# LLAMA 2 – Fine-Tuning Playbook

**Step 1 – Install Requirements**

pip install -r "/content/drive/MyDrive/Colab Notebooks/mlopsproject/requirements.txt"

**Content:**

-f https://download.pytorch.org/whl/torch\_stable.html

torch==2.0.1+cu118

accelerate

appdirs

loralib

bitsandbytes==0.39.1

black

black[jupyter]

datasets

fire

peft

transformers>=4.31.0

sentencepiece

py7zr

scipy

**Step 2 – Copy LLAMA2 model checkpoints from Hugging Face into colab**

!pip install huggingface\_hub

cd /content

snapshot\_download(repo\_id="meta-llama/Llama-2-7b-chat",token="<TOKEN>")

**Step 3 – Run script so that , models can be loaded via HF’s transformers SDK:**

python src/transformers/models/llama/convert\_llama\_weights\_to\_hf.py --input\_dir "/root/" --model\_size 7B --output\_dir "/root/7B-hf"

**Step 4 – Preprocess Data: (Data files in appendix)**

**Step 5 – Fine Tuning using PEFT(lora)**

export CUDA\_VISIBLE\_DEVICES=0

python ../llama\_finetuning.py --use\_peft --peft\_method lora --quantization --dataset alpaca\_dataset --model\_name /root/7B-hf --output\_dir /root/7B-lora

*Note: The script above is passed dataset name alpaca\_dataset whch was already configured to pick alpaca\_data.json file. I preproceesed the data to match dat format and ran the script.*

**Step 6 - Create final Model file by Merging weight of PEFT model to base model**

from peft import PeftModel

from transformers import LlamaForCausalLM, LlamaConfig , AutoModelForCausalLM, TrainingArguments, Trainer

import torch

base\_model = LlamaForCausalLM.from\_pretrained(

"/root/mlops\_llama\_model/7B-hf",

load\_in\_8bit=False,

torch\_dtype=torch.float16,

device\_map="auto"

)

peft\_model = PeftModel.from\_pretrained(base\_model, " /root/mlops\_llama\_model/7B-lora")

peft\_model.merge\_and\_unload().save\_pretrained("/root/mlops\_llama\_model/7B-final")

**Step 7 – Copy tokenizer to above folder.**

cp "/root/mlops\_llama\_model/7B-hf/tokenizer.model" "/root/mlops\_llama\_model/7B-final"

**Step 8 – Inference**

**Run Inference on base model**

python inference/chat\_completion.py --model\_name "/root/8/7B-hf" --prompt\_file prompts.json --quantization

**Run Inference on fine tuned model**

python inference/chat\_completion.py --model\_name "/root/8/7B-final" --prompt\_file prompts.json --quantization

**Appendix:**

Source Data:

A screen shot of a computer

Description automatically generated

Pre-processed data:

A screenshot of a computer screen

Description automatically generated

Prompts.json:

A screenshot of a computer screen

Description automatically generated

**Prompt results for base model:**

**A black screen with white text

Description automatically generated**

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Prompt results for fine-tuned model in uploaded ppt (attached video)**