Efficient Optimization of Multimodal Language Models for Processing Radiological Data using the ROCO Dataset

Danilo Felipe Neto
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Ceará
danilo.neto@alu.ufc.br

Anderson Luis Amaral
Scoras Academy
Scoras Digital
Barueri, São Paulo
anderson@scoras.com.br

Victor Hugo C. de Albuquerque
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Ceará
victor.albuquerque@ieee.org

***Abstract*—This study investigates the computational efficiency and performance trade-offs of model optimization techniques like QLoRA, Knowledge Distillation (KD), and Pruning, when applied to large language models (LLMs) and small language models (SLMs) for radiology question-answering tasks. Using the ROCOv2 multimodal dataset, we systematically compare baseline models against their fine-tuned and compressed counterparts. The primary goal is to evaluate whether such methods can significantly reduce memory and computational demands while maintaining acceptable accuracy, enabling deployment on edge devices and in low-resource clinical environments. Experimental results show that SLMs enhanced with QLoRA retain competitive accuracy while reducing GPU usage by up to 80%, and that combining KD and Pruning further improves inference speed and hardware efficiency making these models viable for real world radiological decision support at the edge computing devices.**

***Index Terms*—Large Language Models, Small Language Models, QLoRA, Knowledge Distillation, Pruning, Radiology, Fine-tuning, Edge Computing.**

I. Introduction

The increasing adoption of large language models (LLMs) in healthcare has accelerated the need for more computationally efficient solutions. While LLMs such as GPT-4 and PaLM-2 have demonstrated state-of-the-art performance in natural language understanding and multimodal reasoning [1], their substantial memory and processing requirements pose a barrier to deployment in clinical environments, especially on edge devices with limited hardware capacity [2]. To address this challenge, this study explores whether small language models (SLMs), when enhanced through modern fine-tuning and compression techniques including Quantized Low-Rank Adaptation (QLoRA) [3], Knowledge Distillation (KD) [4], and Pruning, can approximate the performance of LLMs in radiology-specific visual question answering (VQA) tasks. The primary objective is to evaluate whether these optimized models can preserve clinical interoperability and linguistic precision while significantly reducing computational overhead. We perform a comparative analysis of these strategies using the ROCOv2 dataset, focusing on key evaluation metrics relevant to both accuracy and efficiency. This approach aims to support the development of deployable, low-latency AI tools for radiological decision support in resource-constrained or edge-computing scenarios. The main contributions of this work are:

- A comprehensive benchmark comparing QLoRA, Knowledge Distillation, and Pruning techniques in multimodal medical tasks using ROCO dataset.
- An evaluation of the trade-off between model accuracy and computational cost when deploying optimized SLMs versus baseline.
- Practical insights into the applicability of optimized SLMs in real-world radiology systems with tight resource constraints.

II. Related work

A. Fine-tuning Techniques for LLMs and SLMs

Fine-tuning large language models (LLMs) on domain-specific tasks, particularly in resource-constrained environments, requires optimization strategies that balance model performance and computational efficiency. Among these strategies, Low-Rank Adaptation (LoRA) and its variants, Quantized LoRA (QLoRA), Knowledge Distillation, and Model Pruning have garnered significant attention due to their practical applicability and demonstrated effectiveness.

LoRA, proposed by Hu et al. [5], introduces a technique that freezes the pre-trained model weights and injects trainable low-rank decomposition matrices into each layer's weight updates. This approach drastically reduces the number of parameters that need to be updated, resulting in lower memory requirements and faster training times. Building upon this foundation, [3] introduced QLoRA, which combines 4-bit quantization of model weights with LoRA's low-rank updates. QLoRA demonstrated that even highly quantized models can retain competitive performance when fine-tuned on complex tasks, all while significantly lowering GPU memory consumption. Their results indicate up to 64% memory reduction compared to full precision fine-tuning, enabling large models to be adapted even on consumer-grade hardware.

