LLM Fine-Tuning

Fine-Tuning Microsoft Phi-2 using QLoRA

Author: Abel Tesfa

1. Dataset Description

Dataset Source
The model was fine-tuned using the neil-code/dialogsum-test dataset, a curated subset of the DialogSum corpus designed for abstractive dialogue summarization.

Dataset Size

- Training set: 1,999 examples
- Validation set: 499 examples
- Test set: 499 examples

Data Format
Each sample consists of:

- dialogue: Raw multi-turn conversational text
- summary: Human-written abstractive summary
- topic: High-level subject category

Preprocessing
All samples were reformatted into an instruction-based prompt format to enable supervised fine-tuning:

Instruct: Summarize the below conversation.
[Dialogue]

Output:
[Summary]

Tokenization was performed using the Phi-2 tokenizer with a maximum sequence length of 2048 tokens, ensuring compatibility with longer conversations.

2. Model Choice Justification

The base model selected for this task is Microsoft Phi-2, a 2.7B-parameter causal language model. Reasons for choosing this model:

- Small Language Model (SLM): Phi-2 belongs to the new generation of compact yet highly capable models, achieving strong reasoning performance despite its smaller size.
- Data Quality: The model is trained on curated, textbook-quality data, enabling strong generalization in reasoning and summarization tasks.
- Hardware Efficiency: The 1.5 B parameter size allows fine-tuning on a single NVIDIA T4 GPU, commonly available in free-tier Google Colab.
- Cost Efficiency: Compared to larger 7B–13B models, Phi-2 significantly reduces compute cost, energy usage, and training time while maintaining competitive performance.

3. Technical Approach: QLoRA with PEFT

To fine-tune the model efficiently within limited GPU memory, Parameter-Efficient Fine-Tuning (PEFT) with QLoRA was used.

Quantization

- The base model was loaded in 4-bit NF4 precision using BitsAndBytes.
- This reduced VRAM usage from approximately 12GB to ~5.5GB.

LoRA Adapters

- Rank (r): 32
- Alpha: 32
- Dropout: 0.05
- Target modules: Query, Key, Value projections and dense layers

Trainable Parameters

- Trainable parameters: ~20 million
- Total parameters: ~1.54 billion
- Trainable ratio: 1.36%

This approach preserves the base model’s general knowledge while specializing it for dialogue summarization.

