Gen_AI

July 30, 2025

0.1 Assignment 12: Generative AI and RAG Pipeline

0.2 Name: Daniel Muthama

0.3 Program: DATA and AI

0.4 Date: July 29, 2025

Introduction This notebook implements a Retrieval-Augmented Generation (RAG) pipeline to process a PDF document, extract relevant information, and generate context-aware answers using a generative model. The pipeline uses `langchain` for document loading and chunking, `sentence-transformers` for embeddings, `faiss-cpu` for vector storage, and `transformers` (FLAN-T5) for answer generation. The goal is to demonstrate how retrieval enhances generative question-answering compared to generic answers.

0.5 Step 1: Set Up the Environment

Install Required Libraries:

langchain: Framework for chaining NLP components.

sentence-transformers: For generating embeddings.

faiss-cpu: Efficient vector similarity search.

pypdf: PDF text extraction.

Import Libraries:

0.6 Step 2: Load and Preprocess the PDF

Load the PDF:

Split into Chunks:

chunk_size=500: Split text into 500-character segments.

chunk_overlap=50: Ensures context continuity between chunks.

0.7 Step 3: Create Embeddings and Vector Store

Initialize Embeddings:

Uses the lightweight all-MiniLM-L6-v2 model for sentence embeddings.

Build FAISS Vector Store:

FAISS enables fast similarity search for retrieval.

0.8 Step 4: Initialize the Generative Model

Load FLAN-T5 (Text-to-Text Model):

FLAN-T5 is a powerful open-source model for generative tasks.

0.9 Step 5: Implement the RAG Pipeline

Define the Query Function:

Test the Pipeline:

