```
!nvidia-smi
```

```
Thu Jul  4 17:24:07 2024
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.104.05             Driver Version: 535.104.05   CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  Tesla T4                       Off | 00000000:00:04.0 Off |                    0 |
| N/A   51C    P8              9W /  70W |      0MiB / 15360MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|  No running processes found                                                           |
+---------------------------------------------------------------------------------------+
```

```
!pip install -Uqqq pip --progress-bar off
!pip install -qqq torch==2.0.1
!pip install -qqq transformers
!pip install -qqq langchain
!pip install -qqq chromadb
!pip install -qqq pypdf
!pip install -qqq xformers==0.0.20
!pip install -qqq sentence_transformers
!pip install -qqq InstructorEmbedding
!pip install -qqq pdf2image
```

```
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 54.6/54.6 MB 12.1 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 102.6/102.6 MB 8.3 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 173.2/173.2 MB 7.8 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 177.1/177.1 MB 6.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 98.6/98.6 kB 8.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 63.3/63.3 MB 12.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 96.4/96.4 kB 8.6 MB/s eta 0:00:00
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviou
torchaudio 2.3.0+cu121 requires torch==2.3.0, but you have torch 2.0.1 which is incompatible.
torchtext 0.18.0 requires torch>=2.3.0, but you have torch 2.0.1 which is incompatible.
torchvision 0.18.0+cu121 requires torch==2.3.0, but you have torch 2.0.1 which is incompatible.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 50.4/50.4 kB 3.2 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 975.5/975.5 kB 4.0 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 337.4/337.4 kB 22.9 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.5/127.5 kB 11.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 141.1/141.1 kB 12.8 MB/s eta 0:00:00
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 67.3/67.3 kB 6.3 MB/s eta 0:00:00
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 559.5/559.5 kB 37.2 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.4/2.4 MB 78.3 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 283.7/283.7 kB 24.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 92.0/92.0 kB 8.6 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 75.6/75.6 kB 6.9 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 77.9/77.9 kB 7.6 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.7/1.7 MB 79.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 67.6/67.6 kB 6.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.8/6.8 MB 114.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 59.9/59.9 kB 5.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 52.5/52.5 kB 4.9 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 130.5/130.5 kB 12.3 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 107.0/107.0 kB 10.1 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 41.3/41.3 kB 3.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.4/62.4 kB 5.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 58.3/58.3 kB 5.2 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 341.4/341.4 kB 27.0 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 71.9/71.9 kB 6.9 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 53.6/53.6 kB 4.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.4/3.4 MB 89.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.2/1.2 MB 56.6 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 130.2/130.2 kB 12.1 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 46.0/46.0 kB 3.8 MB/s eta 0:00:00
```

```
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 227.1/227.1 kB 15.8 MB/s eta 0:00:00
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
```

```
!pip install langchain_community
```

```
Collecting langchain_community
  Downloading langchain_community-0.2.6-py3-none-any.whl.metadata (2.5 kB)
Requirement already satisfied: PyYAML>=5.3 in /usr/local/lib/python3.10/dist-packages (from langchain_community) (6.0.1)
Requirement already satisfied: SQLAlchemy<3,>=1.4 in /usr/local/lib/python3.10/dist-packages (from langchain_community)