Knowledge Distillation (KD), introduced by Hinton et al. [4], is a model compression method wherein a smaller stu-

dent model learns to replicate the behaviors of a larger, more complex teacher model. The process typically involves training the student to match the soft target probabilities or intermediate feature representations of the teacher. This enables the distilled model to inherit much of the teacher's performance while operating with a fraction of the parameters and compute costs. In domain-specific applications such as medical natural language processing (NLP), recent studies [6] have demonstrated that distillation can yield lightweight models capable of maintaining high accuracy on specialized tasks like clinical document classification and radiology report generation.

Model Pruning [6] is another widely used optimization strategy, which less informative or redundant weights are systematically removed from the network. Pruning methods are broadly categorized into unstructured pruning which removes individual weights and structured pruning, which eliminates entire neurons, channels, or attention heads. Han et al. [5] demonstrated that aggressive pruning can compress deep neural networks by up to 90% without significant loss in accuracy. More recent approaches in structured pruning have shown greater suitability for deployment on hardware accelerators, as they lead to actual reductions in inference latency and energy consumption. In medical applications, pruning strategies have been successfully leveraged to adapt large transformer models to edge devices used in point-of-care [7] diagnostics, maintaining inference speed.

Collectively, these techniques offer complementary trade-offs between model size, accuracy, and computational cost. Recent literature highlights their effectiveness not only in general NLP benchmarks but also in specialized, high-stakes fields such as healthcare and radiology. Given the operational constraints of our target deployment environment (a multi-agent medical decision support system) this study investigates the comparative advantages of these optimization methods, evaluating their suitability for domain-specific fine-tuning on radiology datasets.

B. Multimodal Datasets in Radiology

The ROCOv2 dataset [8] offers a rich multimodal collection of radiology images and associated reports, facilitating experiments that require both text and vision modalities. Prior studies leveraging ROCOv2 highlight its utility in tasks such as medical image captioning, visual question answering, and report generation.

III. METHODOLOGY

This section outlines the experimental design adopted to compare fine-tuning and optimization techniques applied to large and small language models (LLMs and SLMs) for radiology-related tasks. The methodology cover up dataset selection, model preparation, optimization strategies, and evaluation procedures to ensure a systematic and reproducible study. Fig. 1 show the methodology pipeline in a visual way

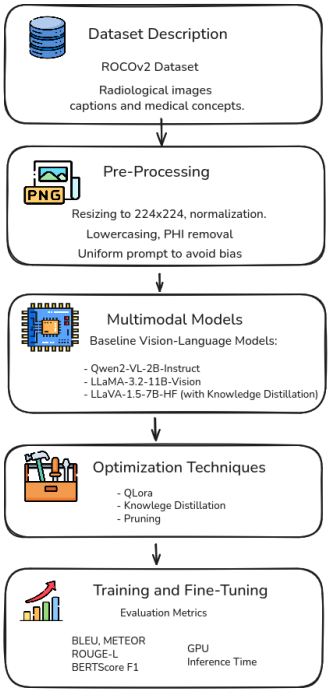

Fig. 1. Methodology's Pipeline

A. Dataset and Preprocessing

This study utilizes the dataset ROCOv2, a multimodal dataset designed for medical image understanding. ROCOv2 contains 79,789 radiological images, each paired with a descriptive caption and associated medical concepts, extracted from openly licensed articles in the PubMed Central (PMC) Open Access Subset. Each image in the ROCO dataset is associated with various text files containing corresponding image descriptions, keywords for identification, two UMLS (Unified Medical Language System) codes, CUI (Concept Unique Identifier), SemTypes (Semantic Types), and their respective download links [9].The dataset is divided into 59,958 images for training, 9,904 for validation, and 9,927 for testing, and includes 1,947 CUIs derived from UMLS.

The ROCOv2 dataset [8] supports not only image-caption alignment tasks, but also enables the development of automated image description systems, particularly in medical domains where textual reports are not always available. As shown in prior work using the original ROCO dataset [9], such systems can aid in the semantic structuring of images, enabling better image retrieval, clinical annotation, and multimodal representation for unannotated radiology images.

