This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

From Manuals to Maintenance: Fine-Tuning MedGemma for Multi-Modal Imaging System Support in Low-Resource Settings

Bernes Lorier Atabonfack[1,2], Zion Kongbi Nfo[2,*], Ahmed Tahiru Issah[1,2,*], Tolulope Olusuyi[2,3,*], Clemence Ingabire[1,2], Mohammed Hardi Abdul Baaki[1,2], Mawuli Deku[1,2], Abdulrazaq Zubair[4], Alyasaa Anas[5,], Confidence Raymond[2,3,7], Maruf Adewole[2,3,6], and Udunna C. Anazodo[3,7]

[1] Carnegie Mellon University Africa, Kigali, Rwanda
[2] IngenziAI, Kigali, Rwanda
[3] Medical Artificial Intelligence Laboratory, Lagos, Nigeria
[4] Federal University of Health Sciences Azare, Nigeria.
[5] University of Maiduguri Teaching Hospital, Borno State, Nigeria
[6] University of Pennsylvania
[7] McGill University, Montréal, Canada
batabonf@alumni.cmu.edu, kongbizion3@gmail.com, aissah@alumni.cmu.edu, tolusuyi@mailab.io, ingabireclec@gmail.com, abdulbaakimohammedhardi@gmail.com, kofi.mawulideku@gmail.com, abdulrazaq.zubair@nkdc.ng, alyasaa390@gmail.com, raymondconfidence@gmail.com, madewole@mailab.io, udunna.anazodo@mcgill.ca

Abstract. Imaging device downtime is a major barrier to healthcare delivery in low- and middle-income countries (LMICs), often driven by limited access to specialized biomedical engineering support. We present a multi-modality medical equipment maintenance question-answering (QA) framework and demonstrate the fine-tuning of a medical foundation model for specialized technical troubleshooting tasks. Guided by a multi-country survey across nine LMICs, we curated technical manuals from MRI and ultrasound systems to generate the INGENZI_DatasetV1, containing 10,294 high-quality, filtered QA-context pairs. Using QLoRA-based parameter-efficient fine-tuning, we adapted the MedGemma-4b-it model to interpret system error logs and generate step-by-step equipment repair instructions. Compared to the baseline model, the fine-tuned system achieved substantial improvements across metrics, including F1 score (0.22 to 0.38), ROUGE-2 (0.18 to 0.41), and BERTScore F1 (0.86 to 0.91). These metric gains demonstrate that the model generates significantly more precise and procedurally accurate technical responses to new troubleshooting queries. This work establishes a reliable foundation for AI-assisted diagnostic and maintenance tools in resource-constrained settings.

Keywords: Medical Foundation Models · MRI Maintenance · Biomedical Engineering · Parameter-Efficient Fine-Tuning · QLoRA · Medical

* Co-second authors

Corresponding author: Udunna C. Anazodo, udunna.anazodo@mcgill.ca

Question Answering · Imaging Equipment Sustainability · Low-Resource Settings

1 Introduction

Medical imaging technologies such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and X-ray systems are essential to modern healthcare, enabling diagnosis, treatment planning, and disease monitoring across a wide range of conditions. Global demand for imaging continues to grow [1] [2], yet access and reliability remain highly uneven, particularly in low- and middle-income countries (LMICs), where equipment downtime significantly limits its diagnostic capacity [2][3]. In resource-constrained settings, prolonged downtime is often driven by limited availability of trained biomedical engineers, restricted access to service documentation, costly manufacturer-dependent servicing models, and infrastructure challenges such as unstable power and connectivity [5]. These factors reduce system availability and negatively impact patient care.

