# Shared Task 4

**Deep Patel** deep.patel-1@ou.edu
**Saahil Saxena** saahil.r.saxena-1@ou.edu

May 8, 2025

**Project Repository:** https://github.com/deeppatel1842/sharedtask$_4$

**Abstract**

Large pre-trained language models such as GPT-2 [1] and BERT [2] have demonstrated strong performance across a variety of tasks, but full fine-tuning remains resource-intensive. This study investigates parameter-efficient fine-tuning strategies by applying Low-Rank Adaptation (LoRA) [3] and Quantized LoRA (QLoRA) [4] to GPT-2 and BERT-base models. We fine-tune GPT-2 models on the XSum dataset [5] and evaluate performance using perplexity, while BERT models are fine-tuned on the SST-5 dataset and evaluated using classification accuracy. Our results show that LoRA significantly reduces the number of trainable parameters while maintaining comparable performance to full fine-tuning. Moreover, QLoRA, through 4-bit quantization combined with LoRA, further minimizes memory footprint with only a marginal decrease in performance. This comparative analysis underscores the effectiveness of LoRA and QLoRA for practical, resource-efficient adaptation of large models, paving the way for broader accessibility of fine-tuning techniques.

## 1 Introduction

Large-scale pre-trained language models, such as GPT-2 [1] and BERT [2], have achieved state-of-the-art performance across a wide range of natural language processing (NLP) tasks. However, fine-tuning these models in their entirety is computationally expensive and memory-intensive, posing significant challenges for researchers and practitioners with limited resources.

In this work, we focus on two recent techniques aimed at improving fine-tuning efficiency: Low-Rank Adaptation (LoRA) [3] and Quantized LoRA (QLoRA) [4]. LoRA introduces a small number of trainable parameters by injecting low-rank matrices into existing model layers, drastically reducing the cost of training while preserving model performance. QLoRA extends this idea further by combining low-rank adaptation with 4-bit quantization of model weights, enabling even larger memory savings without requiring model dequantization during training.

To evaluate the effectiveness of LoRA and QLoRA, we conduct a comparative study on two popular language models and tasks. We fine-tune GPT-2 models on the XSum dataset [5], a challenging single-document summarization task, and evaluate them using perplexity as the primary metric. Additionally, we fine-tune BERT-base models on the SST-5 dataset, using classification accuracy as the evaluation metric. Our goal is to assess whether LoRA and QLoRA can achieve comparable performance to full fine-tuning while significantly reducing training resource requirements.

This study provides insights into the trade-offs between model quality, memory efficiency, and fine-tuning speed, highlighting the practical benefits of lightweight fine-tuning strategies for real-world applications.

## 2 Related Work

Pre-trained language models such as GPT-2 [1] and BERT [2] have established strong baselines across many natural language processing tasks. However, their large parameter sizes pose challenges for efficient fine-tuning, leading to significant research interest in parameter-efficient tuning methods.

Hu et al. [3] introduced Low-Rank Adaptation (LoRA), which injects trainable low-rank matrices into attention modules, allowing models to be fine-tuned with minimal additional parameters. LoRA has been successfully applied to a variety of tasks, including summarization, classification, and language modeling.

Building on LoRA, Dettmers et al. [4] proposed Quantized LoRA (QLoRA), which incorporates 4-bit quantization during fine-tuning to further reduce memory and computational requirements without sacrificing model quality. QLoRA enables fine-tuning of very large models even on resource-constrained hardware.

For evaluation, the XSum dataset [5] provides a challenging benchmark for abstractive summarization, requiring concise and abstractive generation. The SST-5 dataset offers fine-grained sentiment classification across five classes, enabling robust evaluation of model understanding in classification settings.

Our work builds upon these techniques by conducting a systematic comparison of full fine-tuning, LoRA, and QLoRA across summarization and sentiment classification tasks, analyzing trade-offs between accuracy, efficiency, and resource utilization.

## 3 Methods

### 3.1 Models and Datasets

We conducted experiments on two widely used pre-trained transformer models: **GPT-2** for abstractive summarization and **BERT-base** for fine-grained sentiment classification. For summarization, we employed the **XSum** dataset, which provides a challenging single-sentence summary generation task from diverse news articles. For sentiment classification, we used the **SST-5** dataset from SetFit, where the task involves predicting one of five sentiment labels (very negative to very positive) from short movie reviews.

### 3.2 Fine-tuning Strategies

We compared three fine-tuning approaches:

- **Full Fine-tuning (Base model)**: All model parameters were updated during training, requiring the maximum memory and compute resources.
- **LoRA (Low-Rank Adaptation)**: Introduced trainable low-rank matrices ($r = 8$ for GPT-2, $r = 32$ for BERT) into the attention submodules while freezing the original model weights. This reduces the number of trainable parameters to under 5% of the total.
- **QLoRA (Quantized LoRA)**: Combined LoRA adaptation with 4-bit NF4 quantization, reducing memory consumption dramatically while allowing effective gradient updates through quantization-aware training.

