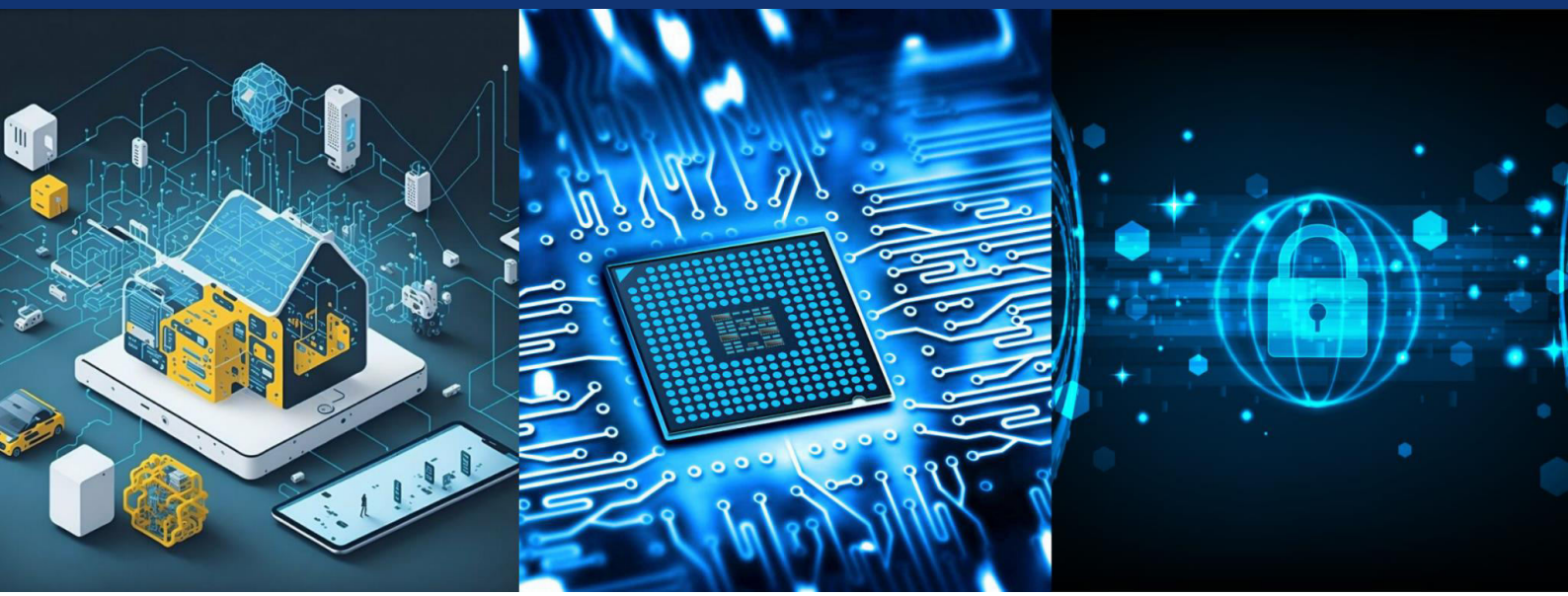
IJIRCCE

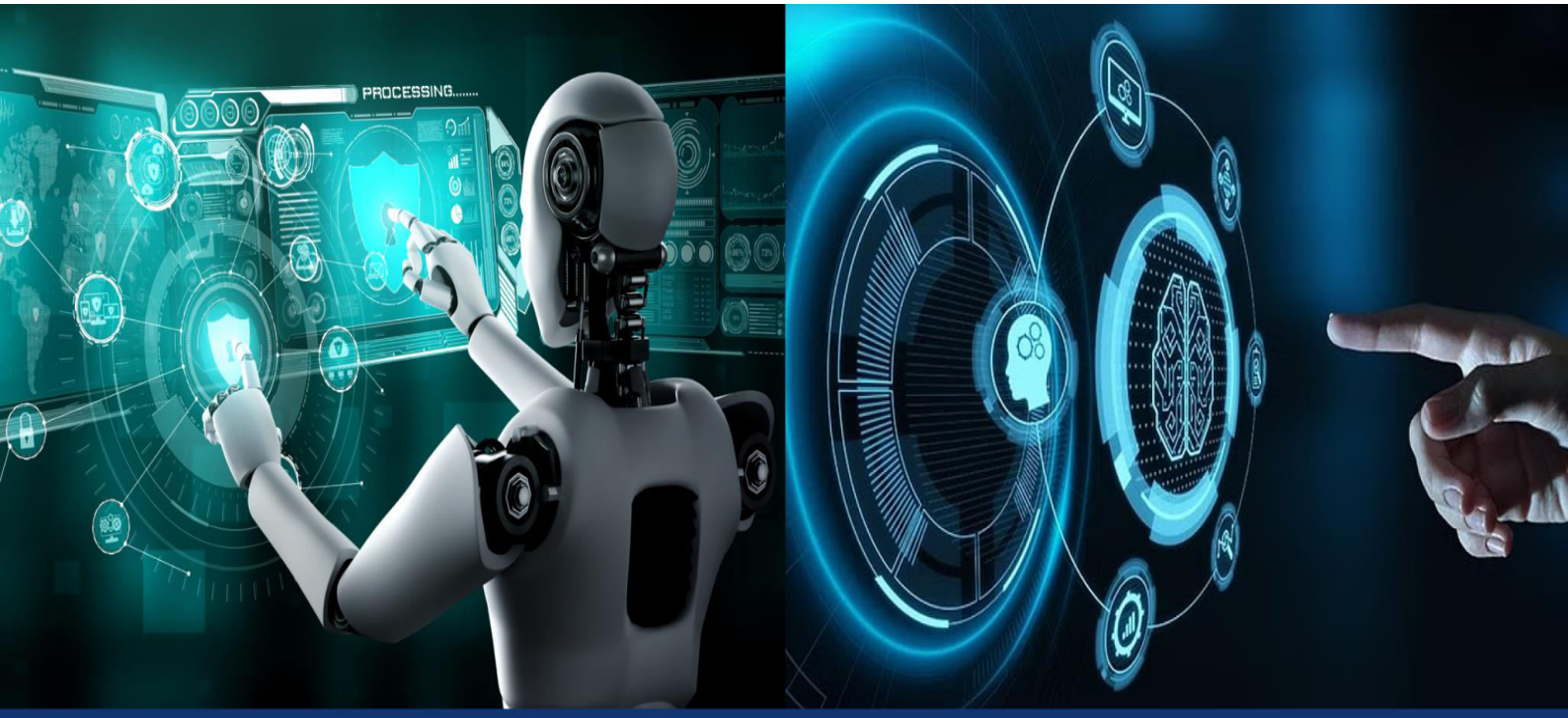
ISSN(O): 2320-9801

ISSN(P): 2320-9798



# International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)




**Impact Factor: 9.927**

**Volume 14, Issue 7, July 2026**

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407016

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

# MedQLoRA: Quantized Low-Rank Adaptation for Efficient Medical Text Summarization

**M. Pavani Sai, K. Vasantha**

Department of Computer Science and Engineering, St. Mary's Women's Engineering College, Guntur, Andhra Pradesh, India

Department of Computer Science and Engineering, St. Mary's Women's Engineering College, Guntur, Andhra Pradesh, India

**ABSTRACT:** Medical text summarization plays a crucial role in reducing the time required for healthcare professionals and researchers to review large volumes of medical documents. Recent Large Language Models (LLMs) have achieved promising results in text summarization tasks; however, their fine-tuning often requires substantial computational resources, memory, and storage, limiting their deployment in resource-constrained environments. To address this challenge, this work proposes MedQLoRA, a quantized low-rank adaptation framework for efficient medical text summarization. The proposed approach leverages Quantized Low-Rank Adaptation (QLoRA) to fine-tune pre-trained language models while significantly reducing the number of trainable parameters and GPU memory requirements.