[1]:

```
# ## Step 1: Install Required Packages
# Uninstall conflicting packages, install numpy first, and restart runtime to␣
 ↪avoid binary incompatibilities.

#!pip uninstall -y faiss-cpu faiss-gpu numpy transformers tokenizers␣
 ↪huggingface_hub sentence-transformers langchain langchain-community pypdf␣
 ↪pydantic langsmith
!pip install --upgrade pip
!pip install numpy==1.26.4 packaging>=24.1  # Pin numpy and upgrade packaging

# After restart, run installations again
!pip install faiss-cpu  # Install latest prebuilt faiss-cpu
!pip install pydantic==2.9.2 langsmith==0.1.13  # Use compatible pydantic␣
 ↪version
!pip install sentence-transformers==2.2.2 huggingface_hub==0.23.0  # Use␣
 ↪compatible huggingface_hub
!pip install -q langchain==0.1.13 langchain-community==0.0.29
!pip install -q transformers==4.41.1 pypdf==3.17.4 tenacity  # Add tenacity,␣
 ↪let pip select tokenizers
```

```
Requirement already satisfied: pip in
/home/daniel/anaconda3/lib/python3.12/site-packages (25.1.1)
Requirement already satisfied: faiss-cpu in
/home/daniel/anaconda3/lib/python3.12/site-packages (1.11.0.post1)
Requirement already satisfied: numpy<3.0,>=1.25.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from faiss-cpu) (1.26.4)
Requirement already satisfied: packaging in
/home/daniel/anaconda3/lib/python3.12/site-packages (from faiss-cpu) (23.2)
Requirement already satisfied: pydantic==2.9.2 in
/home/daniel/anaconda3/lib/python3.12/site-packages (2.9.2)
Collecting langsmith==0.1.13
  Using cached langsmith-0.1.13-py3-none-any.whl.metadata (13 kB)
Requirement already satisfied: annotated-types>=0.6.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from pydantic==2.9.2)
(0.6.0)
Requirement already satisfied: pydantic-core==2.23.4 in
```

```
/home/daniel/anaconda3/lib/python3.12/site-packages (from pydantic==2.9.2)
(2.23.4)
Requirement already satisfied: typing-extensions>=4.6.1 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from pydantic==2.9.2)
(4.14.1)
Requirement already satisfied: orjson<4.0.0,>=3.9.14 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from langsmith==0.1.13)
(3.11.1)
Requirement already satisfied: requests<3,>=2 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from langsmith==0.1.13)
(2.32.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
requests<3,>=2->langsmith==0.1.13) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
requests<3,>=2->langsmith==0.1.13) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
requests<3,>=2->langsmith==0.1.13) (2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
requests<3,>=2->langsmith==0.1.13) (2024.8.30)
Using cached langsmith-0.1.13-py3-none-any.whl (64 kB)
Installing collected packages: langsmith
  Attempting uninstall: langsmith
    Found existing installation: langsmith 0.1.147
    Uninstalling langsmith-0.1.147:
      Successfully uninstalled langsmith-0.1.147
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.

langchain 0.1.13 requires langsmith<0.2.0,>=0.1.17, but you have langsmith

0.1.13 which is incompatible.

Successfully installed langsmith-0.1.13
Requirement already satisfied: sentence-transformers==2.2.2 in
/home/daniel/anaconda3/lib/python3.12/site-packages (2.2.2)
Requirement already satisfied: huggingface_hub==0.23.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (0.23.0)
Requirement already satisfied: transformers<5.0.0,>=4.6.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from sentence-
transformers==2.2.2) (4.41.1)
Requirement already satisfied: tqdm in
/home/daniel/anaconda3/lib/python3.12/site-packages (from sentence-
transformers==2.2.2) (4.66.5)
```

```
Requirement already satisfied: torch>=1.6.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from sentence-
transformers==2.2.2) (2.5.1)
Requirement already satisfied: torchvision in
/home/daniel/anaconda3/lib/python3.12/site-packages (from sentence-
transformers==2.2.2) (0.20.1)
Requirement already satisfied: numpy in
/home/daniel/anaconda3/lib/python3.12/site-packages (from sentence-
transformers==2.2.2) (1.26.4)
Requirement already satisfied: scikit-learn in
/home/daniel/anaconda3/lib/python3.12/site-packages (from sentence-
transformers==2.2.2) (1.3.2)
Requirement already satisfied: scipy in
/home/daniel/anaconda3/lib/python3.12/site-packages (from sentence-
transformers==2.2.2) (1.11.4)
Requirement already satisfied: nltk in
/home/daniel/anaconda3/lib/python3.12/site-packages (from sentence-
transformers==2.2.2) (3.9.1)
Requirement already satisfied: sentencepiece in
/home/daniel/anaconda3/lib/python3.12/site-packages (from sentence-
transformers==2.2.2) (0.2.0)
Requirement already satisfied: filelock in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
huggingface_hub==0.23.0) (3.13.1)
Requirement already satisfied: fsspec>=2023.5.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
huggingface_hub==0.23.0) (2024.6.1)
Requirement already satisfied: packaging>=20.9 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
huggingface_hub==0.23.0) (23.2)
Requirement already satisfied: pyyaml>=5.1 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
huggingface_hub==0.23.0) (6.0.1)
Requirement already satisfied: requests in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
huggingface_hub==0.23.0) (2.32.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
huggingface_hub==0.23.0) (4.14.1)
Requirement already satisfied: regex!=2019.12.17 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
transformers<5.0.0,>=4.6.0->sentence-transformers==2.2.2) (2024.9.11)
Requirement already satisfied: tokenizers<0.20,>=0.19 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
transformers<5.0.0,>=4.6.0->sentence-transformers==2.2.2) (0.19.1)
Requirement already satisfied: safetensors>=0.4.1 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
transformers<5.0.0,>=4.6.0->sentence-transformers==2.2.2) (0.4.5)
```

```
Requirement already satisfied: networkx in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (3.3)
Requirement already satisfied: jinja2 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (3.1.2)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (12.4.127)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (12.4.127)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (9.1.0.70)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (12.4.5.8)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (11.2.1.3)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (10.3.5.147)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (11.6.1.9)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (12.3.1.170)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (12.4.127)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (12.4.127)
Requirement already satisfied: triton==3.1.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (3.1.0)
Requirement already satisfied: setuptools in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (69.1.0)
```

```
Requirement already satisfied: sympy==1.13.1 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
torch>=1.6.0->sentence-transformers==2.2.2) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
sympy==1.13.1->torch>=1.6.0->sentence-transformers==2.2.2) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
jinja2->torch>=1.6.0->sentence-transformers==2.2.2) (2.1.3)
Requirement already satisfied: click in
/home/daniel/anaconda3/lib/python3.12/site-packages (from nltk->sentence-
transformers==2.2.2) (8.1.7)
Requirement already satisfied: joblib in
/home/daniel/anaconda3/lib/python3.12/site-packages (from nltk->sentence-
transformers==2.2.2) (1.3.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
requests->huggingface_hub==0.23.0) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
requests->huggingface_hub==0.23.0) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
requests->huggingface_hub==0.23.0) (2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from
requests->huggingface_hub==0.23.0) (2024.8.30)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from scikit-
learn->sentence-transformers==2.2.2) (3.2.0)
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in
/home/daniel/anaconda3/lib/python3.12/site-packages (from torchvision->sentence-
transformers==2.2.2) (10.1.0)
```

