```
!nvidia-smi
```

```
Sat Feb 17 20:36:13 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   38C    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                                                           |
+---------------------------------------------------------------------------------------+
```

Double-click (or enter) to edit

```
!pip install -Uqqq pip --progress-bar off
!pip install -qqq torch==2.0.1 --progress-bar off
!pip install -qqq transformers==4.31.0 --progress-bar off
!pip install -qqq langchain==0.0.266 --progress-bar off
!pip install -qqq chromadb==0.4.5 --progress-bar off
!pip install -qqq pypdf==3.15.0 --progress-bar off
!pip install -qqq xformers==0.0.20 --progress-bar off
!pip install -qqq sentence_transformers==2.2.2 --progress-bar off
!pip install -qqq InstructorEmbedding==1.0.1 --progress-bar off
!pip install -qqq pdf2image==1.16.3 --progress-bar off
```

```
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Installing backend dependencies ... done
  Preparing metadata (pyproject.toml) ... done
  Building wheel for lit (pyproject.toml) ... done
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source
torchaudio 2.1.0+cu121 requires torch==2.1.0, but you have torch 2.0.1 which is incompatible.
torchdata 0.7.0 requires torch==2.1.0, but you have torch 2.0.1 which is incompatible.
torchtext 0.16.0 requires torch==2.1.0, but you have torch 2.0.1 which is incompatible.
torchvision 0.16.0+cu121 requires torch==2.1.0, 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 package manager. It
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source
lida 0.0.10 requires fastapi, which is not installed.
lida 0.0.10 requires kaleido, which is not installed.
lida 0.0.10 requires python-multipart, which is not installed.
lida 0.0.10 requires uvicorn, which is not installed.
llmx 0.0.15a0 requires cohere, which is not installed.
llmx 0.0.15a0 requires openai, which is not installed.
llmx 0.0.15a0 requires tiktoken, which is not installed.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Installing backend dependencies ... done
  Preparing metadata (pyproject.toml) ... done
  Building wheel for chroma-hnswlib (pyproject.toml) ... done
  Building wheel for pypika (pyproject.toml) ... done
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source
lida 0.0.10 requires kaleido, which is not installed.
lida 0.0.10 requires python-multipart, which is not installed.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It
  Preparing metadata (setup.py) ... done
  Building wheel for sentence_transformers (setup.py) ... done
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source
xformers 0.0.20 requires torch==2.0.1, but you have torch 2.1.0 which is incompatible.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It
```

Double-click (or enter) to edit

```
!wget -q https://github.com/PanQiWei/AutoGPTQ/releases/download/v0.4.1/auto_gptq-0.4.1+cu118-cp310-cp310-linux_x86_64.whl
```

```
!pip install -qqq auto_gptq-0.4.1+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 behaviour is the source
ibis-framework 7.1.0 requires pyarrow<15,>=2, but you have pyarrow 15.0.0 which is incompatible.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It
```

```
!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 33 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 22.02.0-2ubuntu0.3 [186 kB]
Fetched 186 kB in 0s (680 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/Debconf/FrontEnd
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 ... 121749 files and directories currently installed.)
Preparing to unpack .../poppler-utils_22.02.0-2ubuntu0.3_amd64.deb ...
Unpacking poppler-utils (22.02.0-2ubuntu0.3) ...
Setting up poppler-utils (22.02.0-2ubuntu0.3) ...
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
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.vectorstores import Chroma
from pdf2image import convert_from_path
from transformers import AutoTokenizer, TextStreamer, pipeline, AutoModelForQuestionAnswering

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

```
!rm -rf "db"
```

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

```
1382
```

```
embeddings = HuggingFaceInstructEmbeddings(
    model_name="hkunlp/instructor-large", model_kwargs={"device": DEVICE}
)
```

```
load INSTRUCTOR_Transformer
max_seq_length  512
```

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

```
1611
```

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

```
CPU times: user 2min 25s, sys: 1.32 s, total: 2min 26s
Wall time: 2min 49s
```

