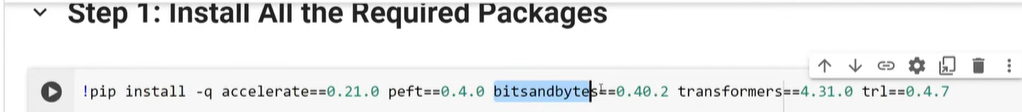
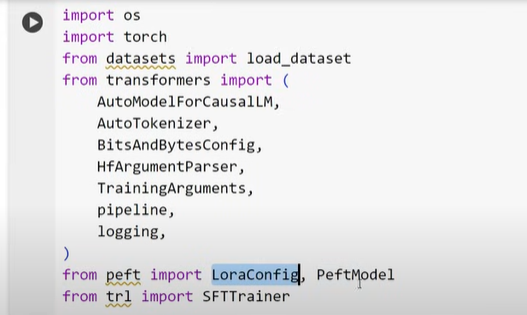
**Parameter Efficient Transfer Technique (PEP) for Finetuning LLM**

It will freeze most of the weights of the LLM Model and only some of the weights will be re-trained.

****

****

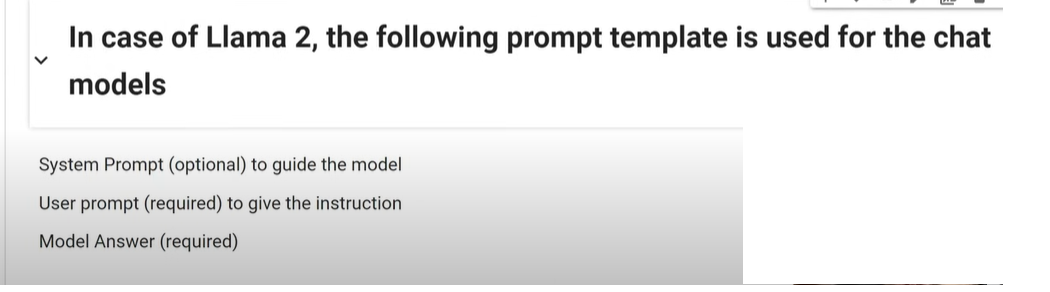
**Note – Bits and Bytes are meant for quantization.**

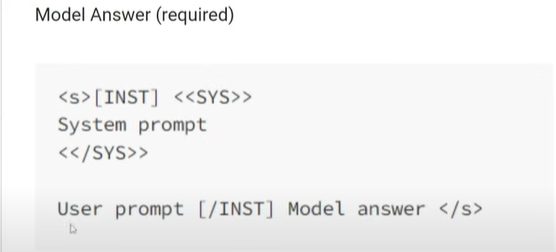
**QUANtization 🡪 convert from higher memory format to lower memory format.**

**Weights in neural network are stored in Matrix and each value is stored in the form of Floating precision 32 bits. We are converting during the quantization process from 32 bits to 8 bits.**

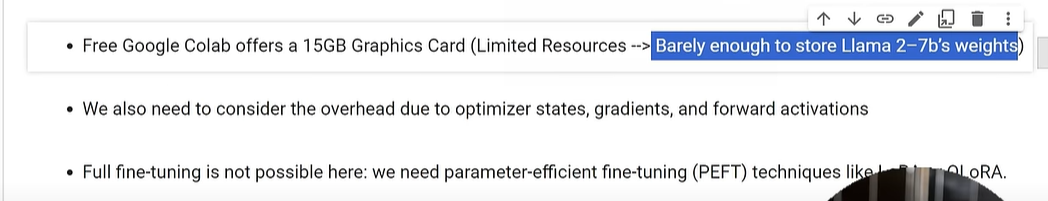
**Inferencing (result from LLM) will be faster.**

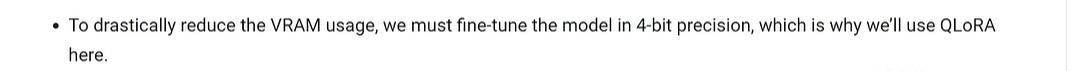
**Each LLM Pre-trained Models are trained on 70 billion parameters. We cannot fine tune it directly as wehave limited RAM in our machine.**

