



Efficient PEFT for LLMs Under Low-Bit Quantization

Student: Mohamed Miqdad
Course: AI (MSc) — Week 2
Date: 28 Oct 2025

Efficient PEFT for LLMs Under Low-Bit Quantization

1. Introduction

The rapid expansion of Large Language Models (LLMs) has significantly increased the computational cost associated with full fine-tuning, particularly in research environments with limited hardware resources. Parameter-Efficient Fine-Tuning (PEFT) has emerged as a practical alternative, enabling model adaptation using a small number of supplementary parameters without modifying the full model weights. In parallel, low-bit quantization has progressed to the point where 4-bit and 8-bit representations can support both training and inference with acceptable levels of precision loss.

The combination of PEFT and quantization has shown considerable promise, yet the performance and stability of recently introduced methods—such as QDyLoRA and LoQT—remain insufficiently examined. This study aims to assess the efficiency and practical trade-offs of these approaches compared to established baselines such as QLoRA.

2. Problem Statement

Although QLoRA demonstrated that 4-bit quantization could support high-quality fine-tuning with minimal hardware requirements, several open questions persist:

- How do newer PEFT methods behave under different quantization levels?
- What performance differences emerge between 4-bit, 6-bit, and 8-bit configurations?
- Do dynamic-rank or quantization-aware PEFT techniques offer measurable advantages over fixed-rank LoRA?
- Which approach provides the most reliable balance between quality and memory efficiency for small-scale academic setups?

Addressing these questions is essential to establishing a clearer understanding of the practical limits and capabilities of low-bit PEFT.

3. Objectives

The study aims to:

1. Benchmark contemporary PEFT techniques under varying quantization settings.

2. Compare QLoRA, QDyLoRA, and LoQT in terms of model quality, memory usage, and training stability.
 3. Evaluate the influence of quantization noise on convergence and performance.
 4. Identify optimal configurations for researchers working with constrained compute.
 5. Produce a reproducible and transparent evaluation suitable for academic use.
-

4. Background and Related Work

4.1 Parameter-Efficient Fine-Tuning

PEFT techniques such as LoRA and its derivatives have become widely adopted due to their ability to reduce computational overhead while preserving strong downstream performance.

Recent extensions—including QLoRA and QDyLoRA—build on this foundation by incorporating quantization into the fine-tuning process.

4.2 Low-Bit Quantization

Low-bit quantization methods (e.g., NF4, INT4, INT8) aim to minimize memory consumption while maintaining sufficient representational capacity for gradient updates.

The stabilization of 4-bit training in particular has attracted attention due to its suitability for consumer-grade GPU hardware.

4.3 Recent Literature

Recent surveys published between 2023 and 2025 highlight the shift toward efficient training methodologies and examine the role of quantization in enabling scalable experimentation.

Several studies have introduced new quantization-aware optimization strategies, but comparative evaluations remain limited.

5. Methodology

Datasets

Evaluation will be conducted on representative NLP tasks, including summarization, instruction-following, and reasoning tasks (e.g., XSum, CNN-DM, GSM8K).

Models

The experiments will employ mid-sized open-access models such as LLaMA-3-8B, Mistral-7B, or Qwen-2-7B, depending on hardware constraints.

Fine-Tuning Methods

The study will compare:

- QLoRA (baseline)
- QDyLoRA
- LoQT
- Full-precision LoRA (reference model)

Evaluation Metrics

Key metrics include:

- Model accuracy and task-specific performance
- Memory usage and peak VRAM consumption
- Training time and throughput
- Convergence behavior
- Sensitivity to bit-width changes

Experimental Setup

Training will be performed on one or two consumer-grade GPUs, ensuring reproducibility and accessibility for academic environments.

6. Expected Contribution

This research intends to provide:

- A clear comparative analysis of recent PEFT+quantization methods.
 - Practical guidance for selecting efficient setups under limited hardware.
 - Empirical evidence illustrating the relationship between bit-width, model quality, and training stability.
 - A benchmark that can be replicated and extended by other researchers.
-

7. Preliminary Conclusion

Early findings from related work suggest that combining PEFT with low-bit quantization allows substantial reductions in memory requirements with only modest performance degradation.

While QLoRA remains a widely adopted baseline, newer methods such as QDyLoRA and LoQT introduce modifications that may improve stability and adaptability, particularly when working at very low bit-widths.

The planned experiments will help clarify these differences and identify configurations that offer the strongest balance between efficiency and performance.

8. Next Steps

- Implement the QLoRA baseline to establish a performance reference.
- Collect preliminary results for VRAM consumption and convergence behavior.
- Integrate QDyLoRA and LoQT into the training pipeline.
- Conduct initial evaluations on a compact dataset.
- Prepare findings for inclusion in the next project report.