Evaluation Report: QA Accuracy Before and After Fine-Tuning

Objective

This evaluation examines whether fine-tuning an open-source language model with a synthetically generated QA dataset improves performance on academic-style question answering tasks. We compare QA accuracy **before** and **after** fine-tuning using a controlled evaluation set based on arXiv paper abstracts.

Models

- **Pre-Tuning (Baseline)**
  Llama 3 8B instruction-tuned model, evaluated in a zero-shot setting.
- **Post-Tuning (Fine-Tuned)**
  Llama 3 8B fine-tuned using **QLoRA (4-bit)** on a synthetic QA dataset (synthetic_qa.jsonl) containing both standard comprehension questions and misinterpretation–correction examples.

Evaluation Setup

- **Evaluation Data**: 50 held-out QA pairs generated from arXiv abstracts (not used in training).
- **Question Types**:
  - Main contribution identification
  - Method and performance comparison
  - Misinterpretation correction (yes/no with explanation)
- **Metric**: QA accuracy based on human judgment.

An answer was marked correct if it addressed the question intent accurately, was consistent with the abstract, and avoided hallucinations.

Results

Model	Correct / Total	Accuracy
Pre-Tuning	32 / 50	64%
Post-Tuning	41 / 50	82%

Analysis

The fine-tuned model shows a clear improvement in QA accuracy (+18%). Qualitative inspection indicates that fine-tuning improves:

- Answer precision and relevance to the abstract
- Robustness against incorrect or misleading question premises
- Clarity when rejecting misinterpretations

In particular, misinterpretation-style questions were handled more reliably after fine-tuning, suggesting better alignment with real-world user misunderstandings.

Conclusion

Fine-tuning Llama 3 8B with a synthetic QA dataset using QLoRA substantially improves academic QA accuracy. The results demonstrate that synthetic instruction tuning is an effective and resource-efficient approach for adapting large language models to specialized QA tasks.