Requirement already satisfied: aiohttp<4.0.0,>=3.8.3 in /usr/local/lib/python3.10/dist-packages (from langchain_communit
Collecting dataclasses-json<0.7,>=0.5.7 (from langchain_community)
  Downloading dataclasses_json-0.6.7-py3-none-any.whl.metadata (25 kB)
Requirement already satisfied: langchain<0.3.0,>=0.2.6 in /usr/local/lib/python3.10/dist-packages (from langchain_commun
Requirement already satisfied: langchain-core<0.3.0,>=0.2.10 in /usr/local/lib/python3.10/dist-packages (from langchain_
Requirement already satisfied: langsmith<0.2.0,>=0.1.0 in /usr/local/lib/python3.10/dist-packages (from langchain_commun
Requirement already satisfied: numpy<2,>=1 in /usr/local/lib/python3.10/dist-packages (from langchain_community) (1.25.2
Requirement already satisfied: requests<3,>=2 in /usr/local/lib/python3.10/dist-packages (from langchain_community) (2.3
Requirement already satisfied: tenacity!=8.4.0,<9.0.0,>=8.1.0 in /usr/local/lib/python3.10/dist-packages (from langchain
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->lan
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3-
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->la
Requirement already satisfied: async-timeout<5.0,>=4.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=
Collecting marshmallow<4.0.0,>=3.18.0 (from dataclasses-json<0.7,>=0.5.7->langchain_community)
  Downloading marshmallow-3.21.3-py3-none-any.whl.metadata (7.1 kB)
Requirement already satisfied: typing-inspect<1,>=0.4.0 in /usr/local/lib/python3.10/dist-packages (from dataclasses-jso
Requirement already satisfied: langchain-text-splitters<0.3.0,>=0.2.0 in /usr/local/lib/python3.10/dist-packages (from l
Requirement already satisfied: pydantic<3,>=1 in /usr/local/lib/python3.10/dist-packages (from langchain<0.3.0,>=0.2.6->
Requirement already satisfied: jsonpatch<2.0,>=1.33 in /usr/local/lib/python3.10/dist-packages (from langchain-core<0.3.
Requirement already satisfied: packaging<25,>=23.2 in /usr/local/lib/python3.10/dist-packages (from langchain-core<0.3.0
Requirement already satisfied: orjson<4.0.0,>=3.9.14 in /usr/local/lib/python3.10/dist-packages (from langsmith<0.2.0,>=
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2-
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2->langchain_c
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2->langc
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2->langc
Requirement already satisfied: typing-extensions>=4.6.0 in /usr/local/lib/python3.10/dist-packages (from SQLAlchemy<3,>=
Requirement already satisfied: greenlet!=0.4.17 in /usr/local/lib/python3.10/dist-packages (from SQLAlchemy<3,>=1.4->lan
Requirement already satisfied: jsonpointer>=1.9 in /usr/local/lib/python3.10/dist-packages (from jsonpatch<2.0,>=1.33->l
Requirement already satisfied: annotated-types>=0.4.0 in /usr/local/lib/python3.10/dist-packages (from pydantic<3,>=1->l
Requirement already satisfied: pydantic-core==2.20.0 in /usr/local/lib/python3.10/dist-packages (from pydantic<3,>=1->la
Requirement already satisfied: mypy-extensions>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from typing-inspect<1,
Downloading langchain_community-0.2.6-py3-none-any.whl (2.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.2/2.2 MB 77.1 MB/s eta 0:00:00
Downloading dataclasses_json-0.6.7-py3-none-any.whl (28 kB)
Downloading marshmallow-3.21.3-py3-none-any.whl (49 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 49.2/49.2 kB 4.7 MB/s eta 0:00:00
Installing collected packages: marshmallow, dataclasses-json, langchain_community
Successfully installed dataclasses-json-0.6.7 langchain_community-0.2.6 marshmallow-3.21.3
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
```

```
!wget -q https://github.com/AutoGPTQ/AutoGPTQ/releases/download/v0.4.2/auto_gptq-0.4.2+cu118-cp310-cp310-linux_x86_64.whl
```

```
!pip install -qqq auto_gptq-0.4.2+cu118-cp310-cp310-linux_x86_64.whl --progress-bar off
```

```
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviou
cudf-cu12 24.4.1 requires pyarrow<15.0.0a0,>=14.0.1, but you have pyarrow 16.1.0 which is incompatible.
google-colab 1.0.0 requires requests==2.31.0, but you have requests 2.32.3 which is incompatible.
ibis-framework 8.0.0 requires pyarrow<16,>=2, but you have pyarrow 16.1.0 which is incompatible.
torchtext 0.18.0 requires torch>=2.3.0, but you have torch 2.0.1 which is incompatible.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
```

```
!sudo apt-get install poppler-utils
```