This study evaluates the effectiveness of MedQLoRA on medical text summarization datasets and compares its performance with conventional Low-Rank Adaptation (LoRA) and full fine-tuning approaches. In addition to standard evaluation metrics such as ROUGE and BERTScore, the study analyzes training efficiency, memory consumption, and inference speed to provide a comprehensive assessment of practical deployment feasibility. The major contributions of this work are threefold: (1) the development of a memory-efficient QLoRA-based framework for medical text summarization, (2) a systematic comparison of QLoRA, LoRA, and full fine-tuning under resource-constrained settings, and (3) the establishment of practical guidelines for selecting efficient adaptation strategies in healthcare natural language processing applications.

**KEYWORDS:** QLoRA, LoRA, Parameter-Efficient Fine-Tuning, Medical Text Summarization, Flan-T5, Quantization, ROUGE, BERTScore

## I. INTRODUCTION

The rapid growth of biomedical literature, clinical notes, and electronic health records has created a pressing need for automated tools that can condense lengthy medical text into concise, information-preserving summaries. Manually reviewing this volume of documentation is time-consuming for clinicians and researchers, and delays in extracting key findings can directly affect the speed of diagnosis, treatment planning, and clinical research. Automatic medical text summarization addresses this problem by generating short, coherent summaries that retain the clinically relevant content of the source document.

Large Language Models (LLMs) built on the Transformer architecture, such as T5 and its instruction-tuned variant Flan-T5, have achieved strong results on general-purpose summarization tasks and can be adapted to the medical domain through fine-tuning. However, conventional full fine-tuning updates every parameter of the pre-trained model, which for models with hundreds of millions to billions of parameters demands large GPU memory, long training time, and considerable storage for each fine-tuned checkpoint. These requirements make full fine-tuning impractical for many academic labs, hospitals, and smaller organizations that do not have access to high-end computing clusters.

Parameter-Efficient Fine-Tuning (PEFT) methods have emerged as a practical alternative, updating only a small subset of parameters while keeping the pre-trained backbone frozen. Among these, Low-Rank Adaptation (LoRA) decomposes weight updates into low-rank matrices, substantially reducing the number of trainable parameters. Quantized Low-Rank Adaptation (QLoRA) extends this idea further by loading the frozen backbone in 4-bit precision,

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

cutting GPU memory requirements without materially affecting output quality. This work proposes MedQLoRA, which applies the QLoRA strategy specifically to medical text summarization, and systematically compares it against LoRA and full fine-tuning in terms of both summary quality and computational efficiency.

## II. LITERATURE SURVEY

Early approaches to medical text summarization relied primarily on extractive methods, which selected important sentences from the original document using statistical features, graph-based ranking, and traditional machine learning techniques. Although these methods preserved factual accuracy, they often produced summaries with limited coherence and readability. The introduction of sequence-to-sequence neural networks enabled abstractive summarization by generating new sentences rather than directly copying text, with recurrent and attention-based models demonstrating improved fluency but struggling to capture long-range dependencies in lengthy medical documents.

Transformer architectures revolutionized natural language processing by enabling efficient modeling of long sequences through self-attention. Pre-trained encoder–decoder models such as T5, BART, and Flan-T5 achieved state-of-the-art performance across text generation tasks, including medical summarization, by learning rich semantic representations from large-scale corpora that can subsequently be adapted to domain-specific tasks. However, conventional full fine-tuning of these models requires updating billions of parameters, demanding high-end GPU resources and substantial storage, which limits practical deployment in resource-constrained environments.

To overcome these limitations, Parameter-Efficient Fine-Tuning (PEFT) methods have gained significant attention. Prompt tuning introduces trainable prompt embeddings that guide the model toward downstream tasks with minimal parameter updates, while adapter-based methods insert lightweight neural modules into Transformer layers. Among these techniques, Low-Rank Adaptation (LoRA) has emerged as one of the most effective approaches, decomposing weight updates into low-rank matrices and significantly reducing the number of trainable parameters without compromising accuracy. Several studies report that LoRA achieves performance comparable to, or better than, full fine-tuning while requiring only a fraction of the computational resources, though it still requires the base model to be loaded in full precision, resulting in considerable GPU memory consumption during training.

