Generative AI Assignment 3

Name : Zain Al Abidin
Roll No : 21L-6260

General Discussion Based on the results :

Full Fine-Tuning:
This method updates all parameters of the model, resulting in the highest number of trainable parameters and relatively high training time and memory usage. While it performs decently in terms of accuracy, it is resource-intensive and not ideal for limited compute environments.

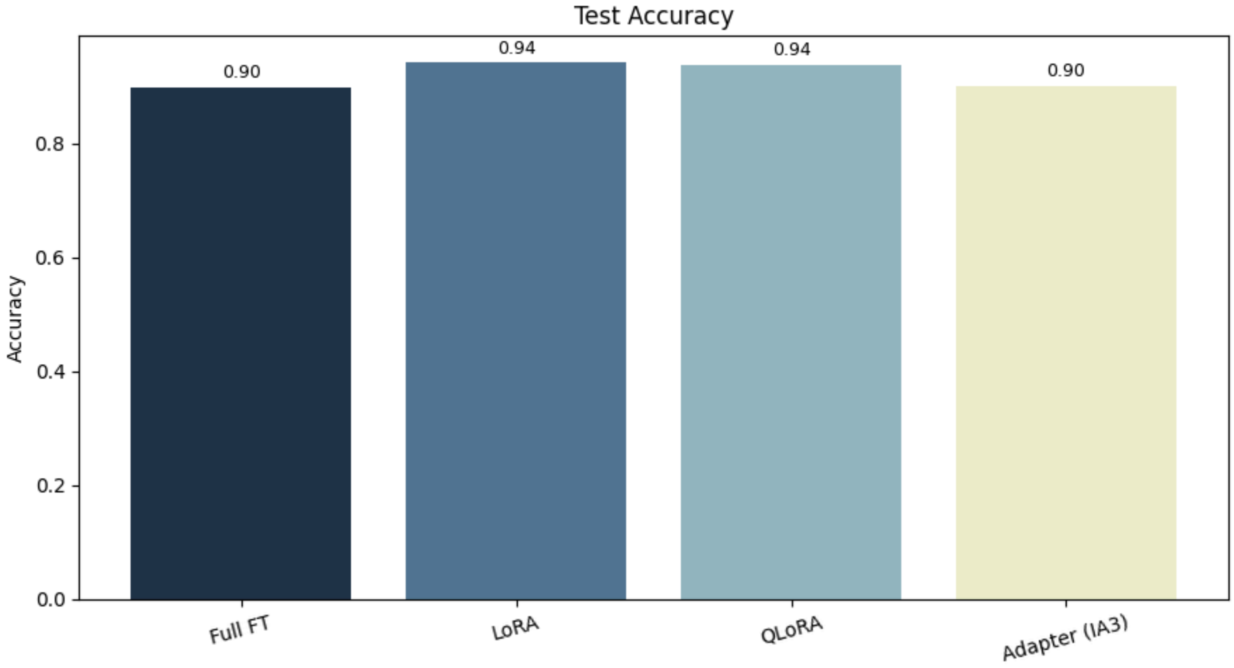
LoRA Fine-Tuning (PEFT):
LoRA achieves the best test accuracy among all methods while using a fraction of the parameters compared to full fine-tuning. It is also faster to train and uses less GPU memory, making it highly efficient and practical for real-world applications.

