_stores.chroma import ChromaVectorStore
from llama_index.core import VectorStoreIndex, StorageContext

# Local persistent directory
persist_dir = "/content/chroma_lic_index"

# Initialize Chroma in local (embedded) mode
chroma_client = chromadb.PersistentClient(path=persist_dir)

# Create or get a collection
chroma_collection = chroma_client.get_or_create_collection("lic_policy")

# Wrap into LlamaIndex ChromaVectorStore
vector_store = ChromaVectorStore(chroma_collection=chroma_collection)

# Create storage context
storage_context = StorageContext.from_defaults(vector_store=vector_store)
```

```
index = VectorStoreIndex.from_documents(documents, storage_context=storage_context)
```

```
from llama_index.core.query_engine import RetrieverQueryEngine

# Create a retriever from the index
retriever = index.as_retriever(similarity_top_k=5)

# Initialize the query engine with the retriever
query_engine = RetrieverQueryEngine.from_args(retriever)

response = query_engine.query("What is the surrender value condition for LIC New Jeevan Shanti policy?")
print("Answer:\n", response)
```

```
Answer:
 The Surrender Value payable for LIC's New Jeevan Shanti policy is determined as the higher of the Guaranteed Surrender Value or the Special Surrender Value. The Guaranteed Su
```

```
response = query_engine.query("Provide the formula for 'Additional Benefit on Death")
print("Answer:\n", response)
```

```
Answer:
 The formula for Additional Benefit on Death is: Additional Benefit on Death per month = (Purchase Price * Annuity rate p.a. payable monthly) / 12.
```

## Index the same document 3 times, each with a different embedding model:

- OpenAI (text-embedding-3-small)
- BGE (BAAI/bge-small-en-v1.5)
- Jina (jinaai/jina-embeddings-v2-base-en)

Store each in a separate ChromaDB collection

Allow querying each to compare responses

```
Start coding or generate with AI.
```

```
Start coding or generate with AI.
```

## Initialize All 3 Embedding Models

```
from llama_index.embeddings.openai import OpenAIEmbedding
from llama_index.embeddings.huggingface import HuggingFaceEmbedding
#from llama_index.embeddings.jinaai import JinaEmbedding

openai_embed = OpenAIEmbedding(model="text-embedding-3-small")
bge_embed = HuggingFaceEmbedding(model_name="BAAI/bge-small-en-v1.5")
#jina_embed = JinaEmbedding(model_name="jinaai/jina-embeddings-v2-base-en")
```

```
UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in your Google Colab and restart your
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets. (raised from /usr/local/lib/python3.11/dist-packages/huggingface_hub/uti
```

```
modules.json: 100%  349/349 [00:00<00:00, 34.7kB/s]
config_sentence_transformers.json: 100%  124/124 [00:00<00:00, 13.3kB/s]
README.md: 94.8k/? [00:00<00:00, 8.28MB/s]
sentence_bert_config.json: 100%  52.0/52.0 [00:00<00:00, 5.60kB/s]
config.json: 100%  743/743 [00:00<00:00, 48.6kB/s]
model.safetensors: 100%  133M/133M [00:00<00:00, 171MB/s]
tokenizer_config.json: 100%  366/366 [00:00<00:00, 22.6kB/s]
vocab.txt: 232k/? [00:00<00:00, 9.46MB/s]
tokenizer.json: 711k/? [00:00<00:00, 34.3MB/s]
special_tokens_map.json: 100%  125/125 [00:00<00:00, 7.66kB/s]
config.json: 100%  190/190 [00:00<00:00, 19.9kB/s]
```

Start coding or generate with AI.

## Create Separate Indexes per Embedding

```
import chromadb
from llama_index.vector_stores.chroma import ChromaVectorStore
from llama_index.core import VectorStoreIndex, StorageContext

chroma_client = chromadb.PersistentClient(path="/content/chroma_lic_index")

# Mapping for models → collections
embedding_configs = {
    "openai": (openai_embed, "lic_openai"),
    "bge": (bge_embed, "lic_bge"),
    #"jina": (jina_embed, "lic_jina"),
}