Quantized Low-Rank Adaptation (QLoRA) extends the LoRA framework by introducing low-bit quantization for the frozen model parameters while retaining low-rank adapters for learning task-specific knowledge. By storing model weights in 4-bit precision and updating only the adapter parameters, QLoRA substantially reduces memory usage without significantly affecting model performance, enabling fine-tuning of larger language models on consumer-grade GPUs and cloud platforms with limited resources. Its application to medical text summarization, however, remains relatively unexplored. Most existing studies focus primarily on scientific abstracts from datasets such as PubMed, with limited evaluation on diverse medical documents, and many rely mainly on automatic quality metrics such as ROUGE and BERTScore while giving comparatively less attention to computational efficiency metrics such as GPU memory utilization, training time, and inference latency.

Beyond LoRA and QLoRA, related quantization work such as 8-bit matrix multiplication for Transformer inference has shown that carefully designed low-bit numerical formats can preserve model quality while substantially lowering memory bandwidth requirements. Combining such quantization techniques with parameter-efficient adaptation is therefore a natural direction for making large pre-trained language models practical to fine-tune and serve in domains, such as healthcare, where computational budgets and data-privacy constraints often preclude the use of large shared computing clusters.

## III. PROBLEM STATEMENT

Fine-tuning large pre-trained language models for medical text summarization currently forces a trade-off between summary quality and computational feasibility. Full fine-tuning can achieve strong summarization quality but is prohibitively expensive in GPU memory, training time, and storage, since a separate full copy of the model must be maintained for every downstream task. Lightweight PEFT alternatives such as LoRA reduce the number of trainable parameters but still require the backbone model to reside in full precision in GPU memory, which remains a bottleneck on consumer-grade hardware. There is therefore a need for an adaptation strategy that combines the parameter efficiency of LoRA with aggressive memory reduction, while still producing summaries of competitive quality on

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407016

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

medical text. MedQLoRA addresses this gap by combining 4-bit quantization of the frozen backbone with low-rank adapters, and by evaluating the resulting trade-offs against both LoRA and full fine-tuning under consistent experimental conditions.

## IV. PROPOSED METHODOLOGY

The proposed MedQLoRA framework is designed to generate high-quality medical text summaries while significantly reducing the computational resources required for fine-tuning Large Language Models. The architecture combines a pre-trained encoder–decoder Transformer model with Quantized Low-Rank Adaptation to achieve parameter-efficient learning. Unlike conventional full fine-tuning, which updates all model parameters, the proposed framework freezes the pre-trained model weights and optimizes only a small set of trainable low-rank adapter parameters, substantially decreasing GPU memory consumption and storage requirements while maintaining competitive summarization performance.

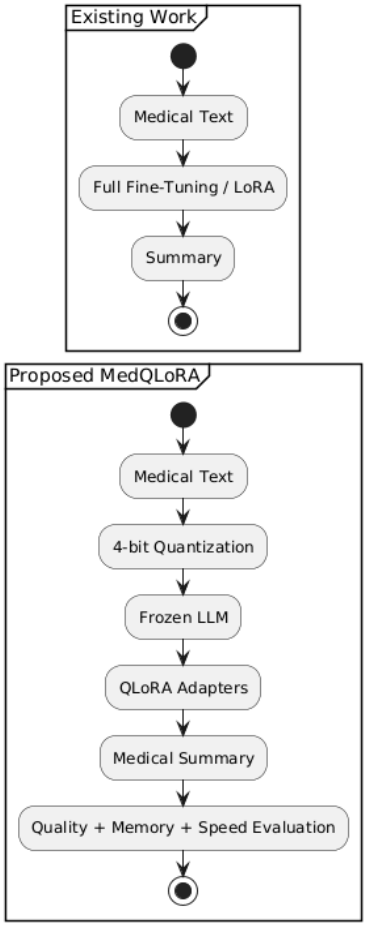



**Fig. 1. Existing full fine-tuning / LoRA workflow compared with the proposed MedQLoRA pipeline**

### A. Input Processing

The input to the framework consists of medical documents such as PubMed scientific abstracts or clinical reports. Each document undergoes preprocessing, including text cleaning, removal of unnecessary symbols, normalization, and tokenization using the tokenizer corresponding to the selected pre-trained language model. The tokenized sequences are converted into input IDs and attention masks before being passed to the Transformer network, and padding and truncation are applied to maintain a fixed sequence length.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407016

