LLM Assignment 3
Aayush Ranjan
2021003

Accuracy Improvement:

Before Fine-Tuning: The model had an accuracy of 41.0%.

After Fine-Tuning: The accuracy improved significantly to 88.0%, indicating a successful fine-tuning process.

Sample Format Check: No samples were in an incorrect format, meaning the dataset preparation and preprocessing steps were handled well.

Sample Breakdown:
1) 49 samples were initially misclassified by the pretrained model but were correctly classified after fine-tuning.
2) 10 samples remained incorrect even after fine-tuning.
3) 2 samples were correctly classified by the pretrained model but became incorrect after fine-tuning, which is a common occurrence in fine-tuning where some trade-offs may happen.

Training Loss and Validation Loss: The training and validation losses show a downward trend initially, stabilizing after a few epochs, which suggests the model was learning effectively without significant overfitting.

Training Time and Resources: The fine-tuning process took approximately 38 minutes with a peak memory usage of 3.34 GB. This memory footprint seems efficient, possibly due to QLoRA's low-rank adaptation approach, which reduces the number of trainable parameters.

Trainable Parameters: The model had 1.36% of parameters as trainable (around 20.97 million trainable out of a total of 1.54 billion), showing that only a small fraction of parameters were fine-tuned, thanks to QLoRA.
Your results show a substantial increase in accuracy, indicating that the QLoRA fine-tuning process effectively adapted the Microsoft Phi model to the NLI task. If

you're interested in further improving the model, you could experiment with additional epochs or consider tuning hyperparameters like learning rate and batch size.

Screenshots of the Results:

```
Accuracy before fine-tuning:  41.0
Samples in incorrect format:  0
CPU times: user 6min 1s, sys: 286 ms, total: 6min 1s
Wall time: 6min 1s
```

```
Trainable Parameters: 20971520
Total Parameters: 1542364160
Percentage of Trainable Parameters: 1.36%
```

[625/625 38:12, Epoch 4/5]

Epoch	Training Loss	Validation Loss
0	0.576300	0.471104
2	0.418700	0.458096
4	0.372700	0.472280

```
Peak Memory Usage (GB): 3.34
Time taken for fine-tuning: 38.29 minutes
```

```
Accuracy after fine-tuning:  88.0
Samples in incorrect format:  0
```

```
The number of samples:

Pretrained: Incorrect, Finetuned: Correct -> 49
Pretrained: Incorrect, Finetuned: Incorrect -> 10
Pretrained: Correct, Finetuned: Incorrect -> 2
```

Example Model Corrections/Failures

Pretrained Model: Incorrect, Finetuned Model: Correct (5 out of 49)

```
{
     "premise": "A man in a black shirt, in a commercial kitchen, holding up meat he
took out of a bag.",
     "hypothesis": "A man in a black shirt, in a commercial kitchen, holding up the
old meat he took out of a bag.",
     "label": 1,
     "pretrained_label": 0,
     "finetuned_label": 1
  },
  {
     "premise": "a woman in a black shirt looking at a bicycle.",
     "hypothesis": "A woman dressed in black shops for a bicycle.",
     "label": 1,
     "pretrained_label": 0,
     "finetuned_label": 1
  },
  {
     "premise": "A group of people stand near and on a large black square on the
ground with some yellow writing on it.",
     "hypothesis": "a group of people wait",
     "label": 1,
     "pretrained_label": 0,
     "finetuned_label": 1
  },
  {
     "premise": "Two men in neon yellow shirts busily sawing a log in half.",
     "hypothesis": "Two men are cutting wood to build a table.",
     "label": 1,
     "pretrained_label": 0,
```

```
        "finetuned_label": 1
    },
    {
        "premise": "A man is renovating a room.",
        "hypothesis": "A man is using a hammer in a room.",
        "label": 1,
        "pretrained_label": 0,
        "finetuned_label": 1
    },
```

Pretrained Model: Incorrect, Finetuned Model: Incorrect (5 out of 10)

```
{
        "premise": "This church choir sings to the masses as they sing joyous songs
from the book at a church.",
        "hypothesis": "The church has cracks in the ceiling.",
        "label": 1,
        "pretrained_label": 0,
        "finetuned_label": 2
    },
    {
        "premise": "Two men climbing on a wooden scaffold.",
        "hypothesis": "Two sad men climbing on a wooden scaffold.",
        "label": 1,
        "pretrained_label": 0,
        "finetuned_label": 2
    },
    {
        "premise": "An older gentleman wearing a hat is walking on crutches next to a
busy street.",
        "hypothesis": "A man with a walking stick is next to the street.",
        "label": 2,
        "pretrained_label": 0,
        "finetuned_label": 1
    },
    {
        "premise": "A small child is riding in a red wagon.",
        "hypothesis": "A kid is in a wagon and someone is pulling it.",
        "label": 1,
        "pretrained_label": 0,
        "finetuned_label": 0
    },
```

```
    {
        "premise": "Two men playing in a beautiful lake surrounded by mountains.",
        "hypothesis": "A couple people enjoy the water near the mountains.",
        "label": 1,
        "pretrained_label": 0,
        "finetuned_label": 0
    },
```

Pretrained: Correct, Finetuned: Incorrect (2)

```
{
        "premise": "A Skier ski-jumping while two other skiers watch his act.",
        "hypothesis": "A skier preparing a trick",
        "label": 0,
        "pretrained_label": 0,
        "finetuned_label": 1
    },
    {
        "premise": "A woman is standing near three stores, two have beautiful artwork
and the other store has Largo written on it.",
        "hypothesis": "A woman standing on a street corner outside beside three
different stores, two of which contain beautiful artwork and one with a Largo sign.",
        "label": 0,
        "pretrained_label": 0,
        "finetuned_label": 1
    }
```

Explanation:

The model corrected some samples because fine-tuning exposed it to relevant entailment patterns, improved its contextual understanding, and reduced biases from pretraining, aligning it more closely with task-specific data.

The model wasn't able to correct some samples as:

Complexity of Language: Subtle nuances and indirect entailments can be challenging for the model.

Limited Fine-Tuning Data: Insufficient diversity in data limits generalization.

Model Limitations: Base architecture and limited trainable parameters can restrict performance.

Pretrained Model Bias: Residual biases from pretraining can affect fine-tuning outcomes.

References:
1) https://www.kaggle.com/code/hari31416/downloading-file-and-directory-from-kaggle
2) https://dassum.medium.com/fine-tune-large-language-model-llm-on-a-custom-dataset-with-qlora-fb60abdeba07