# Store indexes for later use
indexes = {}

for name, (embed_model, collection_name) in embedding_configs.items():
    collection = chroma_client.get_or_create_collection(collection_name)
    vector_store = ChromaVectorStore(chroma_collection=collection)
    storage_context = StorageContext.from_defaults(vector_store=vector_store)

    index = VectorStoreIndex.from_documents(
        documents,
```

```
        storage_context=storage_context,
        embed_model=embed_model,
        show_progress=True,
    )

    indexes[name] = index
```

```
Parsing nodes: 100%|██████████| 21/21 [00:00<00:00, 431.19it/s]
Generating embeddings: 100%|██████████| 21/21 [00:00<00:00, 27.71it/s]
Parsing nodes: 100%|██████████| 21/21 [00:00<00:00, 597.35it/s]
Generating embeddings: 100%|██████████| 21/21 [00:00<00:00, 66.83it/s]
```

Start coding or generate with AI.

## 🔍 Create Query Engines for Each

```
from llama_index.core.query_engine import RetrieverQueryEngine

query_engines = {}

for name, idx in indexes.items():
    # Create a retriever from the index
    retriever = idx.as_retriever(similarity_top_k=5)
    # Initialize the query engine with the retriever
    query_engines[name] = RetrieverQueryEngine.from_args(retriever)
```

## 💬 Ask Same Question Across All Embedding Variants

```
query = "What is the death benefit condition in LIC New Jeevan Shanti policy?"

for name, engine in query_engines.items():
    print(f"\n🔹 {name.upper()} Embeddings Response:\n")
    print(engine.query(query))
```

```
🔹 OPENAI Embeddings Response:

The death benefit condition in LIC New Jeevan Shanti policy includes three options: Lumpsum Death Benefit, Annuitisation of Death Benefit, and In Installment. Under the Lumpsu

🔹 BGE Embeddings Response:

The death benefit condition in LIC New Jeevan Shanti policy states that upon the death of the annuitant, no part of the annuity shall be payable or paid for the period between
```

```
query = "What is the Plan Type of the LIC's New Jeevan Shanti?"

for name, engine in query_engines.items():
```

```
    print(f"\n 🔹 {name.upper()} Embeddings Response:\n")
    print(engine.query(query))
```

```
 🔹 OPENAI Embeddings Response:

The Plan Type of LIC's New Jeevan Shanti is a Deferred Annuity Plan.

 🔹 BGE Embeddings Response:

Deferred Annuity Plan
```

```
query = "What is the Minimum Vesting Age in the New Jeevan Shanti LIC policy?"

for name, engine in query_engines.items():
    print(f"\n 🔹 {name.upper()} Embeddings Response:\n")
    print(engine.query(query))
```

```
 🔹 OPENAI Embeddings Response:

The minimum vesting age in the New Jeevan Shanti LIC policy is not explicitly mentioned in the provided context information.

 🔹 BGE Embeddings Response:

The Minimum Vesting Age in the New Jeevan Shanti LIC policy is not explicitly mentioned in the provided context information.
```

```
query = "What is the policy name?"

for name, engine in query_engines.items():
    print(f"\n 🔹 {name.upper()} Embeddings Response:\n")
    print(engine.query(query))
```

```
 🔹 OPENAI Embeddings Response:

The policy name is "LIC’s New Jeevan Shanti".

 🔹 BGE Embeddings Response:

The policy name is "LIC’s New Jeevan Shanti".
```

```
query = "What is the Free-Look-Period"

for name, engine in query_engines.items():
    print(f"\n 🔹 {name.upper()} Embeddings Response:\n")
    print(engine.query(query))
```

```
 🔹 OPENAI Embeddings Response:

The Free Look period is a 30-day duration from the date of receiving the electronic or physical Policy Document, during which the Policyholder can return the policy to the Cor

 🔹 BGE Embeddings Response:

The Free Look period is a 30-day duration from the date of receiving the electronic or physical Policy Document, during which the Policyholder can return the policy to the Cor
```

```
query = "What is the death benefit under the Joint Life deferred annuity option after the deferment period?"

for name, engine in query_engines.items():
    print(f"\n🔹 {name.upper()} Embeddings Response:\n")
    print(engine.query(query))
```

```
🔹 OPENAI Embeddings Response:

The death benefit under the Joint Life deferred annuity option after the deferment period is the higher of the Purchase Price plus Accrued Additional Benefit on Death minus To

🔹 BGE Embeddings Response:

The death benefit under the Joint Life deferred annuity option after the deferment period is the higher of the Purchase Price plus Accrued Additional Benefit on Death minus To
```

Start coding or generate with AI.

## Retriever Evaluation

✅ Step-by-Step: Retriever Evaluation with MRR 🔍 What's MRR? MRR measures how well the system ranks the first relevant result in response to a query.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

MRR=

Where rank_i is the position of the first correct/relevant document in the retrieved list.

✅ Step 1: Define Ground Truth Questions + Relevant Snippets

```
# Each item: (query, expected text snippet that should be retrieved)
evaluation_set = [
    (
        "What is the death benefit condition?",
        "Death Benefit shall be Higher of"
    ),
    (
        "What is the surrender value formula?",
        "Guaranteed Surrender Value = (GSV Factor * Purchase Price)"
    ),
    (
        "When can a policy loan be availed?",
        "Loan facility shall be available at any time after three months"
    )
]
```

## ✅ Step 2: Build a Function to Compute MRR

```
def compute_mrr_for_retriever(retriever, eval_set, top_k=5):
    reciprocal_ranks = []
    for query, expected_snippet in eval_set:
        nodes = retriever.retrieve(query)
        rank = 0
        for i, node in enumerate(nodes):
            if expected_snippet.lower() in node.text.lower():
                rank = i + 1
                break
        reciprocal_ranks.append(1 / rank if rank > 0 else 0)
    return sum(reciprocal_ranks) / len(reciprocal_ranks)
```

## ✅ Step 3: Evaluate All 2 Embedding Models

```
results = {}

for name, index in indexes.items():
    retriever = index.as_retriever(similarity_top_k=5)
    mrr = compute_mrr_for_retriever(retriever, evaluation_set)
    results[name] = round(mrr, 3)
```

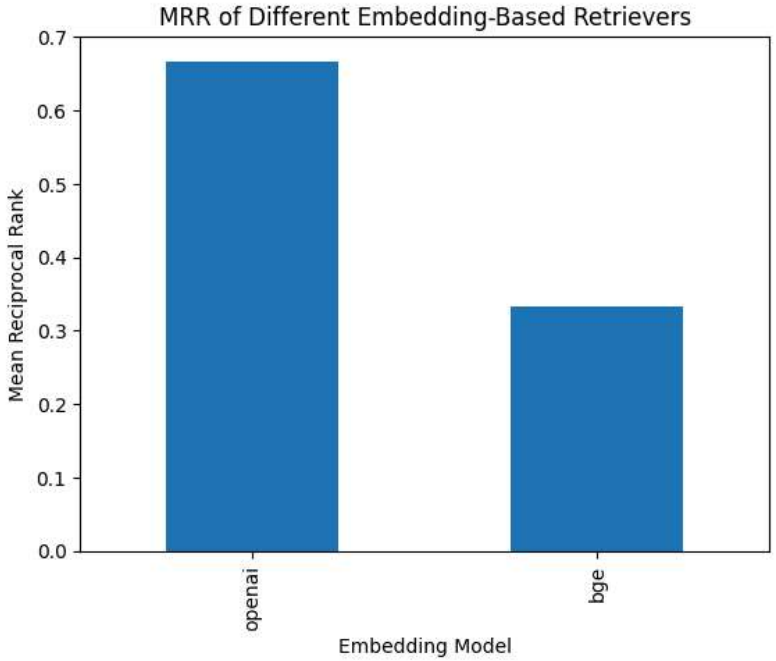
```
import pandas as pd
import matplotlib.pyplot as plt
from IPython.display import display

df = pd.DataFrame(list(results.items()), columns=["Embedding Model", "MRR"])
df.sort_values("MRR", ascending=False, inplace=True)

display(df)

# Optional: plot
df.plot.bar(x="Embedding Model", y="MRR", legend=False, title="MRR of Different Embedding-Based Retrievers")
plt.ylabel("Mean Reciprocal Rank")
plt.show()
```

| | Embedding Model | MRR |
|---|---|---|
| 0 | openai | 0.667 |
| 1 | bge | 0.333 |




Next steps: Generate code with df | View recommended plots | New interactive sheet

higher MRR (Mean Reciprocal Rank) indicates a better retriever.

🔍 Why? MRR evaluates how early the first relevant document appears in the retrieved list:

If a relevant document is ranked 1st, it contributes 1.0

If it's ranked 2nd, it contributes 0.5

If it's not found in top-k, it contributes 0.0

## openai embeddings is doing far better.

But the ground truths are very small statements and are much easier to match literally within retrieved chunks — even irrelevant or partially relevant ones — as long as that exact phrase is found.

This inflates MRR artificially — you're measuring whether the retriever surfaced literal keywords, not whether it retrieved semantically rich and fully relevant context.

The evaluation could be better if the assessment is done with semantic similarity between ground truth and retrieved docs

Start coding or generate with AI.

## Use a much more detailed ground truth. The semantic similarity will mean more with this level grounding instead of a vague phrase