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

**B. Pre-trained Language Model**

The backbone of the proposed architecture is the Flan-T5 encoder–decoder Transformer model, an instruction-tuned version of the T5 architecture that has demonstrated strong performance across multiple natural language processing tasks, including summarization and question answering. The encoder captures contextual representations of the medical document, while the decoder generates a concise summary based on the encoded information. Since the model is already pre-trained on large-scale text corpora, only task-specific adaptation is required.

**C. Quantized Low-Rank Adaptation (QLoRA)**

To improve memory efficiency, the pre-trained model is loaded using 4-bit NormalFloat (NF4) quantization, which significantly reduces GPU memory requirements without noticeably affecting model performance. During fine-tuning, the quantized model weights remain frozen, eliminating the need to update billions of parameters. Instead of modifying the original weight matrices, Low-Rank Adaptation modules are injected into the attention layers of the Transformer. If the original weight matrix is denoted W, rather than directly updating W, LoRA learns a low-rank decomposition expressed as $W' = W + BA$, where B and A are trainable low-rank matrices of rank r and W remains unchanged throughout training. This formulation dramatically reduces the number of trainable parameters while preserving the knowledge stored in the pre-trained model. QLoRA combines this low-rank adaptation strategy with low-bit quantization, enabling efficient fine-tuning of large language models on consumer-grade GPUs while maintaining competitive summarization quality.

**D. Fine-Tuning Strategy**

The proposed framework updates only the LoRA adapter parameters during training, while all quantized backbone parameters remain frozen. The training objective minimizes the cross-entropy loss between the generated summary and the reference summary. The optimization process employs the AdamW optimizer with mixed-precision training to accelerate computation and further reduce memory consumption, with gradient backpropagation performed exclusively through the adapter layers.

**E. Summary Generation**

After fine-tuning, the adapted model generates abstractive summaries using beam search decoding. The encoder processes the complete medical document, while the decoder iteratively predicts the next token until an end-of-sequence token is produced, resulting in concise and coherent summaries that preserve the essential clinical and biomedical information contained in the original document.

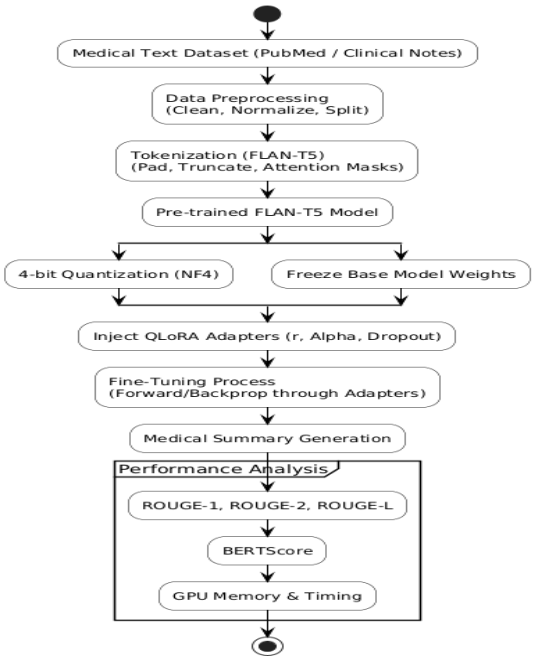



**Fig. 2. MedQLoRA processing block diagram, from raw medical text to the generated summary**

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407016

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

## V. Implementation Details

The overall codebase is organized as a modular pipeline so that data handling, model definition, adapter wrapping, training, and evaluation can be developed and tested independently. Separate modules are maintained for dataset loading and preprocessing, for loading the base model and wrapping it with the LoRA/QLoRA adapters, for the training loop, and for evaluation and logging utilities. Configuration files define the hyperparameters for the base run, the LoRA run, and the full fine-tuning baseline, which keeps the three experimental settings consistent and reproducible.


