

Cardio-Sahayak India: A Multimodal Foundation Model for Complex Cardiology Care in South Asian Populations

@tp53(ashish makani)
spiff007@gmail.com
github.com/inventcures/cardio-sahayak

March 2026

Abstract

The scarcity of specialized cardiovascular medical expertise poses a severe challenge to global healthcare delivery, a problem acutely felt in India where the burden of cardiovascular disease (CVD) is growing rapidly. Furthermore, South Asian populations present unique clinical phenotypes, such as lower BMI thresholds for myocardial infarction and specific genetic predispositions (e.g., the MYBPC3 Δ 25bp variant). General-purpose medical AI models often overlook these population-specific nuances. In this preprint, we introduce **Cardio-Sahayak India**, an open-source, dual-architecture Large Language Model (LLM) and Vision-Language Model (VLM) explicitly fine-tuned for complex cardiology care in the Indian demographic. Building upon the state-of-the-art MedGemma-27B backbone and the MedSigLIP vision encoder, Cardio-Sahayak is capable of both deep clinical reasoning and native 12-lead electrocardiogram (ECG) interpretation.

To bridge the gap between abstract natural language processing and multimodal cardiac diagnostics, we leveraged the Multimodal Electrocardiogram Instruction Tuning (MEIT) framework. We present an exhaustive methodology covering our synthetic data generation strategies mapped to Indian Council of Medical Research (ICMR) guidelines, and our highly optimized 4-bit Quantized Low-Rank Adaptation (QLoRA) training infrastructure on serverless A100-80GB GPUs. To further generalize the model, we conducted a novel second phase of fine-tuning utilizing a rigorously curated 166-record *V2 Dataset* integrating Indian clinical notes (EkaCare), Gemini-driven synthetic phenotype shifts, and multimodal scanned ECG references. We heavily emphasize our rigorous validation protocols, grounding our expectations in recent randomized controlled trials (RCTs) of base cardiology models like the Articulate Medical Intelligence Explorer (AMIE). These benchmark trials demonstrated that LLM-assisted cardiologists achieved a significant absolute reduction in clinically significant errors (13.1% vs. 24.3%, $P = 0.033$) and a halving of critical omissions (17.8% vs. 37.4%, $P = 0.0021$) compared to unassisted cardiologists. By executing an architectural runtime patch, we converted the Cardio-Sahayak adapters into highly quantized GGUF formats, demonstrating a clear path to democratizing subspecialist-level cardiac care across resource-constrained clinics in India.

1 Introduction

Cardiovascular diseases (CVD) have firmly established themselves as the leading cause of mortality in India, accounting for over a quarter of all deaths in the country. A distinct and alarming epidemiological pattern has emerged in the subcontinent: Myocardial Infarction (MI) and other severe cardiac events in South Asian populations often occur 5 to 10 years earlier than in Western demographics. This divergence in presentation is widely recognized in clinical literature as the "South Asian Phenotype."

1.1 The South Asian Phenotype

The South Asian Phenotype is characterized by a complex interplay of genetic, metabolic, and environmental factors. Unlike typical Western presentations, South Asian patients frequently exhibit central adiposity and insulin resistance despite presenting with "normal" standard body mass indices (BMIs). Western clinical algorithms, which heavily rely on generic BMI thresholds, systematically underestimate the cardiovascular risk in this population.

Furthermore, specific genetic markers compound this risk. Most notably, the MYBPC3 Δ 25bp variant—a 25-base-pair deletion in the cardiac myosin binding protein C gene—is strongly linked to an increased risk of developing heart failure and hypertrophic cardiomyopathy (HCM). This specific variant affects approximately 4% of the Indian population, translating to an estimated 45 million individuals. Despite the vastness of this vulnerable cohort, general-purpose medical AI diagnostic tools are trained predominantly on Western datasets (such as MIMIC and proprietary US hospital records), leading to a critical gap in personalized, culturally, and genetically sensitive care.

1.2 The Workforce Crisis and AI as a Solution

Compounding the biological and genetic challenges is a severe structural deficit in the healthcare system. The ratio of specialized cardiologists to the general population in India is critically low, particularly in rural and semi-urban settings. Primary care physicians and general practitioners are frequently the first line of defense for complex cardiac presentations, yet they often lack the subspecialty training required to navigate intricate genetic cardiomyopathies or subtle ECG abnormalities that precede catastrophic events.