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  poppler-utils
0 upgraded, 1 newly installed, 0 to remove and 45 not upgraded.
Need to get 186 kB of archives.
After this operation, 696 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 poppler-utils amd64 22.02.0-2ubuntu0.4 [186 kB]
Fetched 186 kB in 1s (147 kB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based frontend cannot be used. at /usr/share/perl5/D
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package poppler-utils.
(Reading database ... 121925 files and directories currently installed.)
Preparing to unpack .../poppler-utils_22.02.0-2ubuntu0.4_amd64.deb ...
```

```
Unpacking poppler-utils (22.02.0-2ubuntu0.4) ...
Setting up poppler-utils (22.02.0-2ubuntu0.4) ...
Processing triggers for man-db (2.10.2-1) ...
```

```
import torch
from auto_gptq import AutoGPTQForCausalLM
from langchain import HuggingFacePipeline, PromptTemplate
from langchain.chains import RetrievalQA
from langchain.document_loaders import PyPDFDirectoryLoader
from langchain.embeddings import HuggingFaceInstructEmbeddings, HuggingFaceEmbeddings
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.vectorstores import Chroma
from pdf2image import convert_from_path
from transformers import AutoTokenizer, TextStreamer, pipeline
from langchain_community.embeddings import SentenceTransformerEmbeddings


DEVICE = "cuda:0" if torch.cuda.is_available() else "cpu"
```

```
/usr/local/lib/python3.10/dist-packages/transformers/utils/generic.py:481: UserWarning: torch.utils._pytree._register_py
  _torch_pytree._register_pytree_node(
/usr/local/lib/python3.10/dist-packages/transformers/utils/generic.py:338: UserWarning: torch.utils._pytree._register_py
  _torch_pytree._register_pytree_node(
```

```
# Data
```

```
!mkdir pdfs
```

```
!ls
```

```
auto_gptq-0.4.2+cu118-cp310-cp310-linux_x86_64.whl  pdfs  sample_data
```

```
!rm auto_gptq-0.4.2+cu117-cp310-cp310-linux_x86_64.whl
```

```
rm: cannot remove 'auto_gptq-0.4.2+cu117-cp310-cp310-linux_x86_64.whl': No such file or directory
```

```
!ls
```

```
auto_gptq-0.4.2+cu118-cp310-cp310-linux_x86_64.whl  pdfs  sample_data
```

```
!rm ./'- The Official Guide to PTE Academic - Teacher Notes.pdf'
```

```
rm: cannot remove './- The Official Guide to PTE Academic - Teacher Notes.pdf': No such file or directory
```

```
from google.colab import files
```

```
uploaded = files.upload()
```

```
Choose files human_right.pdf
• human_right.pdf(application/pdf) - 326613 bytes, last modified: 04/07/2024 - 100% done
Saving human_right.pdf to human_right.pdf
```

```
! mv  human_right.pdf pdfs
```

```
!ls ./pdfs
```

```
human_right.pdf
```

```
!rm ./pdfs/urban_development.pdf
```

```
rm: cannot remove './pdfs/urban_development.pdf': No such file or directory
```

```
loader = PyPDFDirectoryLoader("pdfs")
docs = loader.load()
len(docs)
```

```
49
```

```
!pip install flash_attn
```

```
Requirement already satisfied: flash_attn in /usr/local/lib/python3.10/dist-packages (2.5.9.post1)
Requirement already satisfied: torch in /usr/local/lib/python3.10/dist-packages (from flash_attn) (2.0.1)
Requirement already satisfied: einops in /usr/local/lib/python3.10/dist-packages (from flash_attn) (0.8.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from torch->flash_attn) (3.15.4)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.10/dist-packages (from torch->flash_attn) (4.
```

```
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from torch->flash_attn) (1.12.1)
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-packages (from torch->flash_attn) (3.3)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from torch->flash_attn) (3.1.4)
Requirement already satisfied: nvidia-cuda-nvrtc-cu11==11.7.99 in /usr/local/lib/python3.10/dist-packages (from torch->f
Requirement already satisfied: nvidia-cuda-runtime-cu11==11.7.99 in /usr/local/lib/python3.10/dist-packages (from torch-
Requirement already satisfied: nvidia-cuda-cupti-cu11==11.7.101 in /usr/local/lib/python3.10/dist-packages (from torch->
Requirement already satisfied: nvidia-cudnn-cu11==8.5.0.96 in /usr/local/lib/python3.10/dist-packages (from torch->flash
Requirement already satisfied: nvidia-cublas-cu11==11.10.3.66 in /usr/local/lib/python3.10/dist-packages (from torch->fl
Requirement already satisfied: nvidia-cufft-cu11==10.9.0.58 in /usr/local/lib/python3.10/dist-packages (from torch->flas
Requirement already satisfied: nvidia-curand-cu11==10.2.10.91 in /usr/local/lib/python3.10/dist-packages (from torch->fl
Requirement already satisfied: nvidia-cusolver-cu11==11.4.0.1 in /usr/local/lib/python3.10/dist-packages (from torch->fl
Requirement already satisfied: nvidia-cusparse-cu11==11.7.4.91 in /usr/local/lib/python3.10/dist-packages (from torch->f
Requirement already satisfied: nvidia-nccl-cu11==2.14.3 in /usr/local/lib/python3.10/dist-packages (from torch->flash_at
Requirement already satisfied: nvidia-nvtx-cu11==11.7.91 in /usr/local/lib/python3.10/dist-packages (from torch->flash_a
Requirement already satisfied: triton==2.0.0 in /usr/local/lib/python3.10/dist-packages (from torch->flash_attn) (2.0.0)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from nvidia-cublas-cu11==11.10.3.6
Requirement already satisfied: wheel in /usr/local/lib/python3.10/dist-packages (from nvidia-cublas-cu11==11.10.3.66->to
Requirement already satisfied: cmake in /usr/local/lib/python3.10/dist-packages (from triton==2.0.0->torch->flash_attn)
Requirement already satisfied: lit in /usr/local/lib/python3.10/dist-packages (from triton==2.0.0->torch->flash_attn) (1
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->torch->flash_att
Requirement already satisfied: mpmath<1.4.0,>=1.1.0 in /usr/local/lib/python3.10/dist-packages (from sympy->torch->flash
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system packa
```

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

