```
!pip install openai==0.28
!pip install tiktoken
```

```
Collecting openai==0.28
  Downloading openai-0.28.0-py3-none-any.whl.metadata (13 kB)
Requirement already satisfied: requests>=2.20 in /usr/local/lib/python3.11/dist-packages (from openai==0.28) (2.32.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from openai==0.28) (4.67.1)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (from openai==0.28) (3.11.15)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.20->openai==0.28) (
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.20->openai==0.28) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.20->openai==0.28) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.20->openai==0.28) (2025.1
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->openai==0.28) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp->openai==0.28) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->openai==0.28) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp->openai==0.28) (1.6.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp->openai==0.28) (6.4.3)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->openai==0.28) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->openai==0.28) (1.20.0)
Downloading openai-0.28.0-py3-none-any.whl (76 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 76.5/76.5 kB 5.7 MB/s eta 0:00:00
Installing collected packages: openai
  Attempting uninstall: openai
    Found existing installation: openai 1.76.0
    Uninstalling openai-1.76.0:
      Successfully uninstalled openai-1.76.0
Successfully installed openai-0.28.0
Collecting tiktoken
  Downloading tiktoken-0.9.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (6.7 kB)
Requirement already satisfied: regex>=2022.1.18 in /usr/local/lib/python3.11/dist-packages (from tiktoken) (2024.11.6)
Requirement already satisfied: requests>=2.26.0 in /usr/local/lib/python3.11/dist-packages (from tiktoken) (2.32.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.26.0->tiktoken) (3.
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.26.0->tiktoken) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.26.0->tiktoken) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.26.0->tiktoken) (2025.1.3
Downloading tiktoken-0.9.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.2 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.2/1.2 MB 47.4 MB/s eta 0:00:00
Installing collected packages: tiktoken
Successfully installed tiktoken-0.9.0
```

```
!pip install datasets
```

```
Requirement already satisfied: requests>=2.32.2 in /usr/local/lib/python3.11/dist-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.11/dist-packages (from datasets) (4.67.1)
Collecting xxhash (from datasets)
  Downloading xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2 kB)
Collecting fsspec<=2025.3.0,>=2023.1.0 (from fsspec[http]<=2025.3.0,>=2023.1.0->datasets)
  Downloading fsspec-2025.3.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (from datasets) (3.11.15)
Requirement already satisfied: huggingface-hub>=0.24.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.30.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from datasets) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from datasets) (6.0.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.6.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.4.3)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.20.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.24.0->d
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets) (2.4.0
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets) (2025.
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.9.0.post0
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->datasets) (1
Downloading datasets-3.5.1-py3-none-any.whl (491 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 491.4/491.4 kB 35.2 MB/s eta 0:00:00
```

```
Installing collected packages: xxhash, fsspec, dill, multiprocess, datasets
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2025.3.2
    Uninstalling fsspec-2025.3.2:
      Successfully uninstalled fsspec-2025.3.2
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the sour
torch 2.6.0+cu124 requires nvidia-cublas-cu12==12.4.5.8; platform_system == "Linux" and platform_machine == "x86_64", but you have nv
torch 2.6.0+cu124 requires nvidia-cuda-cupti-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you hav
torch 2.6.0+cu124 requires nvidia-cuda-nvrtc-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you hav
torch 2.6.0+cu124 requires nvidia-cuda-runtime-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you h
torch 2.6.0+cu124 requires nvidia-cudnn-cu12==9.1.0.70; platform_system == "Linux" and platform_machine == "x86_64", but you have nvi
torch 2.6.0+cu124 requires nvidia-cufft-cu12==11.2.1.3; platform_system == "Linux" and platform_machine == "x86_64", but you have nvi
torch 2.6.0+cu124 requires nvidia-curand-cu12==10.3.5.147; platform_system == "Linux" and platform_machine == "x86_64", but you have
torch 2.6.0+cu124 requires nvidia-cusolver-cu12==11.6.1.9; platform_system == "Linux" and platform_machine == "x86_64", but you have
torch 2.6.0+cu124 requires nvidia-cusparse-cu12==12.3.1.170; platform_system == "Linux" and platform_machine == "x86_64", but you hav
torch 2.6.0+cu124 requires nvidia-nvjitlink-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have
gcsfs 2025.3.2 requires fsspec==2025.3.2, but you have fsspec 2025.3.0 which is incompatible.
Successfully installed datasets-3.5.1 dill-0.3.8 fsspec-2025.3.0 multiprocess-0.70.16 xxhash-3.5.0
```

```
import os
import time
import tiktoken
import openai
from datasets import load_dataset


# === Setup ===
print("[LongRAG Demo] Loading test corpus...")
corpus = load_dataset("TIGER-Lab/LongRAG", "nq_corpus", split="train[:5%]")

# === Retrieve a long unit (simulate LongRAG unit) ===
long_units = [doc['text'] for doc in corpus if 'text' in doc][:4]  # simulate top-4 retrievals
long_context = "\n\n".join(long_units)
print(f"[Info] Combined context from {len(long_units)} units.")
```

```
[LongRAG Demo] Loading test corpus...
