REPORT
Assignment - 3

Prerak Gupta
2021552

- **Accuracy**
  Pretrained - 25%
  Fine-tuned - 45%

- **Time taken** to fine-tune using QLoRA - 2522.0316 or 42 minutes

- **Parameters**
  Total parameters in the models - 1563335680
  Number of parameters fine-tuned - 41943040
  Percentage of trainable parameters - 2.68%

- **Resources** used during fine-tuning
  System RAM - 3.87 / 12.7 GB
  GPU RAM - 3.6 / 15.0 GB
  Disk - 37.5 / 112.6 GB

- **Training statistics**

Epoch	Training Loss	Validation Loss
1	2.361100	2.716889
2	2.010900	2.500636
3	1.968900	2.337155
4	1.770700	2.274220
5	1.751500	2.190732

- **Failure cases of the pretrained model that were corrected by the fine-tuned model**

  Example 1 :-
  Premise - Two men climbing on a wooden scaffold.
  Hypothesis - Two sad men climbing on a wooden scaffold.

  Correct Label - Neutral
  Pretrained Model Label - Entailment
  Finetuned Model Label - Neutral

  Example 2 :-
  Premise - A woman in a black shirt looking at a bicycle.
  Hypothesis - A woman dressed in black shops for a bicycle.
  Correct Label - Neutral
  Pretrained Model Label - Entailment
  Finetuned Model Label - Neutral

- **Failure cases of the pretrained model that were not corrected by the fine-tuned model**

  Example 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.

  Correct Label - Neutral
  Pretrained Model Label - Entailment
  Finetuned Model Label - Contradiction

  Example 2 :-
  Premise - Two men in neon yellow shirts busily sawing a log in half.
  Hypothesis - Two men are cutting wood to build a table.

  Correct Label - Neutral

Pretrained Model Label - Entailment
Finetuned Model Label - Contradiction

- **Explanation**
  The fine-tuned model became better at recognizing when slight differences in description (such as mood or context) do not necessarily change the relationship between premise and hypothesis. Fine-tuning helped the model distinguish nuanced meanings and avoid automatically concluding that overlapping words imply an exact match.
  Some failures arose because the model struggled with implicit meanings, like waiting versus simply standing, which are not directly stated in text. Distinguishing such nuances often requires more context than is available in a single sentence.
  Fine-tuning did not fully resolve these cases because understanding implied actions often requires deeper reasoning or additional examples that explicitly deal with implied behaviors.

Saved model and checkpoints
https://drive.google.com/drive/folders/1p1Nqzsn-D2vHSlUP7BufrmA3faWYyZwD?usp=drive_link

Resources

https://medium.com/@prasadmahamulkar/fine-tuning-phi-2-a-step-by-step-guide-e672e7f1d009

https://dassum.medium.com/fine-tune-large-language-model-llm-on-a-custom-dataset-with-qlora-fb60abdeba07

https://github.com/brevdev/notebooks/blob/main/mistral-finetune-own-data.ipynb

https://medium.com/@levxn/lora-and-qlora-effective-methods-to-fine-tune-your-llms-in-detail-6e56a2a13f3c

https://wandb.ai/byyoung3/ml-news/reports/How-to-Fine-Tune-LLaVA-on-a-Custom-Dataset--Vmlldzo2NjUwNTc1