Summary of the Assignment:

Classification:

- Finetuned Llama3.2 model of 1.2 Billion parameters for classification task on SST-2 dataset.

- Achieved accuracy of 13% for zero shot classification and 95% for finetuned model after training for 2 epochs.

- Accuracy score for finetuned model is more than zero shot classification on Llama model, because:
    - in zero shot classification, it relies on its pre-trained knowledge and general understanding of tasks.
    - during fine-tuning, the model learns directly from task-relevant labeled data, aligning its representations more closely to the specific requirements of the classification task.

- In pretraining the **entire model** is trained, which includes all its parameters (1.2 billion in our case).

- Unlike pretraining, we have used LoRA (Low-Rank Adaptation) adaptor to train on few parameters (5.6 million) instead of all (1.2 billion) parameters since the approach significantly reduces memory requirements and computational overhead, making it feasible to fine-tune LLMs on smaller hardware setups while retaining comparable performance.

Question-Answering task:

- In zero-shot evaluation, the scores (EM: 26.27%, F1: 26.44%, BLEU: 0.004) are significantly higher than those in fine-tuned evaluation (EM: 4.97%, F1: 5.36%, BLEU: 0.018). Fine-tuned evaluation metrics show a decrease in EM and F1, while metrics like BLEU, ROUGE, and METEOR slightly improve.
- The same number of parameters implies the model's capacity to learn remained constant. The difference in performance arises from how the weights were updated during fine-tuning. Fine-tuning adapts the pre-trained model to task-specific objectives without increasing computational complexity.
- Lower Exact Match (EM) and F1 Scores in fine-tuned evaluation suggest that fine-tuning may have caused the model to overfit or diverge from generalizable representations.

- The fine-tuning data may have been small, imbalanced, or noisy, leading to reduced performance.
- Zero-shot evaluation typically leverages the model's generalized pretraining capabilities, which may explain the initially higher scores.
- Higher BLEU and ROUGE Scores in fine-tuning may indicate a marginal improvement in token-level or sentence-level similarity for the fine-tuned data. However, the difference is negligible.