```
# embeddings = SentenceTransformerEmbeddings(model_name="Alibaba-NLP/gte-large-en-v1.5", model_kwargs={"trust_remote_code":T
embeddings = HuggingFaceEmbeddings(
    model_name="hkunlp/instructor-large", model_kwargs={"device": DEVICE}
)
```

```
/usr/local/lib/python3.10/dist-packages/langchain_core/_api/deprecation.py:139: LangChainDeprecationWarning: The class `
  warn_deprecated(
config_sentence_transformers.json: 100%  122/122 [00:00<00:00, 5.18kB/s]
README.md: 100%  66.3k/66.3k [00:00<00:00, 1.54MB/s]
sentence_bert_config.json: 100%  53.0/53.0 [00:00<00:00, 3.02kB/s]
/usr/local/lib/python3.10/dist-packages/huggingface_hub/file_download.py:1132: FutureWarning: `resume_download` is depre
  warnings.warn(
config.json: 100%  1.53k/1.53k [00:00<00:00, 54.0kB/s]
pytorch_model.bin: 100%  1.34G/1.34G [00:07<00:00, 111MB/s]
tokenizer_config.json: 100%  2.41k/2.41k [00:00<00:00, 195kB/s]
spiece.model: 100%  792k/792k [00:00<00:00, 34.9MB/s]
tokenizer.json: 100%  2.42M/2.42M [00:00<00:00, 5.48MB/s]
special_tokens_map.json: 100%  2.20k/2.20k [00:00<00:00, 153kB/s]
1_Pooling/config.json: 100%  270/270 [00:00<00:00, 11.6kB/s]
pytorch_model.bin: 100%  3.15M/3.15M [00:00<00:00, 46.0MB/s]
2_Dense/config.json: 100%  116/116 [00:00<00:00, 4.43kB/s]
```

```
text_splitter = RecursiveCharacterTextSplitter(chunk_size=1024, chunk_overlap=64)
texts = text_splitter.split_documents(docs)
len(texts)
```

```
96
```

```
!rm -rf "dbc"
```

```
%%time
db = Chroma.from_documents(texts, embeddings, persist_directory="dbcd")
```

```
CPU times: user 6.37 s, sys: 196 ms, total: 6.57 s
Wall time: 7.73 s
```

```
## LLAMA
```

```
model_name_or_path = "TheBloke/Llama-2-13B-chat-GPTQ"
model_basename = "model"

tokenizer = AutoTokenizer.from_pretrained(model_name_or_path, use_fast=True)

model = AutoGPTQForCausalLM.from_quantized(
    model_name_or_path,
    revision="gptq-4bit-128g-actorder_True",
```

```
    model_basename=model_basename,
    use_safetensors=True,
    trust_remote_code=True,
    inject_fused_attention=False,
    device=DEVICE,
    quantize_config=None,
)
```