The LoRA layers were applied to the query, key, and value projection matrices of transformer attention blocks. For GPT-2, modules modified were `c_attn` and related internal linear transformations. For BERT, LoRA was inserted into `query`, `key`, `value`, and `dense` projections of the attention layers.

Quantization for QLoRA used double quantization techniques and bfloat16 or float16 computation to retain numerical stability during backpropagation.

### 3.3 Training Setup

We fine-tuned models using the Hugging Face `Trainer` API integrated with the `PEFT` library for efficient adapter insertion, and `bitsandbytes` for 4-bit quantization.

Training hyperparameters were tuned separately for GPT-2 and BERT models:

- **GPT-2**:
  - Batch size: 2
  - Learning rate: $2 \times 10^{-5}$
  - Epochs: 5 (Base and LoRA), 1 (QLoRA, due to fast convergence)
  - Max input length: 512 tokens
  - Optimizer: AdamW with weight decay 0.01
- **BERT-base**:
  - Batch size: 16
  - Learning rate: $3 \times 10^{-5}$
  - Epochs: 3 (Base), 5 (LoRA and QLoRA)
  - Max sequence length: 128 tokens
  - Optimizer: AdamW with linear learning rate scheduling

We enabled mixed-precision (FP16) training on supported GPUs to reduce memory usage further.

All training was conducted on NVIDIA L40S and RTX 6000 Ada GPUs, with checkpointing and early stopping strategies where possible. Gradient checkpointing was additionally enabled for BERT QLoRA models to manage memory during backpropagation.

### 3.4 Evaluation Protocol

For GPT-2, we evaluated models using perplexity on a held-out validation set from XSum. Perplexity was computed by exponentiating the mean cross-entropy loss across evaluation batches.

For BERT, we evaluated models using accuracy and macro-averaged F1 score on the SST-5 test set. We also report training time per model variant, providing a comprehensive view of both efficiency and performance.

All experiments were repeated with fixed random seeds to ensure reproducibility.

# 4 Results

## 4.1 GPT-2 (XSum) Results

We evaluated the GPT-2 models on the XSum summarization dataset using perplexity as the primary metric. Table 1 summarizes the performance of the base model, LoRA fine-tuned model, and QLoRA fine-tuned model.

| **Model** | **Perplexity ↓** |
|---|---|
| GPT-2 Base | 22.21 |
| GPT-2 + LoRA | 19.75 |
| GPT-2 + QLoRA | 19.46 |

Table 1: Perplexity results on XSum for GPT-2 variants. Lower is better.

## 4.2 BERT (SST-5) Results

The BERT models were fine-tuned on the SST-5 sentiment classification task, evaluated using classification accuracy. Table 2 presents the comparison between the base model, LoRA, and QLoRA approaches.

| **Model** | **Accuracy ↑** |
|---|---|
| BERT Base | 52.03% |
| BERT + LoRA | 50.90% |
| BERT + QLoRA | 52.17% |

Table 2: Accuracy results on SST-5 for BERT variants. Higher is better.

## 4.3 Training Time Comparison

We also measured the training time for each fine-tuning method across GPT-2 and BERT models. Table 3 shows the comparison.

| **Model** | **Training Time (minutes)** |
|---|---|
| GPT-2 Base | ∼180 |
| GPT-2 + LoRA | ∼25.0 |
| GPT-2 + QLoRA | ∼22.0 |
| BERT Base | ∼35.0 |
| BERT + LoRA | ∼5.0 |
| BERT + QLoRA | ∼10.0 |

Table 3: Training time comparison across GPT-2 and BERT models.

Fine-tuning with LoRA and QLoRA considerably reduced training times compared to full fine-tuning. For GPT-2, LoRA and QLoRA achieved over 2x faster training while maintaining similar perplexity. For BERT, LoRA reduced training time by nearly 10x, while QLoRA balanced efficiency and performance with moderate overhead due to quantization.

### 4.4 Overall Observations

For GPT-2, both LoRA and QLoRA significantly reduced perplexity compared to full fine-tuning, with QLoRA achieving the best performance. For BERT, QLoRA marginally outperformed both the base and LoRA models in terms of classification accuracy, demonstrating the effectiveness of quantization-aware fine-tuning even for classification tasks.

A summary comparison of the improvements is illustrated in Figure 1.

