Assignment 3: Finetuning LLMs

Student Name: Farah AbuAjameieh **Student ID:** 161430

1. Document your fine-tuning process — include model selection, dataset preparation, training configuration, and compute used:

Model Selection: Llama3

Dataset: 5000 samples of "deepmind/math_dataset", "arithmetic__add_or_sub"

Training Confiuration:

```
training_arguments = SFTConfig(
    output_dir="outputs",
    per_device_train_batch_size=4,
    per_device_eval_batch_size=1,
    gradient_accumulation_steps=2,
    optim="paged_adamw_32bit",
    num_train_epochs=1,
    # eval_strategy ="epoch",
    warmup_steps=10,
    learning_rate=2e-4,
    fp16=True,
    bf16=False,
    group_by_length=True,
    report_to="none",
    max_length=512,
)
```

2. Evaluate the model before and after fine-tuning using a set of testing prompts. Compare its performance with respect to f1, Rouge-L, and Bert-score.

Before fine-tuning:

Average F1 Score:	0.0744
ROUGE-L Score:	0.1221
BERTScore (F1):	0.7855

After fine-tuning:

Average F1 Score:	0.1271	→ +0.0527 increase (≈ 70.83%)
ROUGE-L Score:	0.1932	→ +0.0711 increase (≈ 58.23%)
BERTScore (F1):	0.8695	→ +0.0840 increase (≈ 10.70%)