```
"Welcome to Meta LLAMA 3"
```

```
'Welcome to Meta LLAMA 3'
```

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

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

```
!nvidia-smi
```

```
Sun Jul 14 06:45:11 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. E
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute
|                                         |                      |               MIG
|=========================================+======================+======================
|   0  Tesla T4                       Off | 00000000:00:04.0 Off |
| N/A   69C    P0              29W /  70W |   6071MiB / 15360MiB |      0%      Defau
|                                         |                      |                  N
+-----------------------------------------+----------------------+----------------------

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

```
!pip install -r requirement.txt
```

```
  Downloading transformers-4.40.0-py3-none-any.whl (9.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 9.0/9.0 MB 68.8 MB/s eta 0:00:00
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-p
Requirement already satisfied: psutil in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: pyyaml in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: torch>=1.10.0 in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: huggingface-hub in /usr/local/lib/python3.10/dist-p
Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dis
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: tokenizers<0.20,>=0.19 in /usr/local/lib/python3.10
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packag
```

```
  Using cached nvidia_cuda_runtime_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (8
Collecting nvidia-cuda-cupti-cu12==12.1.105 (from torch>=1.10.0->accelerate==0.29
  Using cached nvidia_cuda_cupti_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (14
Collecting nvidia-cudnn-cu12==8.9.2.26 (from torch>=1.10.0->accelerate==0.29.3->-r
  Using cached nvidia_cudnn_cu12-8.9.2.26-py3-none-manylinux1_x86_64.whl (731.7 MB
Collecting nvidia-cublas-cu12==12.1.3.1 (from torch>=1.10.0->accelerate==0.29.3->
  Using cached nvidia_cublas_cu12-12.1.3.1-py3-none-manylinux1_x86_64.whl (410.6 M
Collecting nvidia-cufft-cu12==11.0.2.54 (from torch>=1.10.0->accelerate==0.29.3->
  Using cached nvidia_cufft_cu12-11.0.2.54-py3-none-manylinux1_x86_64.whl (121.6 M
Collecting nvidia-curand-cu12==10.3.2.106 (from torch>=1.10.0->accelerate==0.29.3
  Using cached nvidia_curand_cu12-10.3.2.106-py3-none-manylinux1_x86_64.whl (56.5
Collecting nvidia-cusolver-cu12==11.4.5.107 (from torch>=1.10.0->accelerate==0.29
  Using cached nvidia_cusolver_cu12-11.4.5.107-py3-none-manylinux1_x86_64.whl (124
Collecting nvidia-cusparse-cu12==12.1.0.106 (from torch>=1.10.0->accelerate==0.29
  Using cached nvidia_cusparse_cu12-12.1.0.106-py3-none-manylinux1_x86_64.whl (196
Collecting nvidia-nccl-cu12==2.20.5 (from torch>=1.10.0->accelerate==0.29.3->-r re
  Using cached nvidia_nccl_cu12-2.20.5-py3-none-manylinux2014_x86_64.whl (176.2 MB
Collecting nvidia-nvtx-cu12==12.1.105 (from torch>=1.10.0->accelerate==0.29.3->-r
  Using cached nvidia_nvtx_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (99 kB)
Requirement already satisfied: triton==2.3.0 in /usr/local/lib/python3.10/dist-pac
Collecting nvidia-nvjitlink-cu12 (from nvidia-cusolver-cu12==11.4.5.107->torch>=1
  Downloading nvidia_nvjitlink_cu12-12.5.82-py3-none-manylinux2014_x86_64.whl (21
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.3/21.3 MB 70.5 MB/s eta 0:00:00
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dis
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dis
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-p
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.10/dis
Installing collected packages: nvidia-nvtx-cu12, nvidia-nvjitlink-cu12, nvidia-nc
  Attempting uninstall: transformers
    Found existing installation: transformers 4.41.2
    Uninstalling transformers-4.41.2:
      Successfully uninstalled transformers-4.41.2
Successfully installed accelerate-0.29.3 bitsandbytes-0.43.1 nvidia-cublas-cu12-12
```

```
import json
import torch
from transformers import (AutoTokenizer, AutoModelForCausalLM,BitsAndBytesConfig,pipeline
```

```
config_data=json.load(open("/content/config.json"))
HF_TOKEN=config_data["HF_TOKEN"]
```

```
model_name="meta-llama/Meta-Llama-3-8B"
```

```
bnb_config=BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)
```

```
tokenizer = AutoTokenizer.from_pretrained(model_name, token =HF_TOKEN)
tokenizer.pad_token=tokenizer.eos_token
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/file_download.py:1132: Future
  warnings.warn(
tokenizer_config.json: 100%  50.6k/50.6k [00:00<00:00, 2.25MB/s]
tokenizer.json: 100%  9.09M/9.09M [00:00<00:00, 9.65MB/s]
special_tokens_map.json: 100%  73.0/73.0 [00:00<00:00, 3.33kB/s]
Special tokens have been added in the vocabulary, make sure the associated word embed
```

```
model=AutoModelForCausalLM.from_pretrained(
    model_name,
    token=HF_TOKEN,
    quantization_config=bnb_config,
    device_map="auto",
)
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/file_download.py:1132: Future
  warnings.warn(
config.json: 100%  654/654 [00:00<00:00, 23.8kB/s]
model.safetensors.index.json: 100%  23.9k/23.9k [00:00<00:00, 399kB/s]
Downloading shards: 100%  4/4 [01:40<00:00, 21.99s/it]
model-00001-of-00004.safetensors: 100%  4.98G/4.98G [00:28<00:00, 242MB/s]
model-00002-of-00004.safetensors: 100%  5.00G/5.00G [00:28<00:00, 58.2MB/s]
model-00003-of-00004.safetensors: 100%  4.92G/4.92G [00:38<00:00, 43.9MB/s]
model-00004-of-00004.safetensors: 100%  1.17G/1.17G [00:04<00:00, 240MB/s]
Loading checkpoint shards: 100%  4/4 [01:15<00:00, 16.30s/it]
generation_config.json: 100%  177/177 [00:00<00:00, 10.3kB/s]
```

```
text_generator = pipeline(
    'text-generation',
    model=model,
    tokenizer=tokenizer,
    max_new_tokens=128,
)
```

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

```
def get_response(prompt):
  sequence=text_generator(prompt)
  return sequence[0]['generated_text']
  return gen_txt
```

```
prompt="What is Large Language Model "
```

```
llama3_response=get_response(prompt)
```

```
llama3_response
```

```
'What is Large Language Model \xa0(LLM)?\nA large language model (LLM) is a type of artificial intelligence (AI) that is trained to understand human language. LLMs are trained on large amounts of text data, and they use this data to learn the patterns and structures of human language. LLMs can be used to generate text, summarize text, answer questions, and perform other natural language processing tasks.\nWhat are the benefits of LLMs?\nLLMs have many benefits, including the ability to generate text, summarize text, answer questions, and perform other natural language processing task s. LLMs are also able to learn from large amounts of'
```

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

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