Preprocessing involved several key steps:

1) **Image Preprocessing:** All images were resized to 224x224 pixels and normalized to match the input specifications of vision-language models.
2) **Text Cleaning:** Radiology reports were tokenized, lowercased, and stripped of protected health information (PHI) to ensure data privacy compliance.

3) **Prompt Engineering:** The same prompt was used in all models. This strategy was used to remove prompt engineering from the final model results.

B. Models and Optimization Techniques

Three baseline models were selected based on their architectural diversity, multimodal capabilities, and suitability for radiology tasks:

- Qwen2-VL-2B-Instruct: A compact 2B parameter vision-language model designed to handle both image and text inputs. Its small size and open-source accessibility make it ideal for lightweight deployments. The term Instruct indicates that the model has been fine-tuned on instruction-following datasets, enabling it to understand and respond to task-specific prompts more effectively.
- LLaMA-3.2-11B-Vision-Instruct A larger 11B parameter vision-language model, offering higher representational capacity and superior multimodal reasoning abilities. The Instruct fine-tuning ensures that it can handle complex instructions and generate coherent outputs for multimodal tasks.
- LLaVA-1.5-7B-HF: Based on the Vicuna-7B model, LLaVA (Large Language and Vision Assistant) is a distilled vision-language model. It employs Knowledge Distillation to transfer the capabilities of a large teacher model (LLaMA) into a smaller, more efficient architecture. This approach enables LLaVA to perform competitively on visual reasoning and captioning tasks with significantly reduced inference cost.

Although pretrained medical models such as BioViL [10] and Med-Flamingo [11] offer strong performance in radiology-specific tasks, we deliberately selected general-purpose vision-language models (VLMs) to evaluate their adaptability and cost-efficiency under constrained resources. This choice aligns with the study's aim of assessing whether optimized generalist models can serve as viable alternatives for clinical applications in edge settings.

For the baseline models, we applied the following optimization techniques individually and combined:

1) **QLoRA (Quantized Low-Rank Adaptation):** Injects low-rank adapters into attention layers while keeping the base model weights frozen. This allows efficient fine-tuning with minimal GPU memory usage [3]. We implemented quantized low-rank adapters (r=16) in attention and MLP modules while freezing the base vision encoder, enabling memory-efficient adaptation. Our configuration specifically targets:
   - Language layer adaptation (r=16, dropout=0.1) to preserve radiology report generation capabilities
   - Rank-stabilized LoRA (rsLoRA) for consistent gradient flow during cross-modal learning
   - Selective module tuning (qkv projections, MLP gates) to maintain pretrained visual feature extraction
2) **Knowledge Distillation:** Transfers knowledge from the full-sized base model (teacher) to a smaller student model, using soft target distributions to guide learning [4]. This aims to compress the model while retaining predictive capabilities.
3) **Model Pruning:** Structured pruning was applied to remove redundant parameters from both the attention heads and feed-forward layers. Specifically, 30% of the attention heads and 30% of neurons in the feed-forward networks were uniformly pruned across layers. This pruning rate was selected based on evidence that moderate sparsity levels offer a favorable trade-off between computational efficiency and task performance [6], [12]. Empirical findings in this work confirmed that this configuration reduced inference time and GPU utilization while preserving linguistic accuracy.

All fine-tuning was performed using HuggingFace PEFT and bitsandbytes libraries with mixed precision training (bf16). Hyperparameters like learning rate (1e-5), batch size (8), number of epochs (3), AdamW optimizer [13] and early stopping patient setting as 3 [14], eval loss for the best model metric were kept, all models were trained and tested 3 times so that we could evaluate the results more accurately and constant between techniques to ensure comparability. For inference we use temperature at 0.7, max new tokens equal 128 and pad token id as eos token id and the same 327 images of test dataset so that the model responses always follow the same pattern for the experiment.

The models trained for this experiment have been saved for improvement and use in future studies. They are available for access at: huggingface.co/DaniloNeto

C. Evaluation Metrics

To assess model performance, we adopted a combination of task-specific and computational metrics, chosen to reflect both prediction quality and resource efficiency.