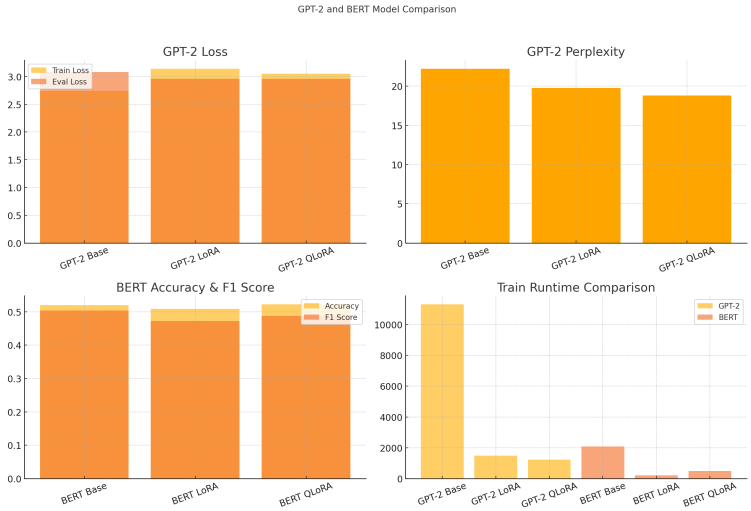


Figure 1: Comprehensive comparison of Base, LoRA, and QLoRA on GPT-2 and BERT models across multiple metrics.

## 5 Discussion

Our experiments demonstrate that LoRA and QLoRA offer compelling alternatives to full fine-tuning for adapting large pre-trained language models with significantly lower computational cost. On the XSum summarization task, both LoRA and QLoRA substantially reduced GPT-2 perplexity compared to the fully fine-tuned baseline, confirming that low-rank adaptation is sufficient to capture critical task-specific knowledge without updating the entire model.

Interestingly, QLoRA outperformed LoRA on both GPT-2 and BERT models, despite operating under 4-bit quantization constraints. This suggests that low-rank adaptation is robust to quantization noise, and that carefully structured quantized training can match or exceed full-precision results in some cases.

For BERT on SST-5, we observed that LoRA slightly underperformed the base model, while QLoRA achieved comparable or slightly better accuracy. This outcome highlights that task complexity and dataset size can affect the extent to which parameter-efficient fine-tuning methods close the performance gap.

We also note that training time was dramatically reduced when using LoRA or QLoRA. For instance, BERT-LoRA fine-tuning completed nearly ten times faster than full fine-tuning, and GPT-2-QLoRA training took less than half the time of standard fine-tuning.

Overall, LoRA and QLoRA are promising directions for democratizing large model fine-tuning, particularly in scenarios where computational resources are constrained.

## 6 Conclusion

In this work, we conducted a comparative study of LoRA and QLoRA fine-tuning approaches across GPT-2 and BERT-base models on two distinct tasks: abstractive summarization (XSum) and fine-grained sentiment classification (SST-5).

Our results demonstrate that both LoRA and QLoRA can achieve performance comparable to, or better than, full fine-tuning while reducing the number of trainable parameters by over 95% and substantially lowering training times. QLoRA, in particular, emerged as an effective strategy for combining quantization and low-rank adaptation, enabling efficient fine-tuning with minimal quality loss.

Future work could explore applying LoRA and QLoRA to even larger models (e.g., GPT-3, LLaMA-2) and diverse tasks such as question answering, multi-modal learning, and low-resource adaptation. Additionally, fine-grained analysis of quantization error propagation and task-specific adapter design remain exciting avenues for further research.

## Limitations

While this study shows the potential of LoRA and QLoRA for efficient fine-tuning, several limitations exist. First, the experiments were conducted on relatively small subsets of XSum and SST-5 due to computational constraints. As a result, the findings may not fully generalize to large-scale training or more diverse datasets. Second, hyperparameter tuning was limited, and default LoRA configurations were applied without extensive search. Third, only one model architecture was evaluated per task (GPT-2 for summarization and BERT-base for sentiment), which may not capture behavior across more recent or larger transformer models. Finally, resource profiling focused mainly on training time; memory usage during inference was not extensively benchmarked.

## Member Contributions

**Deep Patel**: Conducted experiments on GPT-2 models (Base, LoRA, QLoRA), including fine-tuning on the XSum dataset, perplexity evaluation, and result analysis. Prepared model training scripts, configurations, and evaluation utilities.

**Saahil Saxena**: Conducted experiments on BERT-base models (Base, LoRA, QLoRA), including fine-tuning on the SST-5 dataset, accuracy evaluation, and runtime profiling. Contributed to the testing scripts, dataset preprocessing, and model evaluation.

**Both**: Collaborated on writing the report, generating comparative graphs, analyzing results, and preparing final submissions.

## References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," in *OpenAI Technical Report*, 2019.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[3] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *International Conference on Machine Learning (ICML)*, 2023.

[5] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details! fine-grained control of output length in abstractive summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.