```
# Each item: (query, expected text snippet that should be retrieved)
evaluation_set = [
    (
        "What is the death benefit under joint life annuity?",
        "After the deferment period, under the Joint Life deferred annuity option, on the first death, the annuity continues for the surviving annuitant. Upon the death of the las
    ),
    (
        "What is the surrender value formula?",
        "The policy can be surrendered at any time during its term. The surrender value is the higher of Guaranteed Surrender Value (GSV) or Special Surrender Value. GSV is calcul
    ),
    (
        "Is there a maturity benefit in the policy?",
        "No, there is no maturity benefit under this policy. This is explicitly stated in Part C of the policy document."
    ),
    (
        "What options are available to the nominee for death benefit?",
        "The nominee can choose from: (1) Lump sum death benefit, (2) Annuitisation of the benefit amount into an immediate annuity, or (3) Receiving the benefit in installments o
    ),
    (
        "How is the Additional Benefit on Death calculated?",
        "It is calculated as (Purchase Price * Monthly annuity rate) / 12. This accrues at the end of each policy month only during the deferment period."
    )
]
```

Retriver Evaluation based on semantic similarity

```
from sentence_transformers import SentenceTransformer, util

# Use a universal encoder (same one used for BGE or similar)
similarity_model = SentenceTransformer("BAAI/bge-small-en-v1.5")

def compute_mrr_for_retriever(retriever, eval_set, top_k=5, similarity_threshold=0.75):
    reciprocal_ranks = []

    for query, ground_truth in eval_set:
        # Get embeddings
        gt_embedding = similarity_model.encode(ground_truth, convert_to_tensor=True)
        retrieved_nodes = retriever.retrieve(query)

        rank = 0
        for i, node in enumerate(retrieved_nodes[:top_k]):
            retrieved_embedding = similarity_model.encode(node.text, convert_to_tensor=True)
            score = util.cos_sim(gt_embedding, retrieved_embedding).item()
```

```
            if score >= similarity_threshold:
                rank = i + 1
                break

        reciprocal_ranks.append(1 / rank if rank > 0 else 0)

    return sum(reciprocal_ranks) / len(reciprocal_ranks)
```

Start coding or generate with AI.

🔍 Let's add Average Precision as another evaluation criteria in addition to MRR.Adding Average Precision (AP) alongside MRR gives you a more nuanced view of retriever performance.

What is Average Precision (AP)?

For each query:

AP rewards every relevant chunk retrieved and how early they appear.

It averages the precision values at all relevant positions in the ranked list.

$$\text{AP} = \frac{1}{\#\text{ relevant chunks}} \sum_{\text{ranks where relevant chunks appear}} \text{Precision@rank}$$

Then **Mean Average Precision (MAP)** is:

$$\text{MAP} = \frac{1}{|Q|} \sum_{q=1}^{|Q|} AP_q$$

✅ Updated Evaluation Function

This computes both MRR and MAP using semantic similarity:

```
from sentence_transformers import SentenceTransformer, util

similarity_model = SentenceTransformer("BAAI/bge-small-en-v1.5")

def compute_mrr_and_map_for_retriever(retriever, eval_set, top_k=5, similarity_threshold=0.75):
    reciprocal_ranks = []
    average_precisions = []

    for query, ground_truth in eval_set:
        gt_embedding = similarity_model.encode(ground_truth, convert_to_tensor=True)
        retrieved_nodes = retriever.retrieve(query)

        hits = []
        for i, node in enumerate(retrieved_nodes[:top_k]):
            node_embedding = similarity_model.encode(node.text, convert_to_tensor=True)
```

```
            similarity = util.cos_sim(gt_embedding, node_embedding).item()
            hits.append(similarity >= similarity_threshold)

        # MRR: first relevant item
        try:
            first_relevant = hits.index(True)
            reciprocal_ranks.append(1 / (first_relevant + 1))
        except ValueError:
            reciprocal_ranks.append(0)