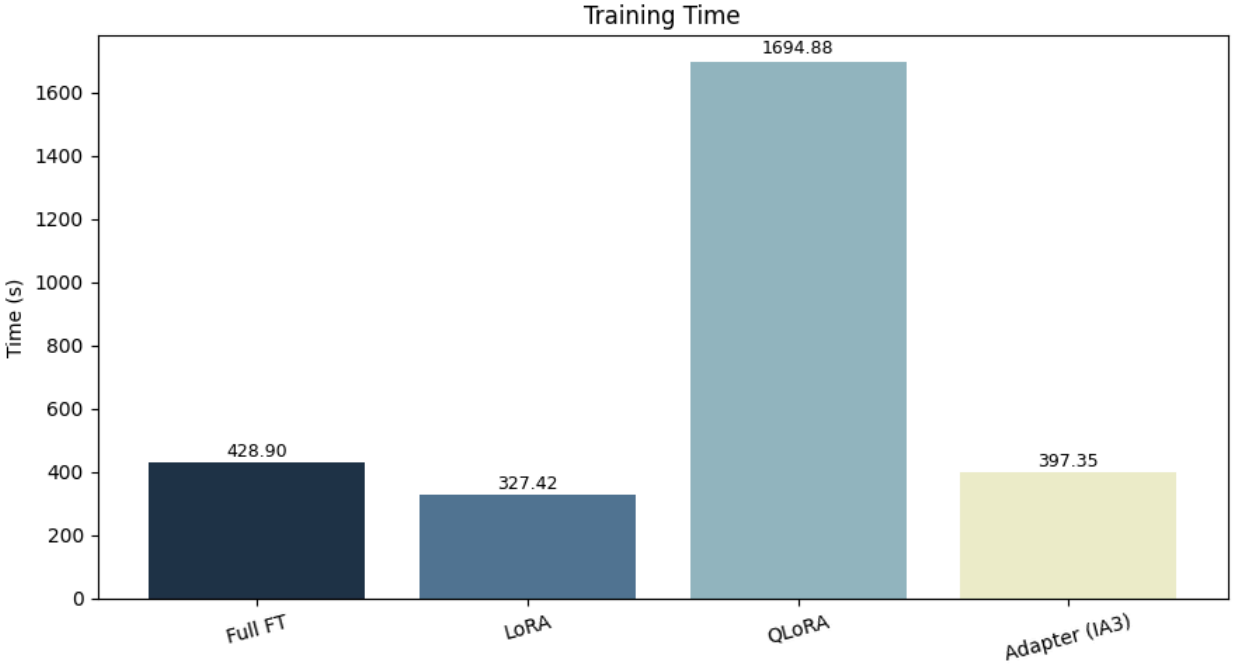
QLoRA Fine-Tuning:
QLoRA maintains competitive accuracy and uses very few trainable parameters, similar to LoRA. However, it requires significantly more GPU memory and has a much longer training time, which may make it less feasible for environments with limited hardware.

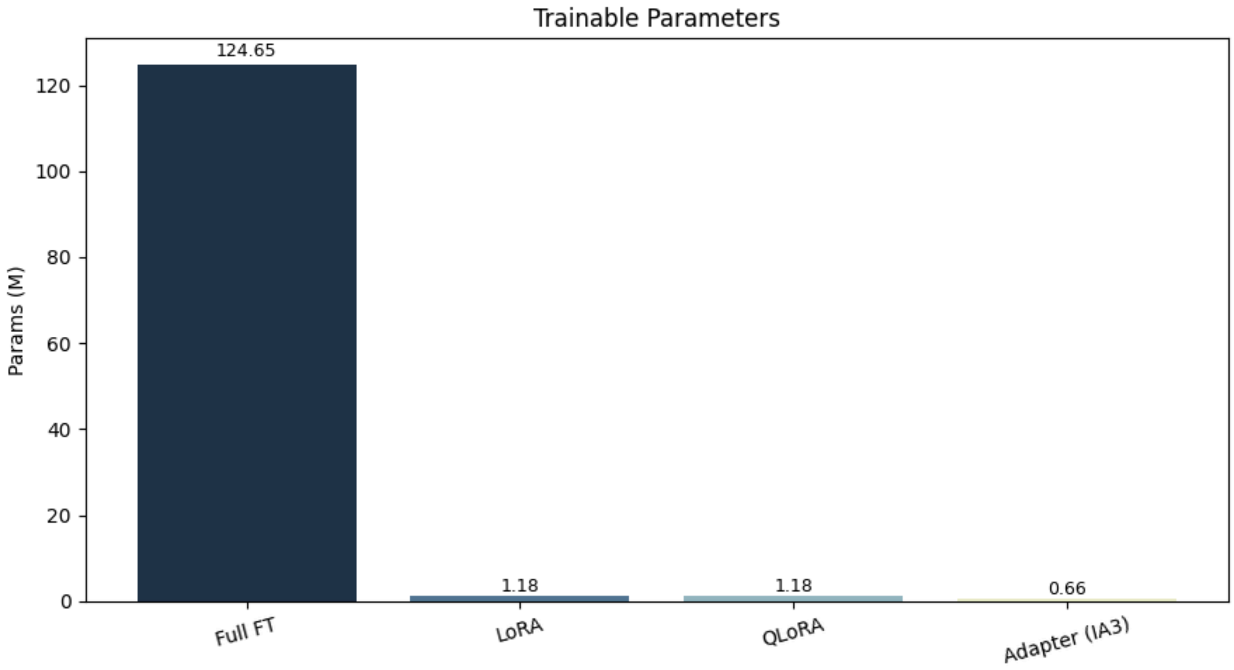
Adapter Tuning (IA3):
IA3 is the most lightweight approach in terms of parameters and uses moderate resources. It provides reasonable accuracy but does not outperform LoRA or QLoRA. It is a suitable option when minimizing model size is a priority, even at the cost of slight performance trade-offs.