[2]:

```python
# ## Step 2: Import Libraries
# Import libraries for document loading, chunking, embeddings, vector storage,␣
 ↪text generation, and retries.

from langchain.document_loaders import PyPDFLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings import HuggingFaceEmbeddings
from langchain.vectorstores import FAISS
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM, pipeline
import os
import requests
from tenacity import retry, stop_after_attempt, wait_exponential,␣
 ↪retry_if_exception_type
```

```
# ## Step 3: Load and Preprocess PDF
# Load the PDF and split it into chunks. Download from Google Drive if not␣
 ↪present.

def load_and_chunk_pdf(pdf_path="document.pdf"):
    if not os.path.exists(pdf_path):
        print("Downloading document from Google Drive...")
        try:
            # Replace with your actual Google Drive file ID
            file_id = "1KMEK23HfyRwGi46bAGNaqdovLRmEwDH33hihb4aEe0k"
            url = f"https://drive.google.com/uc?export=download&id={file_id}"
            session = requests.Session()
            response = session.get(url, stream=True, timeout=30)
            response.raise_for_status()

            for key, value in response.cookies.items():
                if 'download_warning' in key:
                    url = f"https://drive.google.com/uc?
 ↪export=download&confirm={value}&id={file_id}"
                    response = session.get(url, stream=True, timeout=30)
                    break

            with open(pdf_path, 'wb') as f:
                for chunk in response.iter_content(chunk_size=8192):
                    if chunk:
                        f.write(chunk)
            print(f"Downloaded document saved as {pdf_path}")
        except Exception as e:
            print(f"Failed to download document: {e}")
            return None

    try:
        loader = PyPDFLoader(pdf_path)
        docs = loader.load()
        splitter = RecursiveCharacterTextSplitter(
            chunk_size=500,
            chunk_overlap=50,
            separators=["\n\n", "\n", " ", ""]
        )
        chunks = splitter.split_documents(docs)
        print(f"Loaded {len(docs)} pages, split into {len(chunks)} chunks.")
        return chunks
    except Exception as e:
        print(f"Error processing PDF: {e}")
        return None
```

```python
# ## Step 4: Create Embeddings and Vector Store
# Create embeddings with retry logic for model download.

@retry(
    stop=stop_after_attempt(3),
    wait=wait_exponential(multiplier=1, min=4, max=10),
    retry=retry_if_exception_type(Exception)
)
def create_vector_store(chunks):
    if chunks is None:
        print("Error: No chunks to process. Check the PDF loading step.")
        return None
    embeddings = HuggingFaceEmbeddings(
        model_name="sentence-transformers/all-MiniLM-L6-v2",
        model_kwargs={'device': 'cpu'}
    )
    vectorstore = FAISS.from_documents(chunks, embeddings)
    print("Vector store created successfully.")
    return vectorstore

# ## Step 5: Initialize LLM
# Load FLAN-T5 with retry logic for model download.

@retry(
    stop=stop_after_attempt(3),
    wait=wait_exponential(multiplier=1, min=4, max=10),
    retry=retry_if_exception_type(Exception)
)
def initialize_llm(model_name="google/flan-t5-large"):
    tokenizer = AutoTokenizer.from_pretrained(model_name)
    model = AutoModelForSeq2SeqLM.from_pretrained(model_name)
    pipeline_obj = pipeline(
        "text2text-generation",
        model=model,
        tokenizer=tokenizer,
        device=-1  # Use CPU
    )
    print(f"Loaded {model_name} model.")
    return pipeline_obj