Recent advances in large language models (LLMs) and medical foundation models demonstrate strong clinical reasoning [6][7] [8]. However, their potential to assist biomedical technicians with troubleshooting and maintenance remains largely unexplored. Prior work introduced INGENZI Tech, an AI-powered diagnostic support for medical device troubleshooting, using general-purpose LLMs [5]. This initial framework lacked domain-specific device manual datasets for training and benchmarking AI models, limiting real-world technical precision. To address this, we present an improved framework for adapting medical foundation models to support imaging equipment. Guided by a multi-country survey across nine LMICs, we characterized local maintenance challenges and use this information to curate and publicly release the INGENZI_DatasetV1 [4], as a corpus of 10,294 high-quality question-answer pairs derived from Siemens MAGNETOM MRI and Philips HDI 5000 ultrasound operator and service manuals. Using this dataset, we fine-tuned MedGemma-4B[6] via Quantized Low-Rank Adaptation (QLoRA) [9], demonstrating substantial improvements in answering operational and maintenance-related queries.

Our contributions are threefold: (1) a multi-country survey analyzing imaging equipment maintenance challenges in LMICs; (2) the public release of the INGENZI_DatasetV1, a multi-modality technical question-answer (QA) dataset designed for equipment support modeling; and (3) parameter-efficient fine-tuning of a medical foundation model for biomedical imaging equipment troubleshooting.

2 Related Work

2.1 Medical Foundation Models

Medical foundation models are large language models adapted to biomedical and clinical domains through domain-specific pretraining and instruction tun-

ing. Prior work has shown that such models can achieve strong performance on medical question answering and clinical reasoning tasks. For instance, Med-PaLM [7] demonstrated near-expert performance on standardized medical examinations when appropriately fine-tuned and aligned [10]. MedGemma follows this paradigm by providing instruction-tuned medical language models optimized for healthcare applications. While these models incorporate biomedical terminology and domain knowledge, their evaluation has largely focused on clinical reasoning benchmarks. Their use for operational support of medical equipment, such as interpreting system errors or guiding maintenance procedures, remains underexplored. In this work, we adapt MedGemma-4B to medical imaging system operation and maintenance using a curated dataset derived from manufacturer documentation.

2.2 Parameter-Efficient Fine-Tuning of Large Models

Full fine-tuning of large language models is computationally expensive. Parameter-efficient fine-tuning (PEFT) [11] mitigates this by training a small set of additional parameters while freezing the base model weights. Low-Rank Adaptation (LoRA) [12] introduces trainable low-rank matrices into attention layers, significantly reducing the number of updated parameters. Quantized LoRA (QLoRA) [9] further enhances efficiency by combining 4-bit weight quantization with LoRA adapters, enabling multi-billion-parameter models to be fine-tuned on modest hardware without substantial performance degradation [9]. We employ QLoRA to adapt MedGemma-4B, a vision-language model pretrained on both text and image data, to the INGENZI_DatasetV1. In this work, we fine-tune and use only its text-based capabilities, enabling efficient domain specialization while preserving the general medical reasoning capabilities of the base model.

3 Method

3.1 Multi-Country Survey of Imaging Equipment Maintenance Practices

To ground the development of our device operation dataset and model real-world operational challenges, we conducted a cross-sectional survey targeting biomedical engineers, radiographers, and imaging technicians working in LMICs. The survey characterized device availability, breakdown patterns, maintenance practices, access to documentation, and digital readiness across diverse healthcare settings. The survey was administered using Google Forms and structured into five sections covering facility characteristics, device inventory, maintenance practices, fault management, and digital readiness. Participation was voluntary. Responses were anonymized unless participants explicitly opted into follow-up testing. Consent to participate and the privacy statement indicated that responses would remain confidential and would be used solely for platform development and improvement purposes, and participants could withdraw at any

time without penalty. A total of 61 facilities across nine countries (Nigeria, Rwanda, Ghana, Ethiopia, Tanzania, Uganda, Cameroon, Kenya, and Nepal) participated. The survey findings (see Results section) directly informed our dataset scope and modeling pipeline. Common reports of high breakdown rates, prolonged downtime, limited access to documentation, and fragmented logging practices highlight the need for structured technical knowledge covering multiple imaging modalities, accessible through natural language interfaces. Based on these findings, we curated device operator and service manuals for MRI and ultrasound systems into the multi-modality INGENZI_DatasetV1. We used it for parameter-efficient fine-tuning of a medical foundation model (MedGemma-4B) for imaging equipment maintenance tasks.