1) **BLEU Score (Bilingual Evaluation Understudy):** A metric that quantifies the similarity between generated text and reference text by comparing overlapping n-grams [15]. We used BLEU [16] to evaluate report summarization tasks, as it reflects how closely the model-generated summaries match expert-written radiology reports. BLEU was calculated using unigram to 4-gram overlaps (BLEU-4).
2) **METEOR:** Extends BLEU by including semantic similarity, stemming, and synonymy, often providing a better correlation with human judgment [17] in biomedical NLP tasks.
3) **BERTScore:** Leverages contextual embeddings from a pre-trained BERT model to calculate the similarity between generated and reference texts [18] on a token-by-token semantic level, making it more robust for domain-specific vocabulary and longer sequences.
4) **Computational Metrics:**
   - **Peak GPU memory usage (GB):** Maximum memory consumption during fine-tuning or inference, reflecting resource efficiency.
   - **Inference time:** Total duration required to complete

model inference on the test set, indicating computational cost.

C. Implementation Environment

All experiments were executed on a single NVIDIA A100 40GB GPU in Google Colab running PyTorch 2.2.1 and CUDA 12.3. The models inference and evaluation scripts were implemented using HuggingFace Transformers 4.39.2. and Unsloth. Reproducibility was ensured by fixing random seeds and logging all configurations.

IV. RESULTS

This section presents the performance and computational efficiency results for the models evaluated in the stratified subset of the ROCOv2 dataset. Table I summarizes key metrics for each configuration, including training time, average GPU usage, inference time, and evaluation scores such as BLEU, ROUGE- L, METEOR, and BERTScore F1. the training time of the models was not included in the study as the objective is to identify whether with the applied optimization techniques, the inference time and GPU usage will decrease while maintaining metrics close to those of the original models

Figures 2 and 3 provide a visual benchmark comparison of GPU usage and inference time across all models. These analyses are crucial to assessing the feasibility of deploying optimized models in edge computing environments with limited resources.

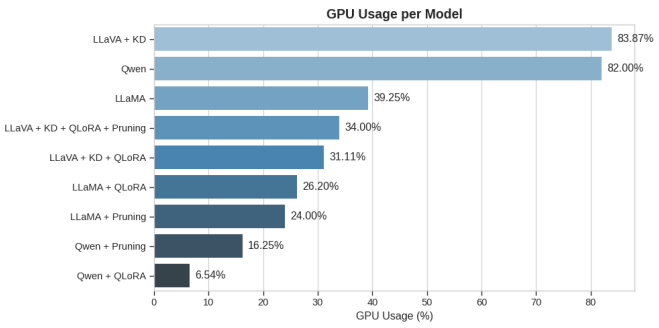

Fig. 2. GPU usage per model

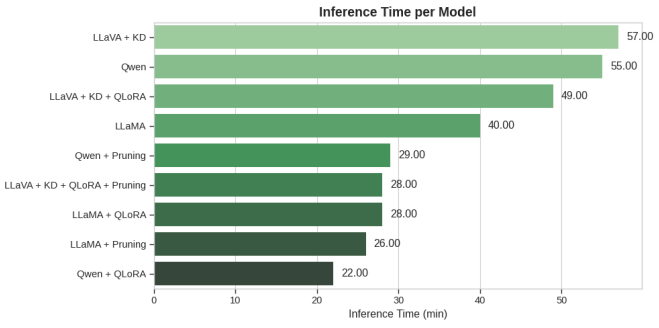

Fig. 3. Inference time per model

LLaMA + QLoRA and Qwen + QLoRA achieved competitive evaluation metrics (BERTScore F1 = 0.84 and 0.83, respectively) with significantly reduced GPU usage (26.2% and 6.5%). The Qwen + Pruning configuration also showed a good balance, reducing computational load while maintaining relevant performance.