        # Average Precision
        num_hits = 0
        precision_at_i = []
        for i, is_relevant in enumerate(hits):
            if is_relevant:
                num_hits += 1
                precision_at_i.append(num_hits / (i + 1))

        ap = sum(precision_at_i) / num_hits if num_hits > 0 else 0
        average_precisions.append(ap)

    mrr = sum(reciprocal_ranks) / len(reciprocal_ranks)
    map_score = sum(average_precisions) / len(average_precisions)

    return round(mrr, 4), round(map_score, 4)


results = {}

for name, index in indexes.items():
    retriever = index.as_retriever(similarity_top_k=5)
    mrr, map_score = compute_mrr_and_map_for_retriever(retriever, evaluation_set)
    results[name] = {"MRR": mrr, "MAP": map_score}


import pandas as pd
from IPython.display import display

df = pd.DataFrame(results).T.reset_index().rename(columns={"index": "Embedding Model"})
display(df)
```

| | Embedding Model | MRR | MAP |
|---|---|---|---|
| 0 | openai | 1.0 | 1.0 |
| 1 | bge | 0.9 | 0.9 |

Next steps: Generate code with df | View recommended plots | New interactive sheet

Start coding or generate with AI.

## Response synthesis

Use a common retriever (e.g. BGE or OpenAI)

Synthesize responses using:

- GPT-4 (via OpenAI)
- GPT 3.5 turbo

Compare generated answers

```
Start coding or generate with AI.
```

⚙️ Step 1: Setup Embedding-Based Retrieval (You already have this)

```
retriever = indexes["bge"].as_retriever(similarity_top_k=5)
retrieved_nodes = retriever.retrieve("What is the death benefit under joint life annuity?")
```

🤖 Step 2: Setup Multiple LLMs for Response Synthesis

- GPT-4 (OpenAI)

```
from llama_index.llms.openai import OpenAI

gpt4_llm = OpenAI(model="gpt-4", temperature=0)
```

## GPT-3.5 turbo

```
from llama_index.llms.openai import OpenAI

gpt3_llm = OpenAI(model="gpt-3.5-turbo", temperature=0)
```

## Mistral 7B (open access)

TheBloke/Mistral-7B-Instruct-v0.2-GGUF

```
!pip install llama-cpp-python
```

```
Collecting llama-cpp-python
  Downloading llama_cpp_python-0.3.11.tar.gz (79.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 79.1/79.1 MB 70.0 MB/s eta 0:00:00
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Installing backend dependencies ... done
  Preparing metadata (pyproject.toml) ... done
Requirement already satisfied: typing-extensions>=4.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-cpp-python) (4.14.0)
Requirement already satisfied: numpy>=1.20.0 in /usr/local/lib/python3.11/dist-packages (from llama-cpp-python) (2.0.2)
Collecting diskcache>=5.6.1 (from llama-cpp-python)
  Downloading diskcache-5.6.3-py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: jinja2>=2.11.3 in /usr/local/lib/python3.11/dist-packages (from llama-cpp-python) (3.1.6)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2>=2.11.3->llama-cpp-python) (3.0.2)
Downloading diskcache-5.6.3-py3-none-any.whl (45 kB)
Building wheels for collected packages: llama-cpp-python
  Building wheel for llama-cpp-python (pyproject.toml) ... done
  Created wheel for llama-cpp-python: filename=llama_cpp_python-0.3.11-cp311-cp311-linux_x86_64.whl size=4065934 sha256=089ac955d46a1a415672002ea9b7c03eed5202833fce0670ddc8f8a
  Stored in directory: /root/.cache/pip/wheels/1a/65/e9/3bd26ec174c6e148cce8a3876d7ed32652e3508ebe262c197a
Successfully built llama-cpp-python
Installing collected packages: diskcache, llama-cpp-python
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2/2 [llama-cpp-python]
Successfully installed diskcache-5.6.3 llama-cpp-python-0.3.11
```

## 📦 ✅ Download Mistral GGUF from HuggingFace

```
from huggingface_hub import hf_hub_download

mistral_path = hf_hub_download(
    repo_id="TheBloke/Mistral-7B-Instruct-v0.2-GGUF",
    filename="mistral-7b-instruct-v0.2.Q5_K_M.gguf"
)
```

## Smaller Model. Use this to make the execution faster. This model runs well on CPU and is lighter than Q5 or Q6

Good tradeoff between speed and quality for local inference