```
tokenizer_config.json: 100%  727/727 [00:00<00:00, 17.9kB/s]
tokenizer.model: 100%  500k/500k [00:00<00:00, 12.4MB/s]
tokenizer.json: 100%  1.84M/1.84M [00:00<00:00, 4.31MB/s]
special_tokens_map.json: 100%  411/411 [00:00<00:00, 9.18kB/s]
You are using the default legacy behaviour of the <class 'transformers.models
config.json: 100%  837/837 [00:00<00:00, 28.5kB/s]
WARNING:auto_gptq.modeling._base:Exllama kernel is not installed, reset disab
WARNING:auto_gptq.modeling._base:CUDA kernels for auto_gptq are not installed
1. You disabled CUDA extensions compilation by setting BUILD_CUDA_EXT=0 when
2. You are using pytorch without CUDA support.
3. CUDA and nvcc are not installed in your device.
config.json: 100%  761/761 [00:00<00:00, 17.6kB/s]
quantize_config.json: 100%  187/187 [00:00<00:00, 4.72kB/s]
model.safetensors: 100%  7.26G/7.26G [01:01<00:00, 208MB/s]
WARNING:auto_gptq.nn_modules.qlinear.qlinear_cuda:CUDA extension not installe
/usr/local/lib/python3.10/dist-packages/transformers/modeling_utils.py:4481:
  warnings.warn(
WARNING:accelerate.utils.modeling:Some weights of the model checkpoint at /ro
WARNING:auto_gptq.nn_modules.fused_llama_mlp:skip module injection for FusedL
```

```
!nvidia-smi
```

```
Thu Jul  4 17:40:20 2024
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.104.05             Driver Version: 535.104.05   CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  Tesla T4                       Off | 00000000:00:04.0 Off |                    0 |
| N/A   48C    P0              26W /  70W |   8509MiB / 15360MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
+---------------------------------------------------------------------------------------+
```

```
DEFAULT_SYSTEM_PROMPT = """
You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers s

If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct.
""".strip()


def generate_prompt(prompt: str, system_prompt: str = DEFAULT_SYSTEM_PROMPT) -> str:
    return f"""
[INST] <>
{system_prompt}
<>

{prompt} [/INST]
""".strip()
```

```
streamer = TextStreamer(tokenizer, skip_prompt=True, skip_special_tokens=True)
```

```
text_pipeline = pipeline(
    "text-generation",
    model=model,
    tokenizer=tokenizer,
    max_new_tokens=1024,
    temperature=0,
    top_p=0.95,
    repetition_penalty=1.15,
    streamer=streamer,
)
```

```
loomForCausalLM', 'CamembertForCausalLM', 'LlamaForCausalLM', 'CodeGenForCausalLM', 'CohereForCausalLM', 'CpmAntForCausal
```

```
llm = HuggingFacePipeline(pipeline=text_pipeline, model_kwargs={"temperature": 0})
```

```
/usr/local/lib/python3.10/dist-packages/langchain_core/_api/deprecation.py:139: LangChainDeprecationWarning: The class `
  warn_deprecated(
```

```
SYSTEM_PROMPT = "Use the following pieces of context to answer the question at the end. If you don't know the answer, just s

template = generate_prompt(
    """
{context}

Question: {question}
""",
    system_prompt=SYSTEM_PROMPT,
)
```

```
prompt = PromptTemplate(template=template, input_variables=["context", "question"])
```

```
qa_chain = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="stuff",
    retriever=db.as_retriever(search_kwargs={"k": 2}),
    return_source_documents=True,
    chain_type_kwargs={"prompt": prompt},
)
```

```
result = qa_chain("In which year was the same sex marriage recognized?")
```

```
/usr/local/lib/python3.10/dist-packages/langchain_core/_api/deprecation.py:139: LangChainDeprecationWarning: The method
  warn_deprecated(
/usr/local/lib/python3.10/dist-packages/transformers/generation/configuration_utils.py:515: UserWarning: `do_sample` is
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/transformers/generation/configuration_utils.py:520: UserWarning: `do_sample` is
  warnings.warn(
 Based on the information provided, same-sex marriage was recognized in Nepal in 2023. This is indicated by the fact tha
```

```
len(result["source_documents"])
```

```
2
```

```
print(result["source_documents"][0].page_content)
```

```
actors based on sexual orientation, gender identity or expression, or sex
characteristics.  Constitutional and legal frameworks outline d protections for
LGBTQI+ persons, but according to LGBTQI+ NGOs  like the country ’s Blue
Diamond Society, there were gaps between those frameworks and
implementation of existing laws.
On March 22, the Supreme Court issued instructions to the government to
recognize and register same- sex marriage documents as legally valid.  The
verdict also issued instructions to the government to review and amend 23
```

```
result = qa_chain("How many Bhutanese refugee live in Nepal?")
```

```
/usr/local/lib/python3.10/dist-packages/transformers/generation/configuration_utils.py:515: UserWarning: `do_sample` is
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/transformers/generation/configuration_utils.py:520: UserWarning: `do_sample` is
  warnings.warn(
 Based on the information provided, I cannot provide an exact number of Bhutanese refugees living in Nepal. However, acc
```

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