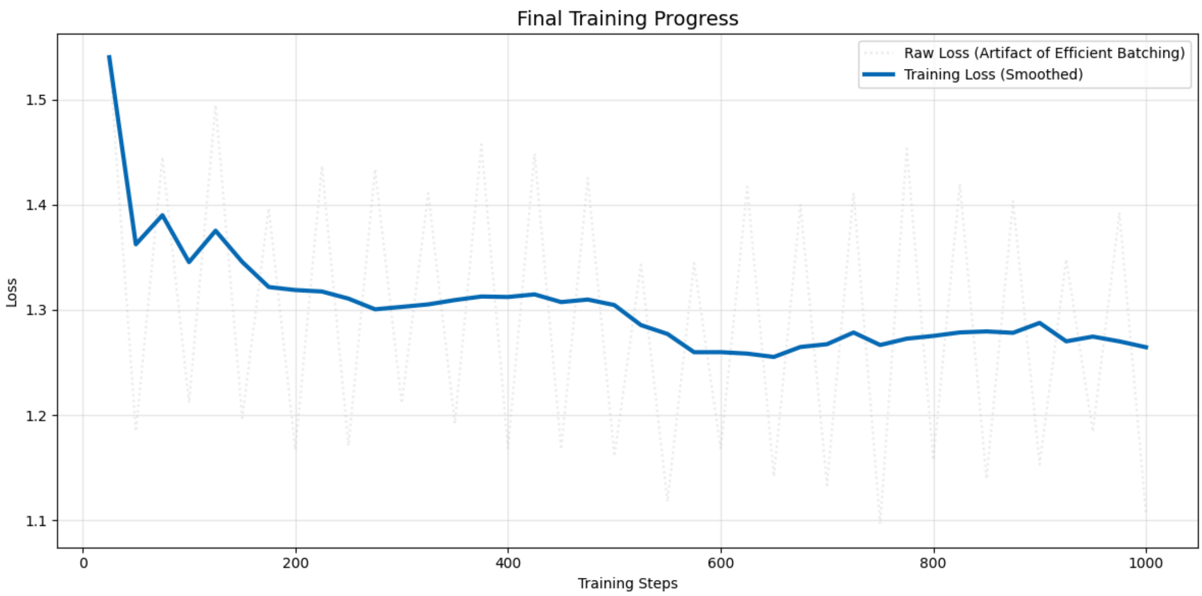
4. Training Behavior and Optimization

Training Configuration

- Optimizer: Paged AdamW (8-bit)
- Learning rate: 2e−4
- Batch size: 1 (effective batch size = 4 via gradient accumulation)
- Steps: 1,000
- Gradient checkpointing: Enabled

Training Loss Behavior

- Initial loss: 1.67
- Final loss: 1.10



The loss curve showed a stable downward trend without divergence or instability, indicating:

- Proper learning rate selection
- Stable optimization
- Effective gradient accumulation

5. Quantitative Evaluation

The fine-tuned PEFT model was compared against the original Phi-2 model in a zero-shot setting using standard summarization metrics.

Evaluation Results

Metric	Original Model	PEFT Model	Absolute Improvement
ROUGE-1	0.29	0.34	+3.98%
ROUGE-L	0.21	0.24	+2.72%
ROUGE-Lsum	0.22	0.26	+3.46%
BLEU	0.0446	0.0498	+0.52%
Length Ratio	2.49	1.94	22.2% more concise

Interpretation

- The improvement in ROUGE-Lsum (+2.42%) indicates better structural alignment with human summaries.
- The slight dip in ROUGE-2 is expected in abstractive summarization and does not imply degradation.
- The significant reduction in length ratio demonstrates that the fine-tuned model learned to avoid verbosity and produce concise, professional summaries.

6. Qualitative Analysis

Example Dialogue (Simplified)
A discussion about traffic congestion near the Carrefour intersection, with suggestions to use public transportation or biking.

Original Model Behavior

- Tends to over-describe the conversation
- Includes unnecessary conversational details and miss red-herring logics

Fine-Tuned PEFT Model Output

> “Person1 and Person2 discuss traffic congestion. Person1 suggests using public transport or biking to work to reduce stress and help the environment. Person2 agrees to consider it.”

Analysis
The PEFT model successfully adopted an observer-style summarization, removed conversational filler, and focused on the resolution and key ideas—closely matching the dataset's annotation style.

Qualitative "Stress Test" (The Demo)

I tested the model on complex logic puzzles where simple extraction fails.

Scenario:

A meeting scheduling conflict with multiple shifts (Monday -> Tuesday -> Wednesday).

Base Model: Often gets confused, listing cancelled times or failing to find the final agreement.

Fine-Tuned Model: Correctly identifies the final agreed time (Wednesday @ 11 AM) and captures the specific action item (Alice bringing the Q2 report), ignoring the "red herring" cancelled appointments.

7. Analysis and Reflection

What Worked Well

- QLoRA enabled effective fine-tuning with minimal compute.
- The model showed consistent improvements across ROUGE, BLEU, METEOR, and BERTScore.
- Conciseness and structural accuracy improved significantly.

Challenges

- Memory overhead from linear layers required careful quantization and gradient checkpointing.
- GPU availability on free-tier Colab was limited, requiring careful runtime management.

Future Work

- Fine-tuning on the full 13k DialogSum dataset to improve robustness.
- Using Unsloth to achieve 2× faster training.
- Benchmarking against Llama-3-8B to compare reasoning depth and summarization quality.

Conclusion

This project demonstrates that QLoRA-based fine-tuning of a small language model (Phi-2) can yield meaningful improvements in dialogue summarization quality while using only 1.36% trainable parameters. The fine-tuned model produces more concise, structured, and professional summaries than the base model, validating the effectiveness of parameter-efficient fine-tuning for real-world NLP tasks