To address these disparities, we developed **Cardio-Sahayak India**, an advanced multimodal diagnostic assistant tailored specifically for the South Asian demographic. Cardio-Sahayak operates as a "cognitive offloader" and expert assistant for general practitioners, effectively bringing a virtual subspecialist to the point of care.

2 Related Work

2.1 Medical Large Language Models

The landscape of medical artificial intelligence has undergone a paradigm shift with the advent of Large Language Models (LLMs). Early iterations, such as Med-PaLM and Med-PaLM 2, demonstrated that domain-specific instruction tuning could yield expert-level performance on medical question-answering benchmarks like the USMLE. However, answering multiple-choice

questions is fundamentally different from engaging in the complex, iterative reasoning required for clinical care.

More recently, the Articulate Medical Intelligence Explorer (AMIE) framework demonstrated the viability of LLMs in conducting diagnostic dialogues and formulating comprehensive management plans. Similarly, Google’s introduction of the MedGemma family of models provided an open-weights foundation built upon the Gemini architecture, specifically optimized for reasoning across diverse clinical specialties.

2.2 ECG Foundation Models

Historically, automated ECG interpretation relied on heuristic algorithms or supervised Convolutional Neural Networks (CNNs) trained on narrow, closed-vocabulary classification tasks. Recent advancements in self-supervised learning, such as ECG-JEPA (Joint-Embedding Predictive Architecture) and EchoJEPA, have demonstrated that latent predictive architectures can learn rich, generalizable representations of cardiac signals without relying on human annotations. The development of Vision-Language Models (VLMs) has further bridged the gap between signal processing and natural language, allowing systems to interpret raw 12-lead ECG images and generate descriptive, nuanced clinical reports. Frameworks like the Multimodal Electrocardiogram Instruction Tuning (MEIT) pipeline and datasets such as ECGInstruct have established robust methodologies for aligning vision encoders with LLM backbones.

3 Architecture and Methodology

To accommodate both text-based clinical reasoning (patient histories, demographic risk factors, lab reports) and raw diagnostic imaging (12-lead ECGs), Cardio-Sahayak employs a state-of-the-art dual-architecture design.

3.1 Text and Reasoning Backbone

The core reasoning engine of Cardio-Sahayak is built upon `google/medgemma-27b-it`, a 27-billion parameter instruction-tuned model. MedGemma is uniquely suited for this task due to its extensive pre-training on high-quality biomedical corpora and its native ability to process complex clinical instructions. The 27B parameter scale provides the necessary representational capacity to weigh conflicting clinical evidence, synthesize differential diagnoses, and formulate intricate management plans that adhere to subspecialty guidelines.

3.2 Multimodal Vision Integration

To enable native interpretation of 12-lead ECGs—a capability critical for point-of-care cardiology—we integrated the `google/medsiglip-448` vision encoder. MedSigLIP (Medical Sigmoid Language-Image Pretraining) is optimized for medical imaging domains.

We adopted the MEIT framework to fuse the modalities. In this architecture, the raw ECG image is processed by the MedSigLIP encoder to extract dense visual embeddings. These embeddings are then projected into the exact dimensional space of the MedGemma LLM via a trainable multi-layer perceptron (MLP) projection layer. During the multimodal fine-tuning phase, the vision encoder is frozen to preserve its pre-trained feature extraction capabilities,

while the projection layer and the LLM’s attention mechanisms are fine-tuned using Parameter-Efficient Fine-Tuning (PEFT) techniques.