3.2 INGENZI Dataset Curation

3.2.1 Data Sources and Scope To construct a domain-specific dataset for imaging equipment support, we curated technical documentation from two imaging modalities: Siemens MAGNETOM MRI systems and Philips HDI 5000 ultrasound systems. All technical manuals were retrieved from open-access online repositories and public engineering platforms. The MRI corpus included 19 operator, service, and maintenance manuals totaling 2,077 pages, covering multiple MAGNETOM models and software versions. The ultrasound corpus included 10 documents totaling 2,322 pages, spanning service manuals, software instructions for use, technical specifications, and cleaning/disinfection guidance. All source PDFs were used strictly for dataset construction and were not redistributed. The released INGENZI_DatasetV1 contains only structured question–answer (QA) pairs extracted from documents, with associated textual context, and excludes proprietary images, diagrams, or raw manuals.

3.2.2 Data processing We implemented a unified pipeline using LlamaIndex to convert unstructured technical documentation into synthetic QA pairs for supervised fine-tuning. Documents were ingested and segmented into overlapping text chunks (512 characters with a 64-character overlap) to promote the generation of grounded, specific questions. Using RagDatasetGenerator, we generated question, answer, and reference context triples. Generation temperature was set to 0.0 to reduce hallucinations, and retry mechanisms were used for robust large-scale processing. The generated samples underwent manual and rule-based filtering to remove entries with missing fields, insufficient context grounding, structural formatting errors, generic responses (e.g., "refer to the manual"), and likely hallucinated answers. For MRI documentation, 8,222 samples were initially generated, of which 5,667 high-quality QA pairs were retained. For Philips HDI 5000 ultrasound documentation, 5,236 samples were generated, with 609 removed during filtering, leaving 4,627 QA pairs. An overview of the data curation and processing is summarized in Fig 1

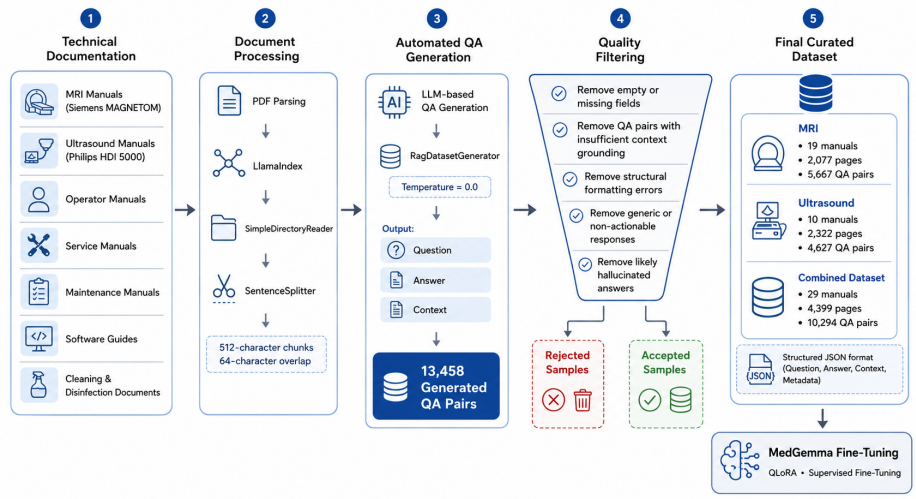

Fig. 1. Workflow for constructing the medical imaging equipment QA dataset from technical manuals.

3.3 Model Fine-Tuning and Experimental Setup

Using the INGENZI dataset, we fine-tuned the MedGemma-4b-it model for technical question answering using parameter-efficient fine-tuning. The task formulation was causal language modeling, where each training example was formatted into an instruction-style prompt containing the question and target answer. The INGENZI dataset was split into training, validation, and test partitions containing 8,235, 1,029, and 1,030 samples, respectively, with a random seed of 42. Splitting was performed at the QA-pair level.

