```
!pip install sentence-transformers langchain_huggingface langchain_pinecone pinecone-client
```

```
Requirement already satisfied: sentence-transformers in /usr/local/lib/python3.11/dist-packages (3.4.1)
Requirement already satisfied: langchain_huggingface in /usr/local/lib/python3.11/dist-packages (0.1.2)
Requirement already satisfied: langchain_pinecone in /usr/local/lib/python3.11/dist-packages (0.2.3)
Requirement already satisfied: pinecone-client in /usr/local/lib/python3.11/dist-packages (6.0.0)
Requirement already satisfied: transformers<5.0.0,≥4.41.0 in /usr/local/lib/python3.11/dist-packages (from sentence-tra
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
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (1.6
Requirement already satisfied: scipy in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (1.14.1)
Requirement already satisfied: huggingface-hub≥0.20.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transfo
Requirement already satisfied: Pillow in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (11.1.0)
Requirement already satisfied: langchain-core<0.4.0,≥0.3.15 in /usr/local/lib/python3.11/dist-packages (from langchain_
Requirement already satisfied: tokenizers≥0.19.1 in /usr/local/lib/python3.11/dist-packages (from langchain_huggingface
Requirement already satisfied: pinecone<6.0.0,≥5.4.0 in /usr/local/lib/python3.11/dist-packages (from langchain_pinecon
Requirement already satisfied: aiohttp<3.11,≥3.10 in /usr/local/lib/python3.11/dist-packages (from langchain_pinecone)
Requirement already satisfied: numpy<2.0.0,≥1.26.4 in /usr/local/lib/python3.11/dist-packages (from langchain_pinecone
Requirement already satisfied: langchain-tests<1.0.0,≥0.3.7 in /usr/local/lib/python3.11/dist-packages (from langchain_
Requirement already satisfied: certifi≥2019.11.17 in /usr/local/lib/python3.11/dist-packages (from pinecone-client) (20
Requirement already satisfied: pinecone-plugin-interface<0.0.8,≥0.0.7 in /usr/local/lib/python3.11/dist-packages (from
Requirement already satisfied: python-dateutil≥2.5.3 in /usr/local/lib/python3.11/dist-packages (from pinecone-client)
Requirement already satisfied: typing-extensions≥3.7.4 in /usr/local/lib/python3.11/dist-packages (from pinecone-client
Requirement already satisfied: urllib3≥1.26.0 in /usr/local/lib/python3.11/dist-packages (from pinecone-client) (2.3.0)
Requirement already satisfied: aiohappyeyeballs≥2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<3.11,≥3
Requirement already satisfied: aiosignal≥1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp<3.11,≥3.10→lan
Requirement already satisfied: attrs≥17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<3.11,≥3.10→langch
Requirement already satisfied: frozenlist≥1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp<3.11,≥3.10→la
Requirement already satisfied: multidict<7.0,≥4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp<3.11,≥3.10→
Requirement already satisfied: yarl<2.0,≥1.12.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp<3.11,≥3.10→la
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub≥0.20.0→senten
Requirement already satisfied: fsspec≥2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub≥0.20.0
Requirement already satisfied: packaging≥20.9 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub≥0.20.0→
Requirement already satisfied: pyyaml≥5.1 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub≥0.20.0→sent
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from huggingface-hub≥0.20.0→senten
Requirement already satisfied: langsmith<0.4,≥0.1.125 in /usr/local/lib/python3.11/dist-packages (from langchain-core<0
Requirement already satisfied: tenacity≠8.4.0,<10.0.0,≥8.1.0 in /usr/local/lib/python3.11/dist-packages (from langchai
Requirement already satisfied: jsonpatch<2.0,≥1.33 in /usr/local/lib/python3.11/dist-packages (from langchain-core<0.4
Requirement already satisfied: pydantic<3.0.0,≥2.5.2 in /usr/local/lib/python3.11/dist-packages (from langchain-core<0
Requirement already satisfied: pytest<9,≥7 in /usr/local/lib/python3.11/dist-packages (from langchain-tests<1.0.0,≥0.3
Requirement already satisfied: pytest-asyncio<1,≥0.20 in /usr/local/lib/python3.11/dist-packages (from langchain-tests<
Requirement already satisfied: httpx<1,≥0.25.0 in /usr/local/lib/python3.11/dist-packages (from langchain-tests<1.0.0,
Requirement already satisfied: syrupy<5,≥4 in /usr/local/lib/python3.11/dist-packages (from langchain-tests<1.0.0,≥0.3
Requirement already satisfied: pytest-socket<1,≥0.6.0 in /usr/local/lib/python3.11/dist-packages (from langchain-tests<
Requirement already satisfied: pinecone-plugin-inference<4.0.0,≥2.0.0 in /usr/local/lib/python3.11/dist-packages (from
Requirement already satisfied: six≥1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil≥2.5.3→pinecon
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch≥1.11.0→sentence-transfo
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from torch≥1.11.0→sentence-transform
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch≥
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch≥
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch≥1.11
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch≥1.11
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch≥1.11
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch≥1
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch≥1
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch≥
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch≥1.1
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch≥1.11.0→
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch≥1.11.0
```

```
from sentence_transformers import CrossEncoder
from pinecone import Pinecone
from langchain_huggingface import HuggingFaceEmbeddings

import os
os.environ['PINECONE_API_KEY'] = "pcsk_5qAp4G_QyLtWLYcjNsy23CDhhgkH4cAvesEhYHMKWTFkC32i8SyzjrDLsWhnLPcWS97PUm"
pinecone_api_key = os.getenv("PINECONE_API_KEY")

pc= Pinecone(api_key= pinecone_api_key)
index = pc.Index("profilestore")

embedder = HuggingFaceEmbeddings(model_name = 'BAAI/bge-base-en-v1.5')
```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94: 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),
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.
  warnings.warn(
modules.json: 100%  349/349 [00:00<00:00, 30.8kB/s]
config_sentence_transformers.json: 100%  124/124 [00:00<00:00, 10.3kB/s]
README.md: 94.6k/? [00:00<00:00, 6.17MB/s]
sentence_bert_config.json: 100%  52.0/52.0 [00:00<00:00, 2.64kB/s]
config.json: 100%  777/777 [00:00<00:00, 39.1kB/s]
model.safetensors: 100%  438M/438M [00:01<00:00, 327MB/s]
tokenizer_config.json: 100%  366/366 [00:00<00:00, 41.4kB/s]
vocab.txt: 232k/? [00:00<00:00, 14.2MB/s]
tokenizer.json: 711k/? [00:00<00:00, 22.9MB/s]
special_tokens_map.json: 100%  125/125 [00:00<00:00, 8.37kB/s]
config.json: 100%  190/190 [00:00<00:00, 18.5kB/s]
```