Note : I only ran 5 epochs for each method as mentioned in the assignment

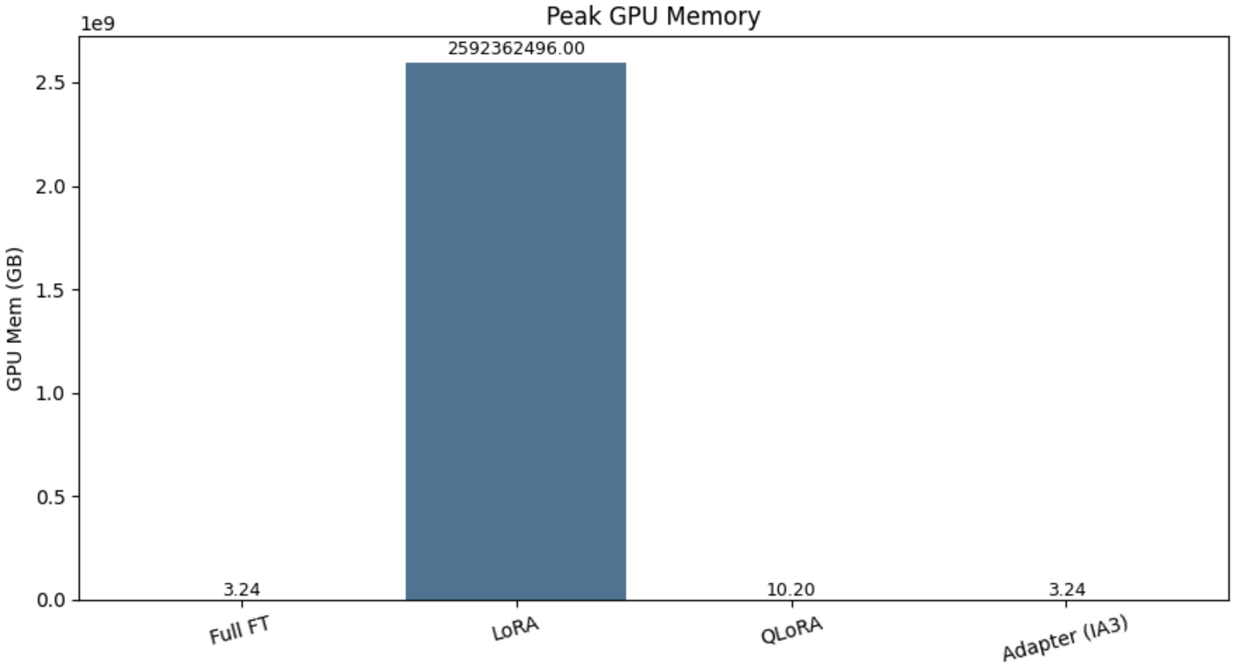
Comparing Metrics of Each Method





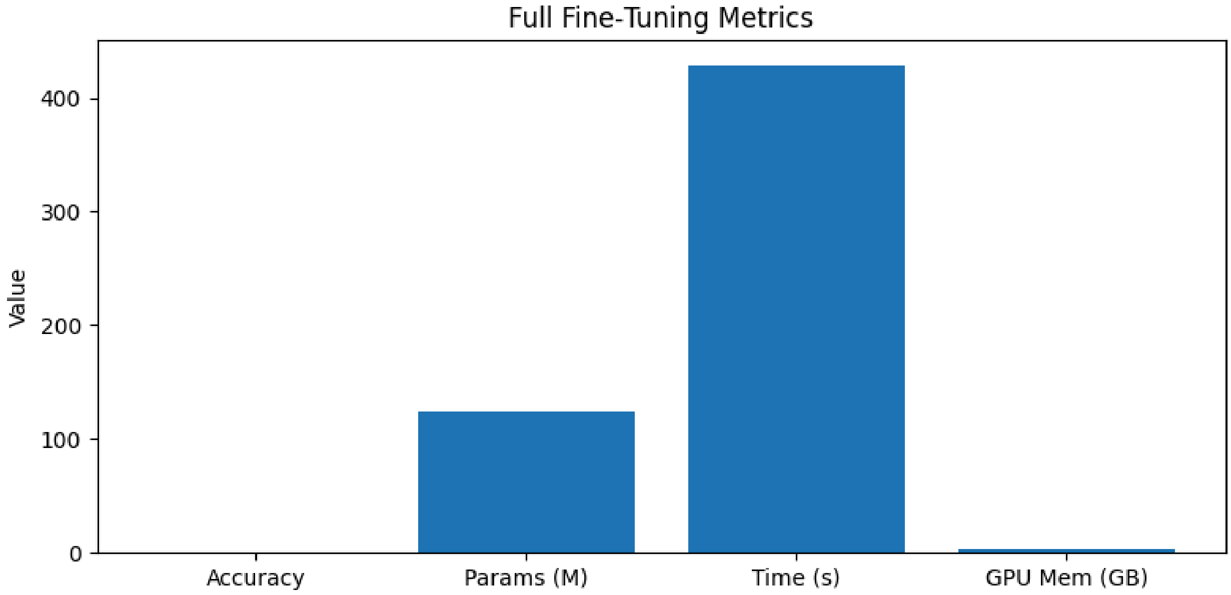


Trainable Parameters