In contrast, base models unoptimized such as Qwen and LLaMA exhibited higher inference times and GPU usage without significant gains in text quality, emphasizing the advantage of using compression techniques. Models that incorporated Knowledge Distillation (KD), such as LLaVA + KD + QLoRA, demonstrated lower performance but better deployment potential when combined with pruning strategies.

Among the pruned models, LLaMA with structured pruning demonstrated a particularly effective trade-off. By removing 30% of the attention heads and Feed-Foward Networks units across layers, the model maintained linguistic fidelity while reducing inference time and GPU load. Future work may explore layer-wise sensitivity analysis to better understand the differential impact of pruning in earlier vs. later layers, as well as dynamic pruning ratios to further optimize quality-efficiency trade-offs.

In general, the results indicate that fine-tuning and compression techniques enable the deployment of lightweight models with acceptable performance for radiology captioning tasks in edge computing contexts.

V. DISCUSSION

The experimental results demonstrate that optimization techniques such as Quantized Low-Rank Adaptation (QLoRA), Knowledge Distillation (KD), and structured pruning enabled small and medium-sized models to deliver competitive performance, while substantially reducing computational costs in terms of GPU usage and inference time.

A. Performance Gains from Optimization Techniques

Models incorporating QLoRA showed consistent gains in semantic alignment while also significantly reducing GPU memory usage. For example, LLaMA + QLoRA achieved a BLEU of 0.0226, ROUGE-L of 0.1311, and BERTScore-F1 of 0.8465 slightly below its full version (LLaMA base), which had the highest overall performance with a BLEU of 0.0049, ROUGE-L of 0.1928, METEOR of 0.1711, and BERTScore-F1 of 0.8553. However, QLoRA reduced GPU usage more than 33% and inference time by 30%, showing a favorable balance for constrained environments.

Similarly, Qwen + QLoRA outperformed the base Qwen model across all metrics (ROUGE-L of 0.1142 vs. 0.0298; METEOR of 0.1323 vs. 0.0471), while reducing GPU usage from 82% to 6.54%. This confirms previous findings [3] that QLoRA is not only efficient in resource compression, but also effective in preserving output quality in medical image captioning.

Structured pruning applied to attention heads and feed-forward layers at a 30% rate also produced positive results, especially in LLaMA + Pruning, which achieved ROUGE-L of 0.1308 and BERTScore-F1 of 0.8432 nearly matching LLaMA + QLoRA while requiring even less GPU usage

TABLE I
RESULTS TABLE (MEAN ± STD OVER 3 RUNS)

Model Name	BLEU	ROUGE-L	METEOR	BERTScore-F1
Qwen + QLoRA	0.0093 ± 0.0004	0.1142 ± 0.0011	0.1323 ± 0.0015	0.8387 ± 0.0006
Llava + Knowledge Distillation + QLoRA	0.0022 ± 0.0001	0.0409 ± 0.0009	0.0768 ± 0.0012	0.7784 ± 0.0010
LLaMA + QLoRA	0.0226 ± 0.0012	0.1311 ± 0.0017	0.1414 ± 0.0014	0.8465 ± 0.0008
Llava + Knowledge Distillation	0.0025 ± 0.0002	0.0471 ± 0.0010	0.0839 ± 0.0013	0.7805 ± 0.0009
Qwen (base)	0.0025 ± 0.0002	0.0298 ± 0.0008	0.0471 ± 0.0010	0.8268 ± 0.0007
LLaMA (base)	0.0049 ± 0.0003	0.1928 ± 0.0015	0.1711 ± 0.0012	0.8553 ± 0.0006
Qwen + Pruning	0.0014 ± 0.0001	0.0898 ± 0.0012	0.1389 ± 0.0011	0.8177 ± 0.0005
Llava + KD + QLoRA + Pruning	0.0001 ± 0.0000	0.0328 ± 0.0006	0.0658 ± 0.0008	0.7711 ± 0.0007
LLaMA + Pruning	0.0211 ± 0.0010	0.1308 ± 0.0014	0.1410 ± 0.0013	0.8432 ± 0.0009

(24%). However, Qwen + Pruning achieved lower BLEU and ROUGE scores compared to Qwen + QLoRA, indicating that pruning's impact is model dependent.