# ## Step 6: Implement RAG Query Function
# Retrieve relevant chunks and generate a context-aware answer.

def query_rag(question, vectorstore, llm_pipeline, k=3):
    if vectorstore is None:
        print("Error: Vector store is not initialized.")
        return None
    relevant_docs = vectorstore.as_retriever().
 ↪get_relevant_documents(question)[:k]
    context = "\n".join([doc.page_content for doc in relevant_docs])

    prompt = f"""Answer the question using only the following context. Provide␣
 ↪a clear, concise, and complete response in full sentences.

Context:
{context}

Question: {question}

Answer:"""

    response = llm_pipeline(
        prompt,
        max_new_tokens=200,
        temperature=0.9,
        top_k=50,
        top_p=0.9,
        do_sample=True
    )
    return response[0]['generated_text']

# ## Step 7: Main Execution
# Execute the pipeline and compare RAG vs. generic answers.

if __name__ == "__main__":
    # 1. Load and chunk PDF
    chunks = load_and_chunk_pdf("document2.pdf")

    # 2. Create vector store
    vectorstore = create_vector_store(chunks)

    # 3. Initialize LLM
    llm_pipeline = initialize_llm()

    # 4. Test RAG pipeline
    question = "Summarize the key points of this document in a paragraph of 100␣
 ↪words."
    answer = query_rag(question, vectorstore, llm_pipeline)

    if answer:
        print("\n" + "="*50)
        print("Question:", question)
        print("-"*50)
        print("Answer:", answer)
```

```python
        print("="*50 + "\n")

    # 5. Compare with generic answer
    generic_prompt = f"Answer the question without any specific context:
  ↪\n\nQuestion: {question}\n\nAnswer:"
    generic_answer = llm_pipeline(
        generic_prompt,
        max_new_tokens=200,
        temperature=0.9,
        top_k=50,
        top_p=0.9,
        do_sample=True
    )[0]['generated_text']
    print("\n" + "="*50)
    print("Generic Answer (No Context):")
    print("-"*50)
    print("Answer:", generic_answer)
    print("="*50 + "\n")
```

```
2025-07-30 10:13:03.297012: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1753859584.408972   10102 cuda_dnn.cc:8579] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1753859584.857201   10102 cuda_blas.cc:1407] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
W0000 00:00:1753859587.269040   10102 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1753859587.269137   10102 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1753859587.269145   10102 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
W0000 00:00:1753859587.269152   10102 computation_placer.cc:177] computation
placer already registered. Please check linkage and avoid linking the same
target more than once.
2025-07-30 10:13:07.421535: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 FMA, in other operations, rebuild
TensorFlow with the appropriate compiler flags.
```

```
Loaded 3 pages, split into 11 chunks.
Vector store created successfully.

model.safetensors:  53%|#####2    | 1.65G/3.13G [00:00<?, ?B/s]

Error while downloading from https://cas-bridge.xethub.hf.co/xet-bridge-us/63526
f4c7e4cc3135fd0ff1a/6acb653f3c05b8398d386c9e96027f0bde0a3e41095eb027851460889132
4f49?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Content-Sha256=UNSIGNED-
PAYLOAD&X-Amz-Credential=cas%2F20250730%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-
Date=20250730T070636Z&X-Amz-Expires=3600&X-Amz-Signature=53a50434f362ca0a78ce993
126433ba25b6b81acbb224bf41c656dedeba4930b&X-Amz-SignedHeaders=host&X-Xet-Cas-
Uid=66b5fb39fab7e0d167dbbf90&response-content-disposition=inline%3B+filename*%3D
