Gen AI
Assignment 4 Report

Group Members

- **21L-5659 - Ammar Ali Khan**
- **21L-6292 - Hamza Ahmed**
- **21L-6234 - Murtaza Ahmed**
- **21L-6260 - Zain Al Abidin**

Dataset

- **Source**: `ccdv/arxiv-summarization`
- **Subset**: 5,000 samples
- **Splits**:
    - **Train**: 4,000
    - **Validation**: 500
    - **Test**: 500
- **Input Format**: Prompted article text
- **Label**: Abstract summary(arXiv)

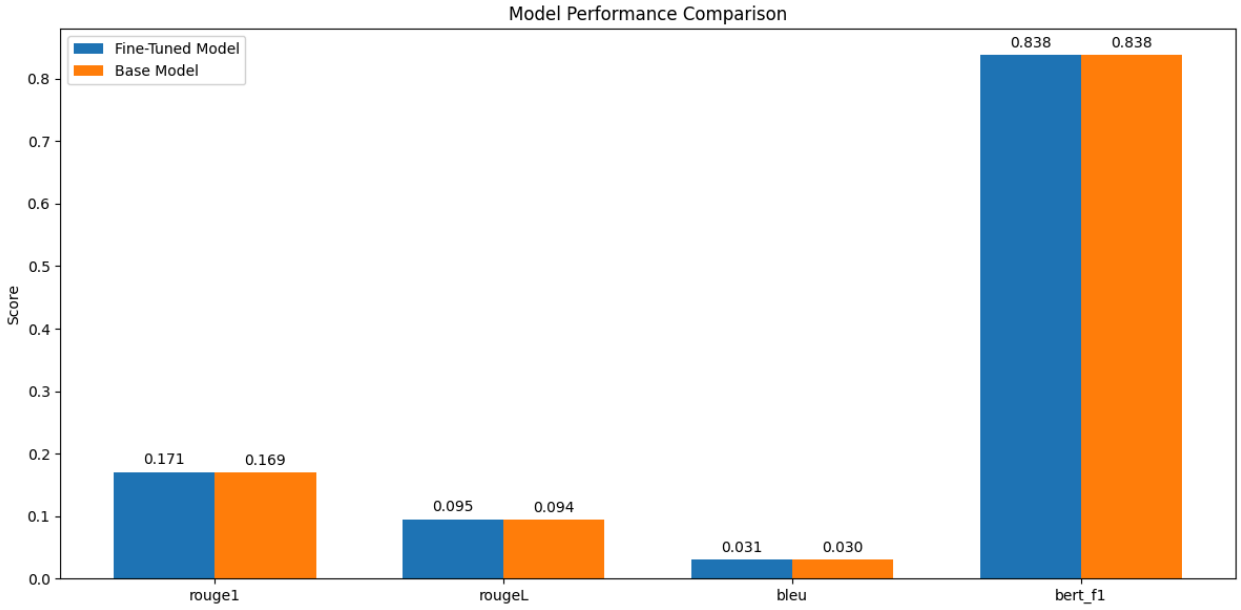
Model & Training

- **Base Model**: LLaMA-based (4-bit, FP16)
- **Trainer**: `SFTTrainer` from `trl`
- **Hyperparameters**:
    - **Batch Size**: 1 (gradient accumulation: 16)
    - **Epochs**: 5
    - **Warmup Steps**: 50
    - **Optimizer**: AdamW (8-bit)
    - **FP16**: Enabled

```
 Epoch  Training Loss  Validation Loss
Done! Runtime: 17.1017 seconds
```

Evaluation Metrics

- **ROUGE-1**:
  - **Fine-Tuned**: 0.1705
  - **Base**: 0.1694
- **ROUGE-L**:
  - **Fine-Tuned**: 0.0949
  - **Base**: 0.0942
- **BLEU**:
  - **Fine-Tuned**: 0.0309
  - **Base**: 0.0304
- **BERTScore F1**:
  - **Fine-Tuned**: 0.8380
  - **Base**: 0.8380



Outputs

- **Comparison File**: `comparisons.json` (10 test samples)

LLM-as-a-Judge prompts

```python
def evaluate_summary_with_llm(article, reference, generated_summary, model="llama-3.1-70b-instant"):
"""
Evaluate the quality of a generated summary using Groq's LLM.

Parameters:
- article: The original research article or excerpt.
- reference: The reference (ground truth) summary.
- generated_summary: The summary to evaluate.
- model: The Groq model used for evaluation.

Returns:
- A structured evaluation response with scores and justifications.
"""
prompt = f"""You are an expert reviewer evaluating the quality of an AI-generated summary of a research paper. Assess the summary based on the following criteria:

1. **Fluency (1-5)**: Is the summary well-written, grammatically correct, and easy to understand?
2. **Factuality (1-5)**: Are the statements accurate and faithful to the original text?
3. **Coverage (1-5)**: Does the summary capture the core problem, methodology, and main findings of the paper?

Score each aspect from 1 (poor) to 5 (excellent), and provide a brief explanation for each score.

---

**Original Paper (excerpt):**
{article[:2000]}... [truncated for brevity]

**Generated Summary:**
{generated_summary}

---

**Your Evaluation:**
Provide scores and justifications for each of the three criteria.
"""
```

```python
response = client.chat.completions.create(
    model=model,
    messages=[
        {"role": "system", "content": "You are an expert evaluator who rates summaries based on how fluent, factually accurate, and complete they are."},
        {"role": "user", "content": prompt}
    ],
```