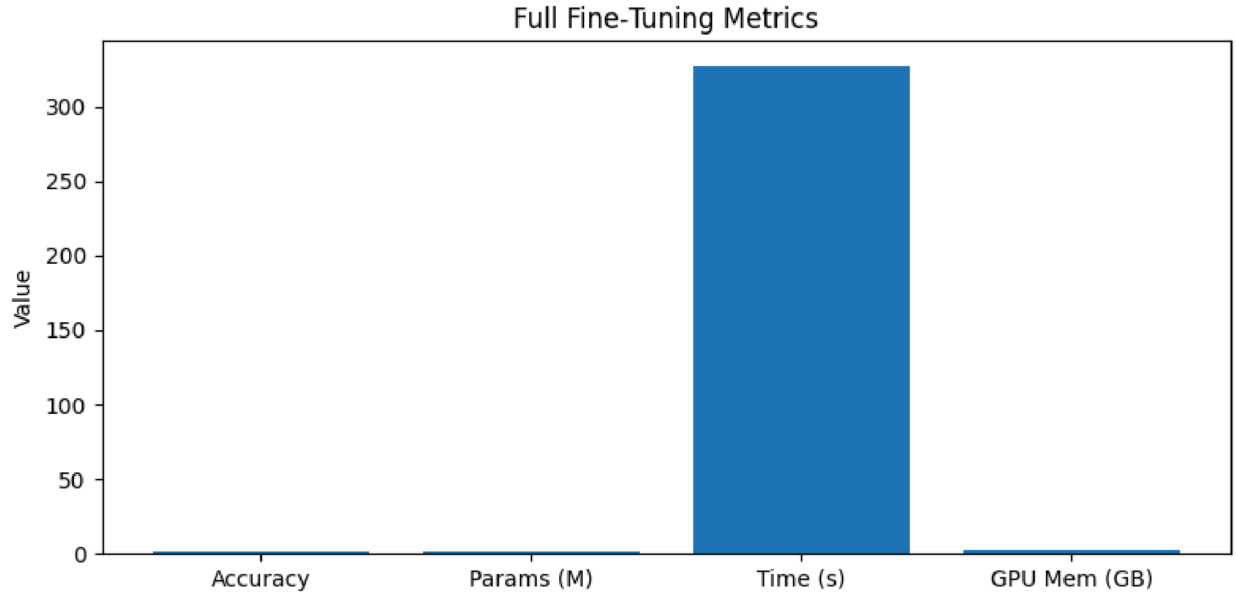
Peak GPU Memory

Method 1: Full Fine-Tuning



Test Accuracy	0.8980
Trainable Parameters	124,647,170 (124.65M)
Training Time (s)	428.9
Peak GPU Memory (GB)	3.24