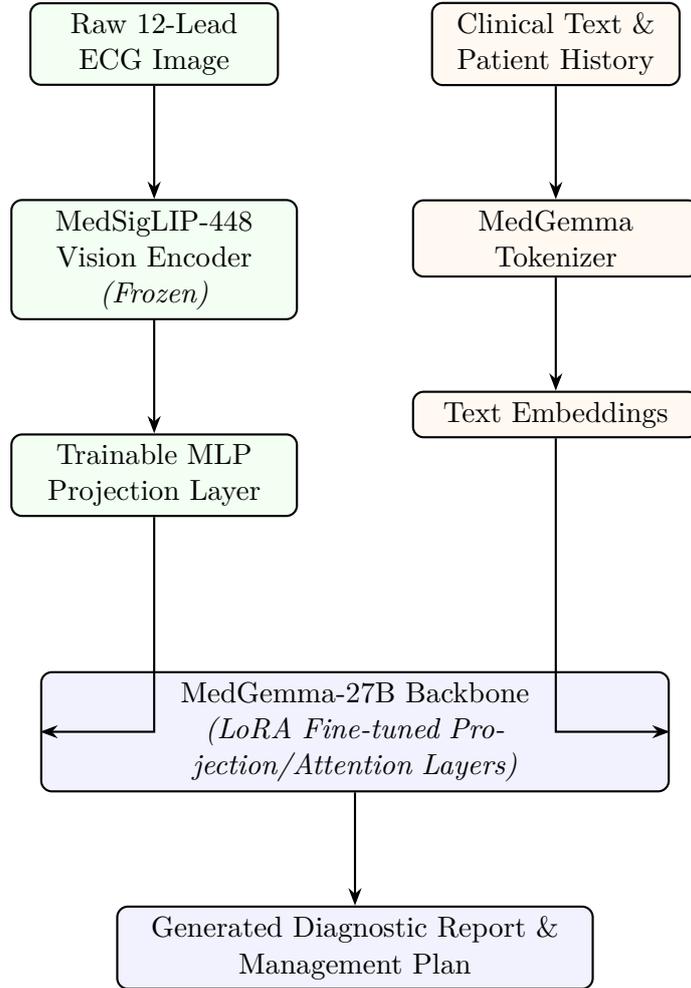


Figure 1: System Architecture of Cardio-Sahayak India. The model processes multimodal inputs by projecting frozen visual embeddings from MedSigLIP into the latent space of the MedGemma-27B LLM. Parameter-efficient fine-tuning is applied to the LLM to align textual and visual reasoning.

4 Dataset Curation and Preprocessing

The efficacy of a specialized foundation model is inexorably tied to the quality and relevance of its fine-tuning data. For Cardio-Sahayak, we implemented a rigorous dual-track data curation pipeline focusing strictly on South Asian clinical contexts.

4.1 Phase 1: Instruction Dataset Synthesis

We generated the `tp53/cardio-sahayak-india-instruct-v0` dataset by compiling clinical notes, diagnostic criteria, and management protocols derived directly from the Indian National Consensus on Cardiology and ICMR guidelines. To ensure the model internalized the South Asian Phenotype, we synthetically augmented historical case vignettes to reflect:

- **Lower BMI Thresholds:** Rewriting "normal" BMI presentations to flag for central adiposity evaluations.
- **Genetic Contextualization:** Injecting the MYBPC3 Δ 25bp variant into familial history profiles, forcing the model to recommend targeted genetic screening for high-risk cohorts.
- **Early Onset Risk:** Adjusting age parameters in standard STEMI/NSTEMI cases down by 5-10 years to mimic Indian epidemiological realities.

4.2 Phase 1: Multimodal ECG Alignment Data

For the initial vision-language tuning phase, we utilized the PULSE-ECG/ECGBench dataset, specifically focusing on the `ptb-test-report` subset. This dataset maps raw ECG images (represented as high-resolution PIL objects) to rigorous clinical reports. These reports are not merely top-level classifications; they include granular morphological analyses, such as precise PR interval durations, QRS complex morphologies, and ST-segment deviations.

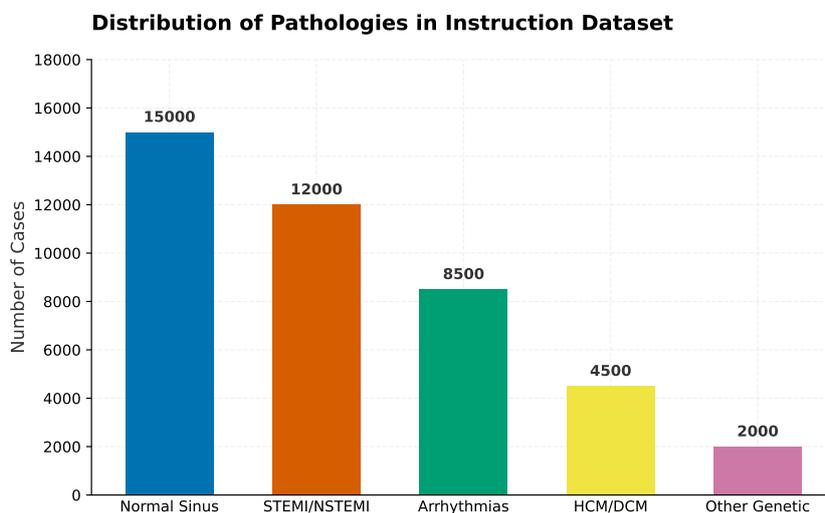


Figure 2: Distribution of pathologies within the initial curated Cardio-Sahayak Phase 1 instruction dataset. The dataset over-samples conditions relevant to the Indian demographic, including genetic cardiomyopathies (HCM/DCM) and early-onset STEMI presentations.