UTF-8%27%27model.safetensors%3B+filename%3D%22model.safetensors%22%3B&x-
id=GetObject&Expires=1753862796&Policy=eyJTdGF0ZW1lbnQiOlt7IkNvbmRpdGlvbiI6eyJEY
XRlTGVzc1RoYW4iOnsiQVdTOkVwb2NoVGltZSI6MTc1Mzg2Mjc5Nn19LCJSZXNvdXJjZSI6Imh0dHBzO
i8vY2FzLWJyaWRnZS54ZXRodWIuaGYuY28veGV0LWJyaWRnZS11cy82MzUyNmY0YzdlNGNjMzEzNWZkM
GZmMWEvNmFjYjY1M2YzYzA1YjgzOThkMzg2YzllOTYwMjdmMGJkZTBhM2U0MTA5NWViMDI3ODUxNDYwO
Dg5MTMyNGY0OSoifV19&Signature=urvr1rMbFOGKMjS9yDB8POL7g5XPIHckQzk2RWAMTgYeHrcuFM
g8w1sM1v7zhtrP3NwQyEsHu3UfgaBfhc2Vozp1tYjuKCkuwZUEaBSQyIalrkF5W1t12OWhBmslv%7E9r
v78hxZzL6qvoWfN-
x4DdOzg4HnLI0jRf0fBJhiHEWUGCrGyTcOZy%7EQXJPButYOsptPzJqLqWEBRCN31H8doKhnzEaXZeN-
71vf5vfPOD-
yb%7EoOEMymohqDbIPn6ZKTT3OHoiYiq3K2G6fLO9zt%7EdGhEJRBZbw6mGWcF3p2czUZRuCef9k5XqY
ftiljqRkvqa%7EO5c2%7EHkjEaWEn24KaVX9Q__&Key-Pair-Id=K2L8F4GPSG1IFC:
HTTPSConnectionPool(host='cas-bridge.xethub.hf.co', port=443): Read timed out.
Trying to resume download…

model.safetensors:  60%|#####9    | 1.88G/3.13G [00:00<?, ?B/s]

Error while downloading from https://cas-bridge.xethub.hf.co/xet-bridge-us/63526
f4c7e4cc3135fd0ff1a/6acb653f3c05b8398d386c9e96027f0bde0a3e41095eb027851460889132
4f49?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Content-Sha256=UNSIGNED-
PAYLOAD&X-Amz-Credential=cas%2F20250730%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-
Date=20250730T070636Z&X-Amz-Expires=3600&X-Amz-Signature=53a50434f362ca0a78ce993
126433ba25b6b81acbb224bf41c656dedeba4930b&X-Amz-SignedHeaders=host&X-Xet-Cas-
Uid=66b5fb39fab7e0d167dbbf90&response-content-disposition=inline%3B+filename*%3D
UTF-8%27%27model.safetensors%3B+filename%3D%22model.safetensors%22%3B&x-
id=GetObject&Expires=1753862796&Policy=eyJTdGF0ZW1lbnQiOlt7IkNvbmRpdGlvbiI6eyJEY
XRlTGVzc1RoYW4iOnsiQVdTOkVwb2NoVGltZSI6MTc1Mzg2Mjc5Nn19LCJSZXNvdXJjZSI6Imh0dHBzO
i8vY2FzLWJyaWRnZS54ZXRodWIuaGYuY28veGV0LWJyaWRnZS11cy82MzUyNmY0YzdlNGNjMzEzNWZkM
GZmMWEvNmFjYjY1M2YzYzA1YjgzOThkMzg2YzllOTYwMjdmMGJkZTBhM2U0MTA5NWViMDI3ODUxNDYwO
Dg5MTMyNGY0OSoifV19&Signature=urvr1rMbFOGKMjS9yDB8POL7g5XPIHckQzk2RWAMTgYeHrcuFM
g8w1sM1v7zhtrP3NwQyEsHu3UfgaBfhc2Vozp1tYjuKCkuwZUEaBSQyIalrkF5W1t12OWhBmslv%7E9r
v78hxZzL6qvoWfN-
x4DdOzg4HnLI0jRf0fBJhiHEWUGCrGyTcOZy%7EQXJPButYOsptPzJqLqWEBRCN31H8doKhnzEaXZeN-
71vf5vfPOD-
yb%7EoOEMymohqDbIPn6ZKTT3OHoiYiq3K2G6fLO9zt%7EdGhEJRBZbw6mGWcF3p2czUZRuCef9k5XqY
ftiljqRkvqa%7EO5c2%7EHkjEaWEn24KaVX9Q__&Key-Pair-Id=K2L8F4GPSG1IFC:
HTTPSConnectionPool(host='cas-bridge.xethub.hf.co', port=443): Read timed out.
Trying to resume download…
```

```
model.safetensors:  63%|######2   | 1.96G/3.13G [00:00<?, ?B/s]
```

```
generation_config.json:   0%|          | 0.00/147 [00:01<?, ?B/s]
```

```
Loaded google/flan-t5-large model.
```

```
/home/daniel/anaconda3/lib/python3.12/site-
packages/langchain_core/_api/deprecation.py:119: LangChainDeprecationWarning:
The method `BaseRetriever.get_relevant_documents` was deprecated in langchain-
core 0.1.46 and will be removed in 0.3.0. Use invoke instead.
  warn_deprecated(
```

```

==================================================
Question: Summarize the key points of this document in a paragraph of 100 words.
--------------------------------------------------
Answer: Hinton shared critical insights about AI’s future in his interview with
the Data and AI course.
==================================================


==================================================
Generic Answer (No Context):
--------------------------------------------------
Answer: There are a number of types of insurance that provide different coverage
for a variety of purposes.
==================================================
```

```
[ ]:
```