```
from google.colab import drive
drive.mount('/content/drive')

folder_path = "/content/drive/My Drive/GenAI/"
```

```
Mounted at /content/drive
```

```
# ═══════════  Generate Test Data ═══════════

import random
import pandas as pd

# Load Excel file
df1 = pd.read_excel(folder_path + "testdata1.xlsx")
df2 = pd.read_excel(folder_path + "testdata2.xlsx")

df= pd.concat([df1, df2], ignore_index=True)

# Create a copy of answers and shuffle them
shuffled_answers = df["Answer"].sample(frac=1, random_state=42).reset_index(drop=True)

# Assign relevance scores: 1.0 for correct pairs, 0.0 for incorrect (shuffled) pairs
positive_samples = list(zip(df["Question"], df["Answer"], [1.0] * len(df)))
negative_samples = list(zip(df["Question"], shuffled_answers, [0.0] * len(df)))

# Combine positive & negative samples
train_data = positive_samples + negative_samples

# Shuffle training data
random.shuffle(train_data)

# Store in excel
df_train = pd.DataFrame(train_data, columns=["Question", "Answer", "Relevance"])
df_train.to_excel(folder_path + "train_data.xlsx", index=False)


from torch.utils.data import DataLoader
from sentence_transformers import InputExample

# Prepare data for training
train_samples = [InputExample(texts=[str(q), str(a)], label=score) for q, a, score in train_data]

model = CrossEncoder('cross-encoder/ms-marco-MiniLM-L-6-v2')

# Convert to DataLoader (batch_size set here)
train_dataloader = DataLoader(train_samples, batch_size=8, shuffle=True)

# Train the model
model.fit(
    train_dataloader=train_dataloader,
```

```
    epochs=3, # CHANGE LATER
    warmup_steps=100
)

model.save(folder_path + "CrossEncoderModel")
```

```
Epoch: 100%  3/3 [05:06<00:00, 102.33s/it]
Iteration: 100%  1352/1352 [01:41<00:00, 13.66it/s]
Iteration: 100%  1352/1352 [01:42<00:00, 13.75it/s]
Iteration: 100%  1352/1352 [01:42<00:00, 12.51it/s]
```

```
# Step 1: Retrieve Top-50 Results from Pinecone (Vector Search)
query = "What does 'Schedule A' determine?"

query_vector = embedder.embed_query(query)
pinecone_results = index.query(vector=query_vector, top_k=500, include_metadata=True)
print(pinecone_results["matches"])

# Step 2: Use a Cross-Encoder for Reranking
cross_encoder = model

# Prepare query-chunk pairs
query_chunk_pairs = [(query, "".join(doc["metadata"]["text"])) for doc in pinecone_results["matches"]]
print(query_chunk_pairs)

cross_encoder_scores = cross_encoder.predict(query_chunk_pairs)

# Step 3: Sort by Cross-Encoder Scores
reranked_results = sorted(
    zip(pinecone_results["matches"], cross_encoder_scores),
    key=lambda x: x[1], reverse=True
)[:10]  # Take top 10

# Output Final Ranked Results
for res, score in reranked_results:
    print(f"Chunk: {res['metadata']['text']}, Score: {score}")
```