Fine-Tuning Method. We employed Quantized Low-Rank Adaptation (QLoRA) to enable efficient fine-tuning under limited computational resources. The base model was loaded in 4-bit precision using BitsAndBytes quantization, and LoRA adapters were applied to all linear layers. The LoRA configuration used rank $r = 16$, scaling factor $\alpha = 16$, and dropout 0.05. Bias parameters were not adapted. Supervised fine-tuning (SFT) was performed using the Transformers Reinforcement Learning (TRL) library's SFTTrainer.

Training Configuration. Training was conducted on Google Colab using NVIDIA A100 Tensor Core GPUs. The model was trained for 15 epochs with a learning rate of 2×10^{-4} using the fused AdamW optimizer. Batch size was 4 per device with gradient accumulation steps of 4, resulting in an effective batch size of 16. Gradient checkpointing was enabled to reduce memory usage. A linear learning rate scheduler with a warmup ratio of 0.03 was applied. The maximum gradient norm was clipped at 0.3. Models were evaluated every 50 steps, and checkpoints were saved at each epoch.

Evaluation Metrics. Model performance was evaluated using token-level F1 score, ROUGE-1[14], ROUGE-2, ROUGE-L, and BERTScore F1 [13]. The baseline corresponds to the original MedGemma-4b-it model without domain-specific fine-tuning.

4 Results

Survey Facility Characteristics and Device Distribution: Among participating facilities, public hospitals constituted the largest group (30/61), followed by private hospitals (18), diagnostic centers (8), research and/or academic centers (4), and NGO/faith-based institutions (1) Fig 2. Across facilities, ultrasound and X-ray systems were the most widely available modalities, whereas SPECT systems were relatively rare. Breakdown frequency was high across commonly used modalities. In total, 63 devices were reported as non-functional across the 61 facilities, corresponding to an average of more than one broken imaging device per facility Fig 3.

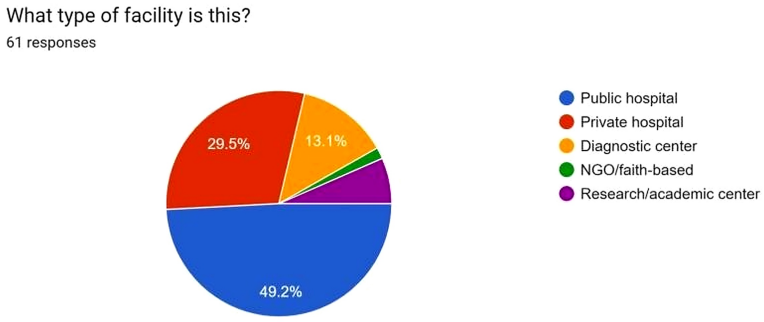

Fig. 2. Survey Responses by healthcare facility type (n = 61).

In terms of documentation, access to technical documentation was limited, with 41% facilities relying exclusively on hard-copy manuals, 11% on digital manuals, while 33% reported no manuals available, and 15% had both hard-copy and digital formats. Despite these limitations, 72% (44/61) expressed willingness to share technical documentation for knowledge-building purposes. Error logging practices were inconsistent, with most facilities reporting either no formal logging system or reliance on paper-based logs, while only a minority used specialized digital systems. Open-ended responses revealed recurring themes, summarized as: 1) software-related failures (calibration errors, system freezes, PACS connectivity issues); 2) electrical instability and power fluctuations; 3) hardware failures (coil failures, tube rotor errors); 4) human resource shortages; financial

constraints and spare part limitations; 5) maintenance culture deficiencies; and 6) operational overload. These findings indicate partial digital readiness, suggesting that AI-based support systems must accommodate heterogeneous connectivity environments.