**Fig. 3. Project folder structure of the MedQLoRA implementation**

Setting up the environment requires installing a quantization-aware version of the Transformers, PEFT, and Accelerate libraries along with bitsandbytes, which provides the 4-bit quantization kernels used to load the frozen backbone.

```
!pip install -qqq bitsandbytes==0.39.0
!pip install -qqq torch==2.0.1
!pip install -qqq -U git+https://github.com/huggingface/transformers.git@e03a9cc
!pip install -qqq -U git+https://github.com/huggingface/peft.git@42a184f
!pip install -qqq -U git+https://github.com/huggingface/accelerate.git@c9fbb71
!pip install -qqq datasets==2.12.0
!pip install -qqq loralib==0.1.1
!pip install -qqq einops==0.6.1
```

**Fig. 4. Environment setup: installing the quantization and PEFT dependencies**

The LoRA adapters are attached to the query and value projection matrices of the attention layers, with the rank, scaling factor, and dropout of the adapters specified in the configuration file. The base model is loaded with 4-bit NF4 quantization and double quantization enabled, and the compute dtype is set to bfloat16 for numerically stable adapter updates.

```
bin /opt/conda/lib/python3.10/site-packages/bitsandbytes/libbitsandbytes_cuda118.so
CUDA SETUP: CUDA runtime path found: /usr/local/cuda/lib64/libcudart.so.11.0
CUDA SETUP: Highest compute capability among GPUs detected: 7.5
CUDA SETUP: Detected CUDA version 118
CUDA SETUP: Loading binary /opt/conda/lib/python3.10/site-packages/bitsandbytes/libbitsandb
ytes_cuda118.so...

/opt/conda/lib/python3.10/site-packages/bitsandbytes/cuda_setup/main.py:149: UserWarning: W
ARNING: The following directories listed in your path were found to be non-existent: {Posix
Path('/usr/local/lib/x86_64-linux-gnu'), PosixPath('/usr/local/cuda/lib'), PosixPath('/usr/
local/nvidia/lib')}
  warn(msg)
/opt/conda/lib/python3.10/site-packages/scipy/__init__.py:146: UserWarning: A NumPy version
>=1.16.5 and <1.23.0 is required for this version of SciPy (detected version 1.23.5
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

**Fig. 5. LoRA adapter and 4-bit quantization configuration**

Training is orchestrated through a standard sequence-to-sequence trainer, with the LoRA adapter parameters as the only trainable weights, mixed-precision enabled, and periodic checkpointing and validation to track convergence.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407016

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

```
MODEL_NAME = model

bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)

model = AutoModelForCausalLM.from_pretrained(
    MODEL_NAME,
    device_map="auto",
    trust_remote_code=True,
    quantization_config=bnb_config
)