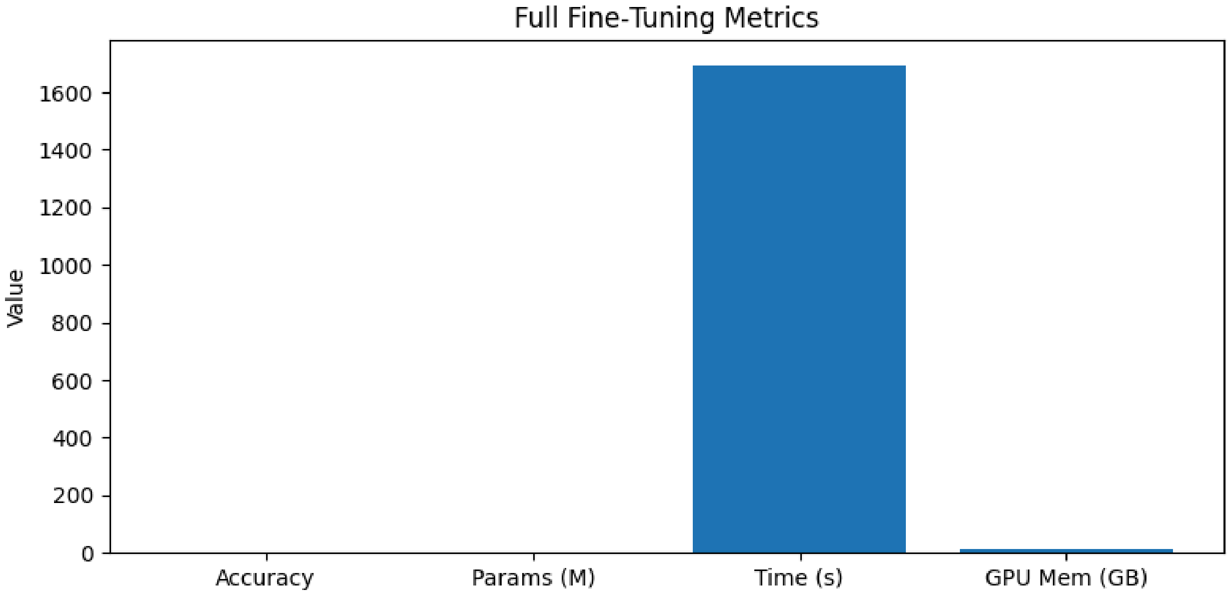
Method 2: LoRA Fine-Tuning using PEFT



Results :

Test Accuracy	0.9415
Trainable Parameters	1,181,954 (1.18M)
Training Time (s)	327.4
Peak GPU Memory (GB)	2.41

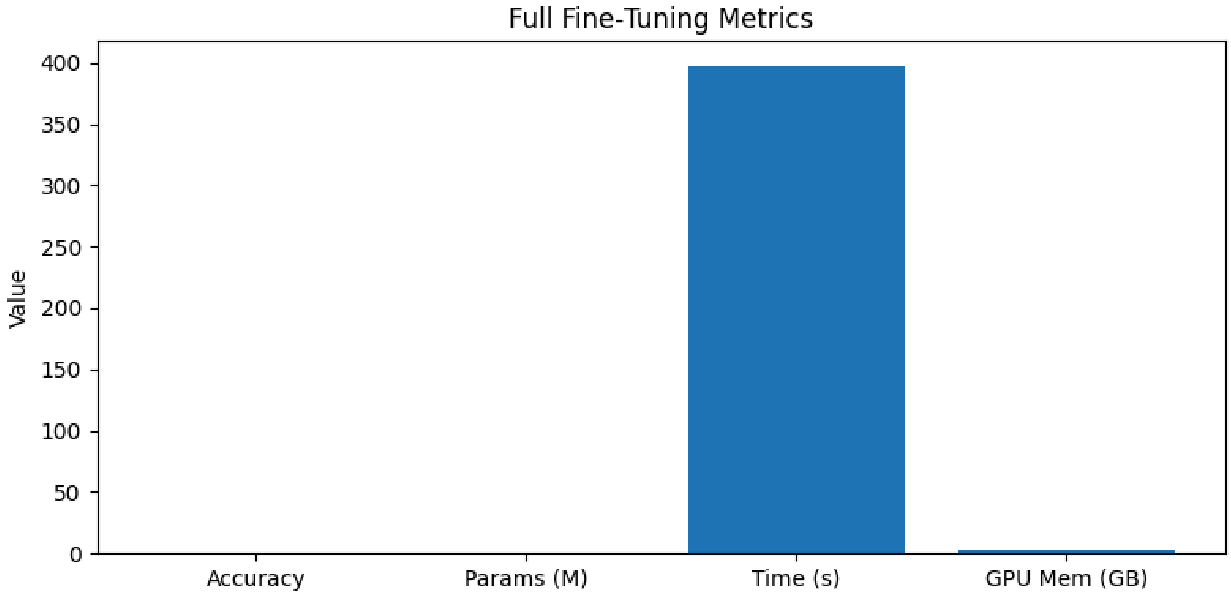
Method 3: QLoRA Fine-Tuning



Results :

Test Accuracy	0.9385
Trainable Parameters	1,181,954 (1.18M)
Training Time (s)	1694.9
Peak GPU Memory (GB)	10.20

Method 4: Adapter Tuning (IA3)



Results :

Test Accuracy	0.9010
Trainable Parameters	656,642 (0.66M)
Training Time (s)	397.3
Peak GPU Memory (GB)	3.24