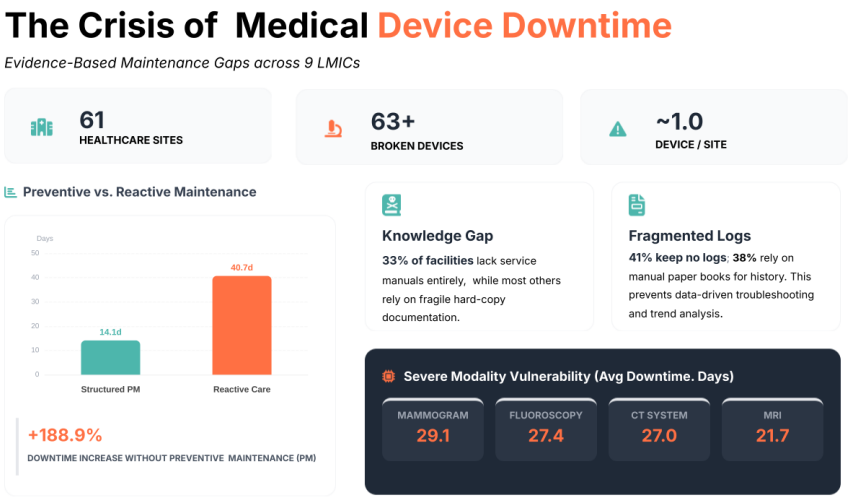

Fig. 3. Summary of survey findings across 61 healthcare facilities in nine LMICs. A total of 63 non-functional imaging devices were reported, corresponding to approximately one broken device per affected site. Facilities with structured preventive maintenance exhibited substantially lower mean downtime (14.1 days) compared to reactive maintenance settings (40.7 days), representing a 188.9% increase in downtime without structured maintenance. Mammography, fluoroscopy, CT, and MRI systems demonstrated the longest average downtime durations. Documentation gaps and fragmented logging practices were common across facilities.

The INGENZI Dataset V1: The combined corpus contains 10,294 QA pairs spanning MRI and ultrasound equipment maintenance, operation, safety, and troubleshooting tasks. Each entry contains an identifier, question, answer, supporting context, and generation metadata Table 1. No raw manuals or proprietary assets are redistributed. To our knowledge, this is among the first openly available datasets derived from medical imaging equipment documentation specifically designed for training and evaluating language models for technical maintenance support. The dataset is publicly available and can be accessed here [4]

Fine-Tuned Foundational Model: Table 2 presents the performance comparison between the base MedGemma model and the fine-tuned model on INGENZI_DatasetV1. Fine-tuning improved all lexical and semantic metrics, with ROUGE-2 increasing from 0.18 to 0.42, token-level F1 from 0.22 to 0.38, and

Table 1. Structure of each entry in the *INGENZI_ DatasetV1*

Field	Description
`id`	Unique identifier for the QA pair
`question`	Natural language technical question
`answer`	Grounded reference answer derived from context
`context`	Source text chunk from MRI documentation
`num_contexts`	Number of context segments used (typically 1)
`query_by`	Source of question generation (AI-generated)
`answer_by`	Source of answer generation (AI-generated)

BERTScore F1 from 0.86 to 0.92. These results indicate strong agreement with the filtered, model-generated reference answers.

Table 2. Baseline vs. fine-tuned performance on the INGENZI_DatasetV1.

Metric	Baseline	Fine-Tuned
F1 Score	0.22	**0.38**
ROUGE-1	0.37	**0.57**
ROUGE-2	0.18	**0.41**
ROUGE-L	0.26	**0.48**
BERTScore F1	0.86	**0.91**

5 Discussion

This work demonstrates that domain-specific fine-tuning substantially improves medical foundation models for multimodal imaging equipment maintenance support. While the base MedGemma model exhibited moderate semantic alignment, it struggled with technical precision. Fine-tuning on the INGENZI dataset yielded major gains across all metrics, notably doubling ROUGE-2 (0.18 to 0.41), suggesting the model's enhanced capacity to reproduce structured technical instructions. In general, the results show that a high-quality, targeted domain dataset such as INGENZI_Dataset_V1 can effectively specialize medical language models for complex operational tasks. This is critical for low-resource settings where original equipment manufacturer (OEM) support is limited. A lightweight model could serve as a vital decision-support tool for frontline technicians. Crucially, these improvements were achieved without retrieval augmentation (RAG), establishing a robust baseline for future hybrid systems.