```
from huggingface_hub import hf_hub_download

mistral_path = hf_hub_download(
    repo_id="TheBloke/Mistral-7B-Instruct-v0.1-GGUF",
    filename="mistral-7b-instruct-v0.1.Q4_K_M.gguf"
)
```

```
mistral-7b-instruct-v0.1.Q4_K_M.gguf:  100%  4.37G/4.37G  [00:37<00:00,  244MB/s]
```

## 🔁 Load Model with llama-cpp

```
from llama_cpp import Llama
```

```
llama_context: n_batch       = 512
llama_context: n_ubatch      = 512
llama_context: causal_attn   = 1
llama_context: flash_attn    = 0
llama_context: freq_base     = 1000000.0
llama_context: freq_scale    = 1
llama_context: n_ctx_per_seq (4096) < n_ctx_train (32768) -- the full capacity of the model will not be utilized
set_abort_callback: call
llama_context:        CPU  output buffer size =     0.12 MiB
create_memory: n_ctx = 4096 (padded)
llama_kv_cache_unified: layer   0: dev = CPU
llama_kv_cache_unified: layer   1: dev = CPU
llama_kv_cache_unified: layer   2: dev = CPU
llama_kv_cache_unified: layer   3: dev = CPU
llama_kv_cache_unified: layer   4: dev = CPU
llama_kv_cache_unified: layer   5: dev = CPU
llama_kv_cache_unified: layer   6: dev = CPU
llama_kv_cache_unified: layer   7: dev = CPU
llama_kv_cache_unified: layer   8: dev = CPU
llama_kv_cache_unified: layer   9: dev = CPU
llama_kv_cache_unified: layer  10: dev = CPU
llama_kv_cache_unified: layer  11: dev = CPU
llama_kv_cache_unified: layer  12: dev = CPU
llama_kv_cache_unified: layer  13: dev = CPU
llama_kv_cache_unified: layer  14: dev = CPU
llama_kv_cache_unified: layer  15: dev = CPU
llama_kv_cache_unified: layer  16: dev = CPU
llama_kv_cache_unified: layer  17: dev = CPU
llama_kv_cache_unified: layer  18: dev = CPU
llama_kv_cache_unified: layer  19: dev = CPU
llama_kv_cache_unified: layer  20: dev = CPU
llama_kv_cache_unified: layer  21: dev = CPU
llama_kv_cache_unified: layer  22: dev = CPU
llama_kv_cache_unified: layer  23: dev = CPU
llama_kv_cache_unified: layer  24: dev = CPU
llama_kv_cache_unified: layer  25: dev = CPU
llama_kv_cache_unified: layer  26: dev = CPU
llama_kv_cache_unified: layer  27: dev = CPU
llama_kv_cache_unified: layer  28: dev = CPU
llama_kv_cache_unified: layer  29: dev = CPU
llama_kv_cache_unified: layer  30: dev = CPU
llama_kv_cache_unified: layer  31: dev = CPU
llama_kv_cache_unified:        CPU KV buffer size =   512.00 MiB
llama_kv_cache_unified: size =  512.00 MiB (  4096 cells,  32 layers,  1 seqs), K (f16):  256.00 MiB, V (f16):  256.00 MiB
llama_context: enumerating backends
llama_context: backend_ptrs.size() = 1
llama_context: max_nodes = 65536
llama_context: worst-case: n_tokens = 512, n_seqs = 1, n_outputs = 0
graph_reserve: reserving a graph for ubatch with n_tokens =  512, n_seqs =  1, n_outputs =  512
graph_reserve: reserving a graph for ubatch with n_tokens =    1, n_seqs =  1, n_outputs =    1
graph_reserve: reserving a graph for ubatch with n_tokens =  512, n_seqs =  1, n_outputs =  512
llama_context:        CPU compute buffer size =   296.01 MiB
llama_context: graph nodes  = 1158
llama_context: graph splits = 1
CPU : SSE3 = 1 | SSSE3 = 1 | AVX = 1 | AVX2 = 1 | F16C = 1 | FMA = 1 | BMI2 = 1 | LLAMAFILE = 1 | OPENMP = 1 | REPACK = 1 |
Model metadata: {'tokenizer.chat_template': "{{ bos_token }}{% for message in messages %}{% if (message['role'] == 'user') != (loop.index0 % 2 == 0) %}{{ raise_exception('Con
Available chat formats from metadata: chat_template.default
Guessed chat format: mistral-instruct
```

## Wrap it into LlamaIndex as a Synthesizer LLM