```
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,
    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, 33.9kB/s]
tokenizer.model: 100%  500k/500k [00:00<00:00, 569kB/s]
tokenizer.json: 100%  1.84M/1.84M [00:00<00:00, 18.6MB/s]
special_tokens_map.json: 100%  411/411 [00:00<00:00, 19.8kB/s]
config.json: 100%  837/837 [00:00<00:00, 53.7kB/s]
WARNING:auto_gptq.modeling._base:Exllama kernel is not installed, reset disable_exllama
WARNING:auto_gptq.modeling._base:CUDA kernels for auto_gptq are not installed, this wil
1. You disabled CUDA extensions compilation by setting BUILD_CUDA_EXT=0 when install au
2. You are using pytorch without CUDA support.
3. CUDA and nvcc are not installed in your device.
quantize_config.json: 100%  188/188 [00:00<00:00, 11.0kB/s]
model.safetensors: 100%  7.26G/7.26G [02:10<00:00, 86.5MB/s]
WARNING:auto_gptq.nn_modules.qlinear.qlinear_cuda_old:CUDA extension not installed.
WARNING:auto_gptq.nn_modules.fused_llama_mlp:skip module injection for FusedLlamaMLPFor
```

```
# from transformers import AutoTokenizer, AutoModelForQuestionAnswering
# from transformers import pipeline, RobertaForCausalLM, RobertaTokenizer

# model_name_or_path = "deepset/roberta-base-squad2"
# model_basename = "model"

# tokenizer = RobertaTokenizer.from_pretrained(model_name_or_path, use_fast=True)

# model = RobertaForCausalLM.from_pretrained(model_name_or_path)
```

```
DEFAULT_SYSTEM_PROMPT = """
You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not inclu

If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't k
""".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,

)
```

```
The model 'LlamaGPTQForCausalLM' is not supported for text-generation. Supported models are ['BartForCausalLM', 'BertLMHeadModel', 'Ber
```

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

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

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},
)
```

Double-click (or enter) to edit

```
result = qa_chain("In what year was 8086 microprocessor developed?")
```

```
Based on the information provided in Sec. 6.7 of "The 8088 Microprocessor: Programming, Interfacing, Software, Hardware, and Applicati
```

```
result = qa_chain("what are the types of micro-processor")
```

```
Based on the information provided in the text, there are five microprocessors and five coprocessors in the Intel family of 16-bit micr

1. 8086 microprocessor
2. 8087 numeric data processor (NDP)
3. 8089 HO processor (IOP)

Therefore, the types of microprocessors are:

1. 8086 microprocessor
2. 8087 NDP
3. 8089 IOP
```

Double-click (or enter) to edit

```
result = qa_chain("Explain basic stored program computer.")
```

```
808 [INST] <>
Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don'
```

```
<>


Interestingly enough, virtually all computers today are still based on the von Neumann
architecture and stored program concept.
Fetch and Execute
Figure 1.2 is a block diagram of a basic stored program computer. There are three major
parts to this system: (1) the central processing unit (CPU), which acts as the “brain”
coordinating all activities; (2) the memory unit, where the program instructions and data
are temporarily stored; and (3) the input/output (I/O) devices, which allow the computer
to input information for processing and then output the result.
The basic timing of the computer is controlled by a square-wave oscillator or clock
generator circuit. This signal is used to synchronize all activities within the computer,
and it determines how fast instructions can be fetched from memory and executed.
The basic processing cycle begins with a memory fetch or read cycle. The instruction
pointer (IP) register (also called the program counter) holds the address of the memory

to fetch its instructions from memory instead of being rewired for each new program.
Sec. 1.2 Stored Program Computer 3

Question: Explain basic stored program computer.
 [/INST]  Based on the provided context, I cannot answer the question because there is no question presented. The text provides a descr



result = qa_chain("Explain foodchain on earth")

 I cannot explain the concept of a "foodchain" because it does not appear in any of the provided texts or contexts. The terms mentioned
```