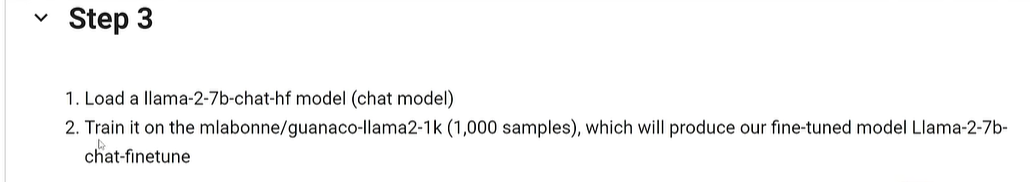
****

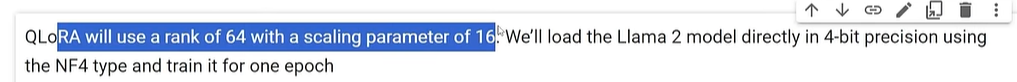
****

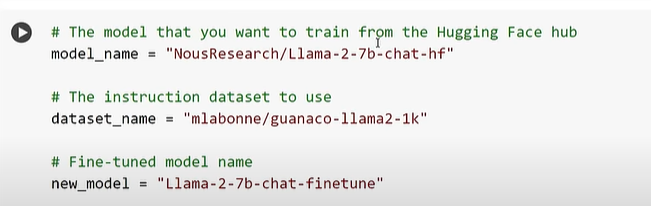
**Steps to Fine Tune LLAMA2**

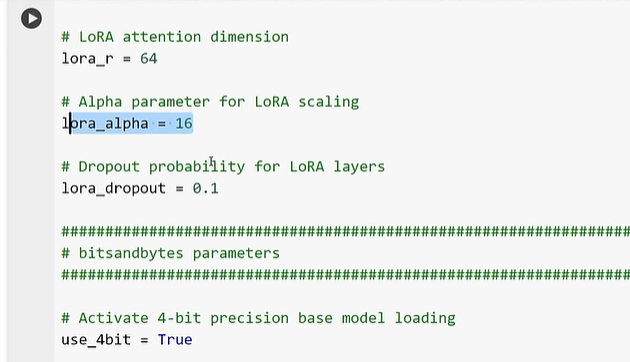
****

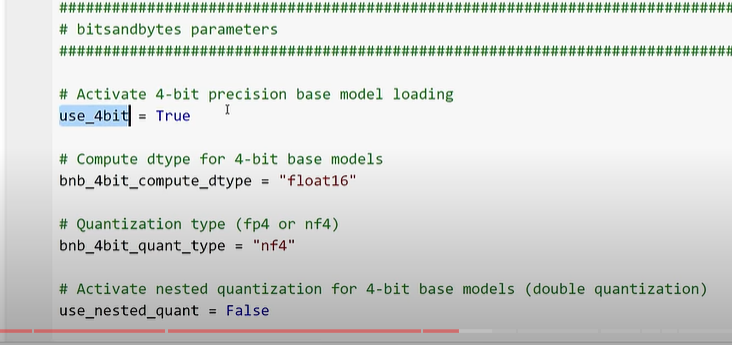
****

****

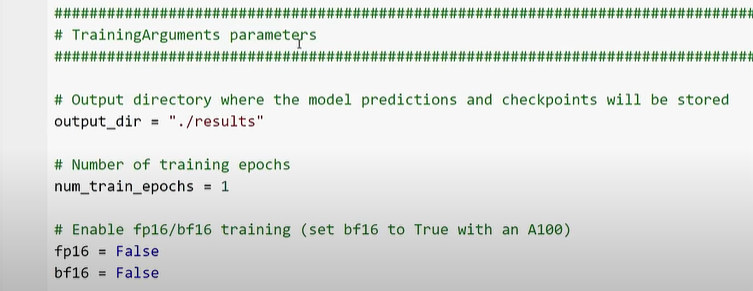
****

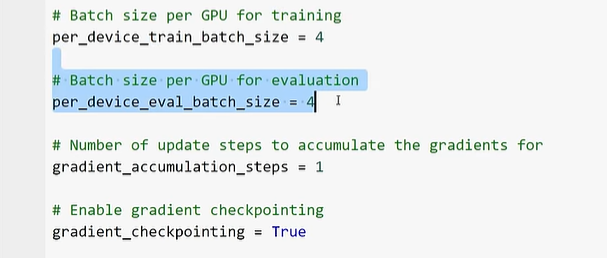
****

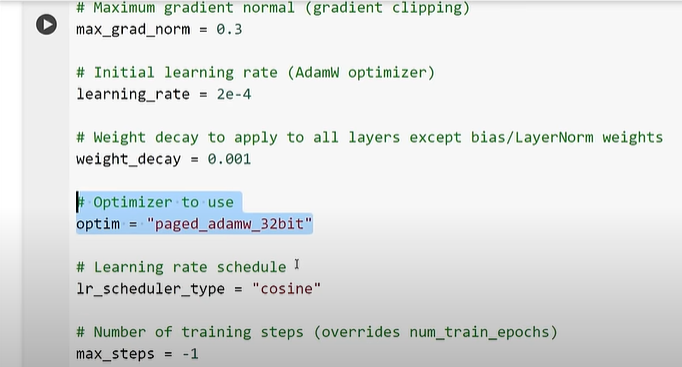
****

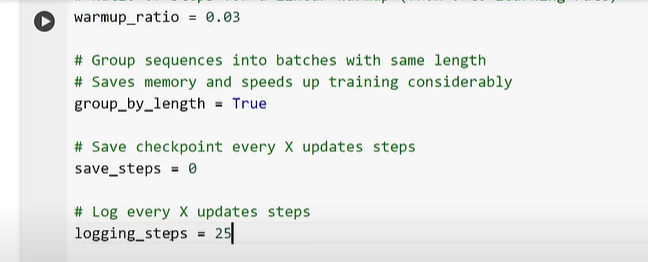
****

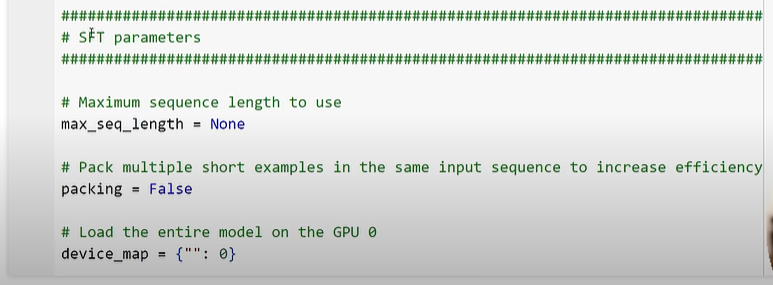
Above are some of the basic parameters we use for fine tuning using LORA Technique











Above is supervised training