```
!pip install llama-index-llms-llama-cpp
```

```
Requirement already satisfied: dirtyjson<2,>=1.0.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (1.0.8)
Requirement already satisfied: filetype<2,>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (1.2.0)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (2025.3.2)
Requirement already satisfied: httpx in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (0.28.1)
Requirement already satisfied: llama-index-workflows<2,>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp)
Requirement already satisfied: nest-asyncio<2,>=1.5.8 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (1.6.0)
Requirement already satisfied: networkx>=3.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (3.5)
Requirement already satisfied: nltk>3.8.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (3.9.1)
Requirement already satisfied: pillow>=9.0.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (11.2.1)
Requirement already satisfied: pydantic>=2.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (2.11.7)
Requirement already satisfied: pyyaml>=6.0.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (6.0.2)
Requirement already satisfied: requests>=2.31.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (2.32.3)
Requirement already satisfied: setuptools>=80.9.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (80.9.0)
Requirement already satisfied: sqlalchemy>=1.4.49 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13.0,>=0.12.0->llama-index-
Requirement already satisfied: tenacity!=8.4.0,<10.0.0,>=8.2.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp)
Requirement already satisfied: tiktoken>=0.7.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (0.9.0)
Requirement already satisfied: tqdm<5,>=4.66.1 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (4.67.1)
Requirement already satisfied: typing-inspect>=0.8.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (0.9.0)
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (1.17.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-c
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llam
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-c
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-
Requirement already satisfied: griffe in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (1.7.3)
Requirement already satisfied: platformdirs in /usr/local/lib/python3.11/dist-packages (from banks<3,>=2.0.0->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (4
Requirement already satisfied: llama-index-instrumentation>=0.1.0 in /usr/local/lib/python3.11/dist-packages (from llama-index-workflows<2,>=1.0.1->llama-index-core<0.13.0,>=
Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.11/dist-packages (from yarl<2.0,>=1.17.0->aiohttp<4,>=3.8.6->llama-index-core<0.13.0,>=0.12.0->llama-index-
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2>=2.11.3->llama-cpp-python<0.4.0,>=0.3.0->llama-index-llms-llama-cpp) (3
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (8.2.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (1.5.1)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk>3.8.1->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-cpp) (202
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-lla
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llam
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.8.0->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-l
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31.0->llama-index-core<0.13.0,>=0.12.0->llama-index-llms-llama-
Requirement already satisfied: greenlet>=1 in /usr/local/lib/python3.11/dist-packages (from sqlalchemy>=1.4.49->sqlalchemy[asyncio]>=1.4.49->llama-index-core<0.13.0,>=0.12.0-
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2/2 [llama-index-llms-llama-cpp]
Successfully installed llama-cpp-python-0.3.11 llama-index-llms-llama-cpp-0.4.0
```

```
from llama_index.llms.llama_cpp import LlamaCPP

mistral_llm = LlamaCPP(
    model_path=mistral_path,
    temperature=0.2,
    n_gpu_layers=20
    context_window=1024,
    max_new_tokens=256,
    generate_kwargs={"top_p": 0.9},
    model_kwargs={"n_threads": 8}
)
```

```
llama_context: enumerating backends
llama_context: backend_ptrs.size() = 1
llama_context: max_nodes = 65536
llama_context: worst-case: n_tokens = 512, n_seqs = 1, n_outputs = 0
graph_reserve: reserving a graph for ubatch with n_tokens =  512, n_seqs =  1, n_outputs =  512
graph_reserve: reserving a graph for ubatch with n_tokens =    1, n_seqs =  1, n_outputs =    1
graph_reserve: reserving a graph for ubatch with n_tokens =  512, n_seqs =  1, n_outputs =  512
llama_context:        CPU compute buffer size =    98.01 MiB
llama_context: graph nodes  = 1158
llama_context: graph splits = 1
CPU : SSE3 = 1 | SSSE3 = 1 | AVX = 1 | AVX2 = 1 | F16C = 1 | FMA = 1 | BMI2 = 1 | LLAMAFILE = 1 | OPENMP = 1 | REPACK = 1 |
Model metadata: {'tokenizer.ggml.unknown_token_id': '0', 'tokenizer.ggml.eos_token_id': '2', 'general.architecture': 'llama', 'llama.rope.freq_base': '10000.000000', 'llama.c
Using fallback chat format: llama-2
```

Start coding or generate with AI.

## 🧠 Step 3: Response Synthesis Using Both LLMs

```
from llama_index.core.response_synthesizers import get_response_synthesizer

def synthesize_response(llm, retrieved_nodes, query):
    synthesizer = get_response_synthesizer(llm=llm)
    return synthesizer.synthesize(query=query, nodes=retrieved_nodes).response
```

Start coding or generate with AI.