6 Limitations

Several limitations remain. First, although INGENZI extends prior work[5] with Siemens MRI and Philips ultrasound documentation, validation across other

modalities (e.g., CT, X-ray) and OEMs is needed. Random QA-level splitting may also place overlapping or near-duplicate contexts across partitions, potentially inflating performance. Second, evaluation was limited to the MedGemma baseline; comparisons with other open-source and commercial LLMs are needed. Third, the absence of live RAG introduces a risk of hallucinated or outdated instructions. Finally, evaluation relied solely on automated metrics, making expert validation by biomedical engineers essential for real-world deployment. Our ongoing work combines the fine-tuned model with a RAG pipeline through a hybrid weighted ensemble to improve factual grounding. Future work will expand modalities, adopt document-level splits, integrate live retrieval, and conduct field safety evaluations.

Acknowledgements

We thank all medical imaging professionals who participated in the survey for their valuable contributions and insights.'This work was partially supported by the European Union – Next Generation EU, Mission 4 Component 2 Line 1.3, through the PNRR MUR project PE0000013 – FAIR "Future Artificial Intelligence Research" (CUP E63C22001940006), through the AFRICAI Mentorship Fellowship. Part of the computing resources used in this work were made possible through the NeurON (The Neuro Open Science& Global Health Platform), partly funded by The Digital Research Alliance Canada (Compute Canada) Research Platforms and Portals Program.

Disclosure of Interests

The authors declare no competing interests relevant to the content of this article.

References

1. "Medical Equipment Maintenance Market," Market.us, Jan. 10, 2025. https://market.us/report/medical-equipment-maintenance-market/ (accessed Feb. 26, 2026).
2. WHO, "Global Atlas of medical devices," www.who.int, Aug. 05, 2017. https://www.who.int/publications/i/item/9789241512312
3. Marks IH, Thomas H, Bakhet M, Fitzgerald E. Medical equipment donation in low-resource settings: a review of the literature and guidelines for surgery and anaesthesia in low-income and middle-income countries. BMJ Glob Health. 2019 Sep 29;4(5):e001785. doi: 10.1136/bmjgh-2019-001785. PMID: 31637029; PMCID: PMC6768372.
4. Atabonfack, B.L., Issah, A.T., Baaki, M.H.A., Ingabire, C., Olusuyi, T., Kongbi, Z., Adewole, M., Anazodo, U.C.: Imaging Equipment Technical QA Dataset. Hugging Face (2026). https://huggingface.co/datasets/BernesLy/INGENZI_DatasetV1

5. Atabonfack, B. L., Issah, A. T., Baaki, M. H., Ingabire, C., Olusuyi, T., Adewole, M., Anazodo, U. C., & Brown, T. X. (2026). Empowering Medical Equipment Sustainability in Low-Resource Settings: An AI-Powered Diagnostic and Support Platform for Biomedical Technicians. Nature, 217-230, 978-3-032-13654. "https://link.springer.com/chapter/10.1007/978-3-032-13654-1_22.
6. Sellergren et al. "MedGemma Technical Report." arXiv preprint arXiv:2507.05201 (2025).
7. Singhal et al.,Med-PaLM, Nature, 2023
8. LLaVA-Med Li et al., NeurIPS Workshop, 2023
9. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv. https://arxiv.org/abs/2305.14314
10. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., . . . Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180. https://doi.org/10.1038/s41586-023-06291-2
11. Balne, C. C., Bhaduri, S., Roy, T., Jain, V., & Chadha, A. (2024). Parameter Efficient Fine Tuning: A Comprehensive Analysis Across Applications. ArXiv. https://arxiv.org/abs/2404.13506
12. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. ArXiv. https://arxiv.org/abs/2106.09685
13. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. arXiv preprint arXiv:1904.09675. https://doi.org/10.48550/arxiv.1904.09675
14. Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics. https://aclanthology.org/W04-1013/