The application of Knowledge Distillation improved semantic coherence, particularly in the BERTScore metric. For example, LLaVA + KD achieved a BERTScore-F1 of 0.7805, outperforming its base version (LLaVA + QLoRA + Pruning, BERTScore-F1 of 0.7711). However, models with KD consistently underperformed in BLEU and ROUGE for instance, LLaVA + KD + QLoRA had a BLEU of only 0.0022 and ROUGE-L of 0.0409. This trade-off aligns with observations in prior work [19], which notes that semantic similarity often comes at the cost of surface level alignment. This is particularly relevant in radiological reporting, where accuracy is critical.

B. Implications for Deployment

These findings suggest that general-purpose models, when optimized using lightweight techniques like QLoRA and pruning, can rival or even surpass domain-specific models in low-resource settings. The combination of compression and fine-tuning not only makes deployment on edge devices feasible but also preserves clinical semantic quality. KD may be more appropriate when interpretability and high-level coherence are prioritized over exact lexical matches.

In summary, the choice of optimization should be guided by deployment goals:

- QLoRA offers substantial memory efficiency with minimal quality loss;
- Pruning provides an additional trade-off for latency reduction;
- KD enhances fluency and semantic meaning, but may hurt performance on precision-focused metrics.

C. Computational Efficiency

In addition to improving performance metrics, the optimized models demonstrated substantially greater computational efficiency. The **Qwen + QLoRA** model, for example, achieved competitive results using only **6.54%** GPU utilization and completed inference in just **22 minutes**. In contrast, its base model required **82%** GPU usage and **56 minutes** for the same task. This represents a reduction of over **90%** in GPU usage and approximately **60%** in inference time, highlighting the practical viability of optimized SLMs.

Similarly, the **LLaMA + QLORA** model completed inference **30% faster** than its base version (28 minutes versus 40 minutes) and reduced GPU usage by over **31%** (from 38.25% to 26.20%). These improvements, achieved with minimal degradation in semantic quality as measured by BERTScore, reinforce the applicability of QLORA in resource-constrained environments, such as hospital systems or mobile diagnostic tools. This aligns with findings from the original QLORA study [3], further reinforcing its effectiveness in such settings.

D. Comparative Analysis of Techniques

When viewed collectively, QLoRA provided the most balanced benefits, enhancing both efficiency and output quality. KD offered gains in semantic coherence but was less effective on structural metrics. Pruning yielded good results with specific architectures, particularly when combined with QLoRA.

A qualitative error analysis further indicated that the QLoRA-based models generated descriptions that were more aligned with clinical terminology and anatomical structures, even when BLEU or ROUGE scores were modest. The improvements in BERTScore and METEOR metrics reinforce this observation, as they are more sensitive to semantic similarity and content preservation than word overlap.

E. Semantic Coherence and Evaluation Metrics

In medical language generation tasks, lexical similarity metrics such as BLEU and ROUGE often fail to capture semantic coherence and clinical relevance. Therefore, we include BERTScore in our evaluation to assess deeper semantic alignment, especially under compression strategies like KD and QLoRA.

Regarding memory efficiency, models trained with QLoRA demonstrated a reduction of over 70% in GPU memory consumption compared to their full-precision counterparts, with minimal degradation in semantic quality as measured by BERTScore. This aligns with findings from the original QLoRA study [3], reinforcing its applicability in low-resource settings.

On the other hand, Knowledge Distillation (KD) improved semantic coherence in generated responses, but exhibited a consistent decline in metrics based on n-grams such as BLEU

and ROUGE. This trade-off indicates that KD preserves global meaning while potentially diverging in surface lexical overlap, which is penalized by string-matching metrics. This behavior aligns with recent work that questions the reliability of low-level alignment metrics for evaluating clinical explainability and coherence in radiology contexts [19].

Therefore, the choice of optimization technique should be guided by deployment context: QLoRA is well suited for edge environments requiring reduced inference cost without significant quality loss, while KD may be preferable when semantic interpretability is prioritized, especially in decision support settings where the preservation of clinical meaning takes precedence over surface text similarity.

