### **Training and Evaluation of Quantized Gemma2 Model**

1. **4 bit Quantized**

#### **Model Size**

The 4-bit quantized Gemma2 model has a size of 6.6 GB.

#### **Training Settings**

The model was trained with the following settings:

* **Batch Size**:
  + Training: 4 per device
  + Evaluation: 4 per device
* **Gradient Accumulation Steps**: 4
* **Optimizer**: Paged AdamW 32-bit
* **Number of Epochs**: 5
* **Evaluation Strategy**: Performed at specific steps
* **Learning Rate**: 2e-4
* **Mixed Precision**:
  + FP16: False
  + BF16: False

#### **Training Output**

* **Global Step**: 195
* **Training Loss**: 0.9153
* **Metrics**:
  + Training Samples Per Second: 2.608
  + Training Steps Per Second: 0.161

#### **Inference Performance**

After fine-tuning, the inference time increased:

* **Before Fine-Tuning**: Approximately 1 second
* **After Fine-Tuning**: Almost 3 seconds

#### **Comments**

Before fine-tuning, the 4-bit quantized Gemma2 model was producing Roman Urdu responses that were generally good, though occasionally mixed with Roman Hindi words. However, the responses were quite robotic and lacked humor. After fine-tuning, the model showed less accuracy compared to both the 8-bit quantized Gemma2 model and the non-fine-tuned 4-bit quantized version.

1. **8 bit Quantized**

#### **Model Size**

The 8-bit quantized Gemma2 model has a size of 9.9 GB.

#### **Training Settings**

The model was trained with the following settings:

* **Batch Size**:
  + Training: 2 per device
  + Evaluation: 2 per device
* **Gradient Accumulation Steps**: 4
* **Optimizer**: Paged AdamW 32-bit
* **Number of Epochs**: 5
* **Evaluation Strategy**: Performed at specific steps
* **Learning Rate**: 2e-4
* **Mixed Precision**:
  + FP16: False
  + BF16: False

#### **Training Output**

* **Global Step**: 237
* **Training Loss**: 1.65
* **Metrics**:
  + Training Samples Per Second: 1.033
  + Training Steps Per Second: 0.128

#### **Inference Performance**

After fine-tuning, the inference time increased:

* **Before Fine-Tuning**: Approximately 5 seconds
* **After Fine-Tuning**: Almost 30 seconds

#### **Comments**

Before fine-tuning, the 8-bit quantized Gemma2 model's Roman Urdu responses were overall good, though they were still robotic, but with a slight twist of humor. After fine-tuning,

**16-bit Quantized**

Was Only able to load 2/4 shards due to memory constraints.

**Issues during fine-tuning:**

* The training loss was constantly zero, applying the following modifications to the tokenizer solved the problem:

model, tokenizer = setup\_chat\_format(model, tokenizer)

* The fine-tuned model wasn’t generating any inference, adding the following code before loading the model and tokenizer resolved the issue:

if torch.cuda.get\_device\_capability()[0] >= 8:

!pip install -qqq flash-attn

torch\_dtype = torch.bfloat16

attn\_implementation = "flash\_attention\_2"

else:

torch\_dtype = torch.float16

attn\_implementation = "eager"

<https://www.datacamp.com/tutorial/fine-tuning-gemma-2>