```
# Assuming retrieved_nodes and gpt4_llm are already defined from previous cells

query = "What is the death benefit under joint life annuity?"

# Define retrieved_nodes by retrieving from the index
retriever = indexes["bge"].as_retriever(similarity_top_k=5)
retrieved_nodes = retriever.retrieve(query)


print("\n🔹 GPT-4 Response:")
gpt4_response = synthesize_response(gpt4_llm, retrieved_nodes, query)
print(gpt4_response)
```

```
🔹 GPT-4 Response:
The death benefit under a joint life annuity in the LIC's New Jeevan Shanti policy is the higher of the Purchase Price plus Accrued Additional Benefit on Death minus the total
```

```
print("\n🔹 Mistral Response:")
mistral_response = synthesize_response(mistral_llm, retrieved_nodes, query)
print(mistral_response)
```

```
🔹 Mistral Response:
llama_perf_context_print:        load time =  227284.01 ms
llama_perf_context_print: prompt eval time =  227280.86 ms /   854 tokens (  266.14 ms per token,     3.76 tokens per second)
llama_perf_context_print:        eval time =    6873.62 ms /    13 runs   (  528.74 ms per token,     1.89 tokens per second)
llama_perf_context_print:       total time =  234164.64 ms /   867 tokens
```

```
Llama.generate: 1 prefix-match hit, remaining 889 prompt tokens to eval
llama_perf_context_print:        load time =  227284.01 ms
llama_perf_context_print: prompt eval time =  233992.42 ms /   889 tokens (  263.21 ms per token,     3.80 tokens per second)
llama_perf_context_print:        eval time =   36876.01 ms /    64 runs   (  576.19 ms per token,     1.74 tokens per second)
llama_perf_context_print:       total time =  270903.55 ms /   953 tokens
Llama.generate: 36 prefix-match hit, remaining 459 prompt tokens to eval
llama_perf_context_print:        load time =  227284.01 ms
llama_perf_context_print: prompt eval time =  118788.44 ms /   459 tokens (  258.80 ms per token,     3.86 tokens per second)
llama_perf_context_print:        eval time =  141592.98 ms /   255 runs   (  555.27 ms per token,     1.80 tokens per second)
llama_perf_context_print:       total time =  260562.39 ms /   714 tokens

Based on the context provided, it appears that the policyholder is not alive. In such cases, a policy of life insurance may be called in question within three years from the d
```

## 📊 Step 4: Compare Answers

```
query = "What is the death benefit under joint life annuity?"
retrieved_nodes = retriever.retrieve(query)

responses = {
    "GPT-4": synthesize_response(gpt4_llm, retrieved_nodes, query),
    "GPT-3.5 Turbo": synthesize_response(gpt3_llm, retrieved_nodes, query),
}

for model, answer in responses.items():
    print(f"\n🔷 {model}:\n{answer}\n")
```

```
🔷 GPT-4:
The death benefit under a joint life annuity in the LIC's New Jeevan Shanti policy is the higher of either the Purchase Price plus Accrued Additional Benefit on Death minus th


🔷 GPT-3.5 Turbo:
The death benefit under joint life annuity is the higher of the Purchase Price plus Accrued Additional Benefit on Death minus Total annuity amount payable till date of death,
```

```
query = "What are the potential options available to the nominee(s) for receiving the death benefit amount?"
retrieved_nodes = retriever.retrieve(query)

responses = {
    "GPT-4": synthesize_response(gpt4_llm, retrieved_nodes, query),
    "GPT-3.5 Turbo": synthesize_response(gpt3_llm, retrieved_nodes, query),
}

for model, answer in responses.items():
    print(f"\n🔷 {model}:\n{answer}\n")
```

```
🔷 GPT-4:
The death benefit amount can be received by the nominee(s) in three ways. The first option is the Lumpsum Death Benefit, where the entire benefit amount payable on death is gi


🔷 GPT-3.5 Turbo:
The potential options available to the nominee(s) for receiving the death benefit amount include receiving the entire benefit amount in lump sum, utilizing the benefit amount
```