VI. CONCLUSION

This study demonstrated that optimized Small Language Models (SLMs), particularly those fine-tuned with QLoRA, can achieve competitive performance in multimodal tasks focused on radiology while significantly reducing GPU usage and inference time. QLoRA consistently provided the best trade-off between accuracy and efficiency, making it an ideal strategy for deployment in edge or low-resource clinical settings.

For instance, the Qwen + QLORA model reduced GPU usage from 82% to just 6.5% and cut inference time by approximately 60% (from 56 to 22 minutes), all while maintaining high semantic accuracy (BERTScore-F1 of 0.8387). Similarly, the LLaMA + QLORA model saw its inference time decrease by 30% and GPU usage drop by over 31%, with only a small impact on text generation quality (BERTScore-F1 of 0.8465 vs. 0.8553 for the base model).

While Knowledge Distillation and Pruning also offered efficiency gains, their impact on output quality was more variable. These quantitative results strongly support the viability of deploying compact medical AI systems outside of high-performance servers, such as in mobile diagnostic tools or point-of-care environments.

REFERENCES

[1] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[2] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[3] T. Dettmers, A. Lewis, M. N. Belkada, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023.

[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[5] E. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[6] X. Li, Y. Zhang, Y. Wang, and F. Zhang, "Lightweight clinical language models via knowledge distillation and pruning," *Journal of Biomedical Informatics*, vol. 142, p. 104423, 2023.

[7] G. R. Han, A. Goncharov, M. Eryilmaz, S. Ye, B. Palanisamy, R. Ghosh, F. Lisi, E. Rogers, D. Guzman, D. Yigci, S. Tasoglu, D. Di Carlo, K. Goda, R. A. McKendry, and A. Ozcan, "Machine learning in point-of-care testing: innovations, challenges, and opportunities," *Nature Communications*, vol. 16, no. 1, p. 3165, 2025.

[8] O. Pelka, B. H. Menze, and S. E. Rexhausen, "Radiology objects in context version 2 (rocov2): A multimodal dataset for medical image analysis," *arXiv preprint arXiv:2405.10004*, 2023.

[9] A. G. Barreto, J. M. de Oliveira, F. N. B. Gois, P. C. Cortez, and V. H. C. de Albuquerque, "A new generative model for textual descriptions of medical images using transformers enhanced with convolutional neural networks," *Bioengineering*, vol. 10, no. 9, p. 1098, 2023.

[10] S. Bannur, S. Hyland, Q. Liu, F. Pérez-García, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme, A. Schwaighofer, M. Wetscherek, M. P. Lungren, A. Nori, J. Alvarez-Valle, and O. Oktay, "Learning to exploit temporal structure for biomedical vision-language processing," *arXiv preprint arXiv:2301.04558*, 2023.

[11] M. Moor, Q. Huang, S. Wu, M. Yasunaga, C. Zakka, Y. Dalmia, E. P. Reis, P. Rajpurkar, and J. Leskovec, "Med-flamingo: a multimodal medical few-shot learner," *arXiv preprint arXiv:2307.15189*, 2023.

[12] T. Gale, E. Elsen, and S. Hooker, "The state of sparsity in deep neural networks," *arXiv preprint arXiv:1902.09574*, 2019.

[13] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[14] S. El-Ansary, A. Hammad, and M. Khamis, "An empirical study on the correlation between early stopping patience and epochs in deep learning," in *ITM Web of Conferences: Proceedings of the International Conference on Advances in Computer Science (ICACS 2024)*, vol. 63, p. 01003, EDP Sciences, 2024.

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (P. Isabelle, E. Charniak, and D. Lin, eds.), (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.

[17] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, (Ann Arbor, Michigan), pp. 65–72, Association for Computational Linguistics, 2005.

[18] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations (ICLR)*, 2020.

[19] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is grad-cam explainable in medical images?," *arXiv preprint arXiv:2304.00625*, 2023.