Resolving data files: 100%                                  25/25 [00:00<00:00, 86.02it/s]
Downloading data: 100%                                      25/25 [04:14<00:00, 16.11s/files]
train-00006-of-00025.parquet: 100%                          280M/280M [00:11<00:00, 26.3MB/s]
train-00007-of-00025.parquet: 100%                          279M/279M [00:11<00:00, 25.2MB/s]
train-00008-of-00025.parquet: 100%                          279M/279M [00:11<00:00, 25.2MB/s]
train-00009-of-00025.parquet: 100%                          278M/278M [00:11<00:00, 24.7MB/s]
train-00010-of-00025.parquet: 100%                          278M/278M [00:11<00:00, 23.4MB/s]
train-00011-of-00025.parquet: 100%                          277M/277M [00:13<00:00, 23.6MB/s]
train-00012-of-00025.parquet: 100%                          276M/276M [00:12<00:00, 23.0MB/s]
train-00013-of-00025.parquet: 100%                          276M/276M [00:12<00:00, 22.0MB/s]
train-00014-of-00025.parquet: 100%                          275M/275M [00:11<00:00, 21.9MB/s]
train-00015-of-00025.parquet: 100%                          276M/276M [00:12<00:00, 22.3MB/s]
train-00016-of-00025.parquet: 100%                          275M/275M [00:12<00:00, 21.0MB/s]
train-00017-of-00025.parquet: 100%                          277M/277M [00:11<00:00, 23.6MB/s]
train-00018-of-00025.parquet: 100%                          278M/278M [00:11<00:00, 25.0MB/s]
train-00019-of-00025.parquet: 100%                          284M/284M [00:11<00:00, 24.4MB/s]
train-00020-of-00025.parquet: 100%                          293M/293M [00:15<00:00, 19.6MB/s]
train-00021-of-00025.parquet: 100%                          306M/306M [00:12<00:00, 24.3MB/s]
train-00022-of-00025.parquet: 100%                          328M/328M [00:13<00:00, 23.6MB/s]
train-00023-of-00025.parquet: 100%                          362M/362M [00:17<00:00, 20.5MB/s]
train-00024-of-00025.parquet: 100%                          442M/442M [00:18<00:00, 24.2MB/s]
Generating train split: 100%                                604351/604351 [01:48<00:00, 6460.65 examples/s]
```

```
# === Token management (Enforce 1028 token limit) ===
encoding = tiktoken.get_encoding("cl100k_base")
```

```
context_tokens = len(encoding.encode(long_context))

MAX_TOTAL_TOKENS = 1028
MAX_OUTPUT_TOKENS = 256

if context_tokens + MAX_OUTPUT_TOKENS > MAX_TOTAL_TOKENS:
    token_budget = MAX_TOTAL_TOKENS - MAX_OUTPUT_TOKENS
    print(f"[Truncate] Context tokens before truncation: {context_tokens}")
    words = long_context.split()
    while len(encoding.encode(" ".join(words))) > token_budget:
        words = words[:-100]
    long_context = " ".join(words)
    context_tokens = len(encoding.encode(long_context))
    print(f"[Truncate] Truncated context tokens: {context_tokens}")

# === User query ===
query = "What is the main idea covered in these documents?"

# === OpenAI Completion ===
openai.api_key = "key"

print("[Query] Sending request to OpenAI...")
start = time.time()
response = openai.ChatCompletion.create(
    model="gpt-4o",
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": f"{long_context}\n\nQuestion: {query}"}
    ],
    max_tokens=MAX_TOTAL_TOKENS - context_tokens,
    temperature=0
)
end = time.time()

output = response['choices'][0]['message']['content']
print("\n=== Response ===")
print(output)
```

```
[Truncate] Context tokens before truncation: 1101
[Truncate] Truncated context tokens: 716
[Query] Sending request to OpenAI...

=== Response ===
The documents cover three distinct topics:

1. **Pulse Foundation**: This document provides an overview of the Pulse Foundation, a Bulgarian non-governmental organization. It detai

2. **Sandy Stewart (coach)**: This document outlines the career of Sandy Stewart, a college volleyball coach who led the Iowa Hawkeyes w

3. **Helsinki Saturday**: This document describes "Helsinki Saturday," a live album by Jandek released in 2009. It details the recording

Overall, the main idea covered in these documents is the description of different entities and individuals, highlighting their activitie
```

```
# === Benchmarking ===
def evaluate(response, keywords):
    score = {
        "coherence": 1.0 if response[0].isupper() and response.strip().endswith('.') else 0.5,
        "relevance": 1.0 if any(k in response.lower() for k in keywords) else 0.5,
        "efficiency": 1.0 if len(response.split()) <= 120 else 0.7
    }
    score["overall"] = round(sum(score.values()) / 3, 2)
    return score

reference_keywords = ["article", "person", "history", "politics", "science", "explains", "research", "data"]
results = evaluate(output, reference_keywords)

print("\n=== Evaluation Scores ===")
for key, val in results.items():
    print(f"{key.capitalize()}: {val}")
print(f"Elapsed time: {round(end - start, 2)} seconds")
```

```
=== Evaluation Scores ===
Coherence: 1.0
Relevance: 1.0
Efficiency: 0.7
Overall: 0.9
```

```
Elapsed time: 4.46 seconds
```