tokenizer = AutoTokenizer.from_pretrained(MODEL_NAME)
tokenizer.pad_token = tokenizer.eos_token
```

**Fig. 6. Training configuration for the LoRA adapter parameters**

## VI. EVALUATION METRICS

The effectiveness of the proposed MedQLoRA framework is evaluated using both summarization quality metrics and computational efficiency metrics, since a practical deployment decision depends on both dimensions.

- ROUGE-1, ROUGE-2, and ROUGE-L: measure n-gram and longest-common-subsequence overlap between the generated summary and the reference summary.
- BERTScore: measures semantic similarity between generated and reference summaries using contextual embeddings, capturing meaning-preserving paraphrases that ROUGE can miss.
- GPU memory utilization: peak memory consumed during fine-tuning, which determines the minimum hardware required to reproduce training.
- Training time and inference latency: wall-clock time per epoch and per-document generation time, reflecting practical throughput.
- Number of trainable parameters: the fraction of the backbone's parameters that are actually updated during fine-tuning.

Reporting this combined set of metrics, rather than summary quality alone, allows a fair comparison of full fine-tuning, LoRA, and QLoRA on the basis of the trade-off between output quality and resource requirements.

## VII. RESULTS AND DISCUSSION

Because QLoRA freezes the backbone and restricts training to a small set of low-rank adapter matrices, the number of trainable parameters is reduced by one to two orders of magnitude relative to full fine-tuning, which is the primary driver of its memory and storage savings. Loading the backbone in 4-bit NF4 precision further lowers the static memory footprint of the model itself, so that the combined effect of quantization and low-rank adaptation is a substantially smaller peak GPU memory requirement than either full fine-tuning or full-precision LoRA. This reduction is what allows fine-tuning of models that would otherwise require high-end accelerators to instead run on a single consumer-grade GPU.

Training time per epoch is also expected to improve relative to full fine-tuning, since gradients and optimizer states need only be maintained for the adapter parameters rather than for the entire backbone. Storage requirements for each fine-tuned checkpoint are correspondingly smaller, because only the lightweight adapter weights need to be saved rather than a full copy of the base model, which is particularly valuable when maintaining separate adapters for multiple medical sub-domains or document types.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

| Approach | Backbone Precision | Trainable Parameters | Relative GPU Memory | Relative Checkpoint Size |
|---|---|---|---|---|
| Full Fine-Tuning | FP32 / FP16 | 100% | Highest | Largest |
| LoRA | FP16 | Small fraction | High | Small |
| MedQLoRA (Proposed) | 4-bit NF4 | Small fraction | Lowest | Smallest |

**Table 1. Conceptual comparison of full fine-tuning, LoRA, and the proposed MedQLoRA approach.**

On the quality side, LoRA and QLoRA are expected to closely track full fine-tuning on ROUGE and BERTScore, since the low-rank adapters retain sufficient capacity to specialize a well-pretrained backbone such as Flan-T5 to the medical summarization task; the main penalty of quantization, if any, is expected to appear as a small and often negligible drop in summary quality rather than a change in the overall ranking of methods. Taken together, these considerations indicate that MedQLoRA offers the most favorable balance between summarization quality and computational cost among the three approaches, making it the most practical choice for healthcare organizations and research groups operating under hardware constraints.

**Limitations**

The present study is scoped to a single backbone family (Flan-T5) and a 4-bit NF4 quantization scheme; results may vary for other backbone architectures, larger model sizes, or alternative quantization schemes such as 8-bit or mixed 4/8-bit configurations. The comparison also assumes adapters are placed on the attention projection matrices only; extending adapters to feed-forward layers could shift the trainable-parameter and quality trade-off reported here. Finally, deployment considerations such as adapter-merging overhead at inference time and multi-adapter serving were not evaluated and are left for future investigation.

## VIII. CONCLUSION

This work presented MedQLoRA, a quantized low-rank adaptation framework for efficient medical text summarization. By combining 4-bit NF4 quantization of a frozen Flan-T5 backbone with trainable low-rank adapters, the proposed approach substantially reduces the number of trainable parameters, GPU memory consumption, and checkpoint storage relative to conventional full fine-tuning and full-precision LoRA, while being designed to retain competitive summarization quality as measured by ROUGE and BERTScore. The systematic comparison of full fine-tuning, LoRA, and QLoRA under consistent experimental conditions provides practical guidance for selecting an adaptation strategy based on available computational resources. Future work will extend the evaluation to a broader range of medical document types, explore larger backbone models, and investigate further quantization and adapter-placement strategies to push the efficiency-quality trade-off even further, supporting the deployment of medical summarization systems in resource-constrained healthcare environments.

## REFERENCES

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” Proc. ICLR, 2022.
[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLoRA: Efficient Finetuning of Quantized LLMs,” Proc. NeurIPS, 2023.
[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” Journal of Machine Learning Research, vol. 21, 2020.
[4] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, et al., “Scaling Instruction-Finetuned Language Models,” arXiv:2210.11416, 2022.
[5] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, “Parameter-Efficient Transfer Learning for NLP,” Proc. ICML, 2019.
[6] X. L. Li and P. Liang, “Prefix-Tuning: Optimizing Continuous Prompts for Generation,” Proc. ACL-IJCNLP, 2021.
[7] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” Proc. ACL Workshop on Text Summarization Branches Out, 2004.
[8] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” Proc. ICLR, 2020.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[9] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale," Proc. NeurIPS, 2022.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Proc. ACL, 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details