Accuracy comparison between the pretrained and fine-tuned models on the test set.

Pretrained -> 44%
Fine tuned -> 64%

Time taken to fine-tune the model using QLoRA.
3360.78 seconds for 5 epochs for the following args

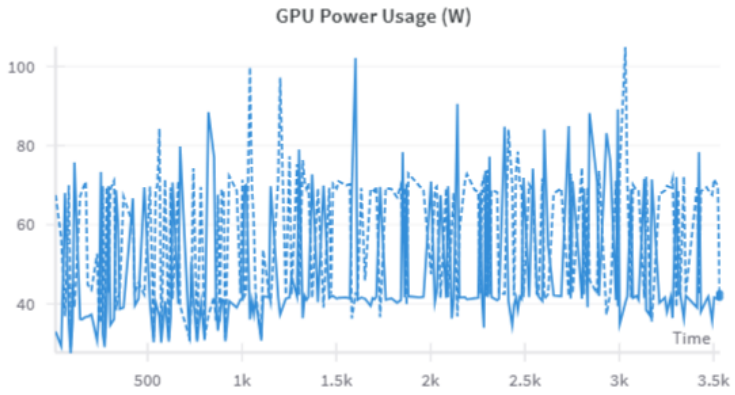
```
# Training arguments
training_args = TrainingArguments(
    output_dir="./models",
    per_device_train_batch_size=2,
    gradient_accumulation_steps=8,
    num_train_epochs=2,
    logging_dir='./logs',
    logging_steps=10,
    save_steps=1,  # Save model after each epoch
    save_total_limit=5,
    fp16=True,
)
```

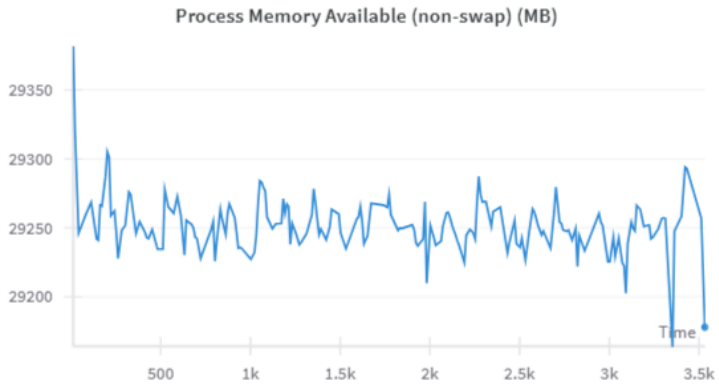
Total parameters in the model and the number of parameters fine-tuned.

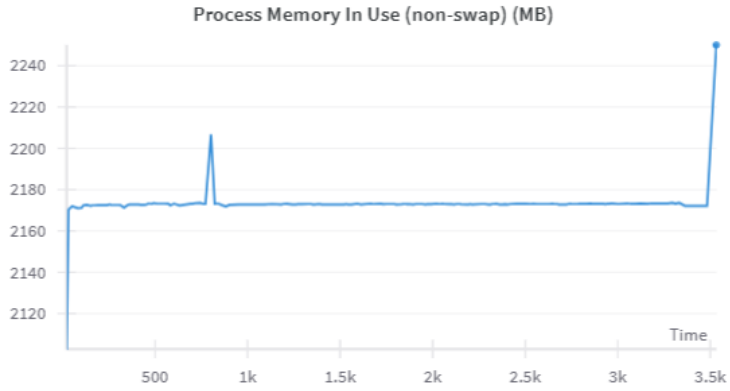
```
25]: model_qlora.print_trainable_parameters()

trainable params: 18,350,080 || all params: 2,798,033,920 || trainable%: 0.6558
```

Resources used (e.g., hardware, memory) during fine-tuning.







Failure cases of the pretrained model that were corrected by the fine-tuned model, as well as those that were not corrected. Provide possible explanations for both.

Examples that pre-trained model got wrong

A person on a horse jumps over a broken down airplane.

A person is training his horse for a competition.

- Possible Issue: The LLM might mistake this for an entailment because "training" and "jumping" could be loosely associated. The model might miss the distinction that while jumping can occur in competition training, it doesn't entail the specific scenario of jumping over an airplane.

A boy is jumping on skateboard in the middle of a red bridge.

The boy skates down the sidewalk.

- Possible Issue: Since both scenarios involve skateboarding, the model might struggle to see this as a contradiction. It may fail to recognize that being on a red bridge and skating down a sidewalk are mutually exclusive in this context.

Two blond women are hugging one another.

The women are on vacation. (Label: Neutral)

- Possible Issue: LLMs may mistakenly infer a connection between relaxation (hugging) and vacation, especially if “vacation” and “hugging” appear frequently together in training data, leading it to wrongly classify this as entailment.

Examples that pre-trained model got right

Premise: "Children smiling and waving at camera"
Hypothesis: "There are children present"
Why it would get it right (Entailm.) This is a straightforward entailment case, where the hypothesis is a direct implication of the premise. LLMs generally perform well with entailment pairs that rely on basic logic and straightforward information retrieval, as it’s clear that if children are smiling at the camera, children are indeed present.

Premise: "A person on a horse jumps over a broken down airplane"
Hypothesis: "A person is at a diner, ordering an omelette"
Why it would get it right (Contrd.) This pair is an obvious contradiction, as there is no plausible overlap between the imagery of someone on a horse and someone ordering food indoors. The LLM is likely to succeed here because it can recognize when two statements are entirely unrelated, leading to a clear contradiction.

Premise: "A boy is jumping on a skateboard in the middle of a red bridge"
Hypothesis: "The boy does a skateboarding trick"
Why it would get it right (Entailm.) In this case, the premise implies that the boy is performing a skateboarding trick, so the hypothesis aligns well. LLMs are effective at detecting entailment when the hypothesis is a reasonable inference based on the action described in the premise.