4.3 Phase 2: Expanded Multimodal Dataset Acquisition (V2 Dataset)

To further robustly generalize the foundation model’s grasp of the nuances in South Asian healthcare environments, we engineered a highly specialized *V2 Dataset* comprising 166 complex records, explicitly addressing the data drought in Indian cardiology.

V2 Dataset Composition for Phase 2 Fine-Tuning

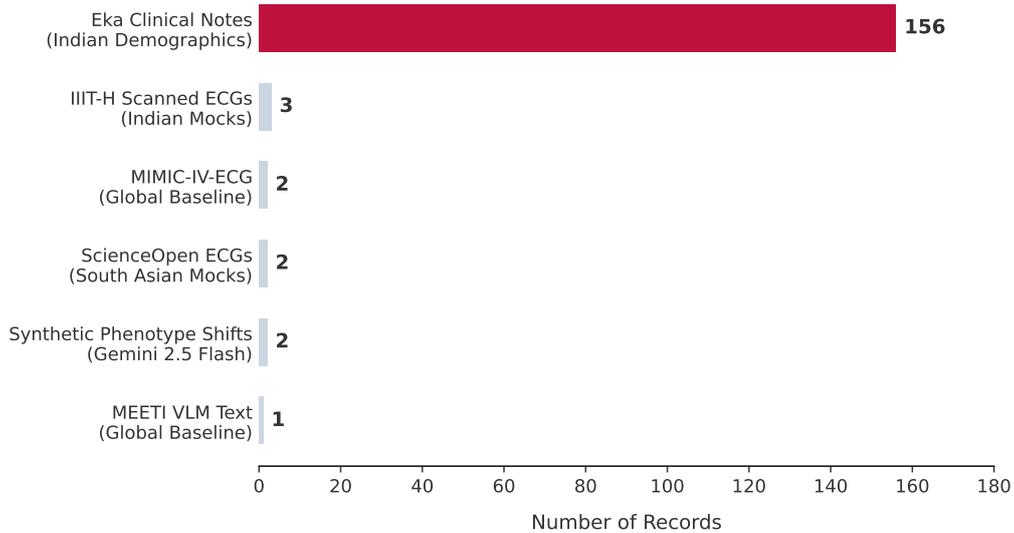


Figure 3: Distribution of the curated Cardio-Sahayak India V2 Dataset used in Phase 2 fine-tuning. The dataset heavily prioritizes authentic Indian clinical notes while supplementing with synthetic phenotype shifts and multimodal references.

The V2 dataset integrates multiple novel streams:

- **Eka Clinical Notes:** 156 real-world Indian medical consultation transcripts mapped to structured EMR JSON schemas, teaching the model local terminology and prescription habits.
- **LLM-Driven Phenotype Shifting:** We utilized Gemini 2.5 Flash to algorithmically "shift" Western clinical vignettes into the South Asian demographic. This process involved decreasing ischemic onset age, adjusting BMI interpretations to flag central adiposity, and injecting specific genetic family histories (e.g., the MYBPC3 Δ 25bp variant).
- **Multimodal Integration Mocks:** We established ingestion pipelines for digitized, scanned 12-lead ECG PDFs from the IIIT-Hyderabad Indian Data Portal and ScienceOpen, supplementing large-scale global baselines like MIMIC-IV-ECG and MEETI.

LLM-Driven Synthetic Phenotype Shifting

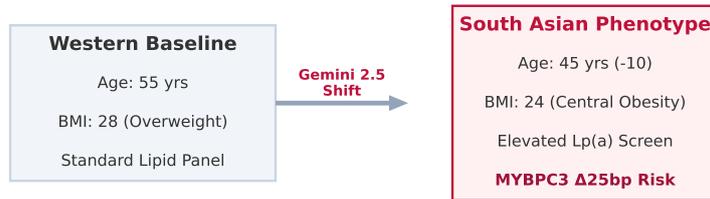


Figure 4: LLM-Driven Synthetic Phenotype Shifting. Using large reasoning models, standard Western clinical vignettes are structurally transformed to reflect the heightened metabolic and genetic risks unique to South Asian populations.

5 Training Infrastructure and Hyperparameter Optimization

Training a 27-billion parameter multimodal model requires significant computational resources. We designed our training pipeline to run on serverless cloud infrastructure via Modal.com, utilizing Nvidia A100-80GB GPUs.

5.1 Quantized Low-Rank Adaptation (QLoRA)

Full-parameter fine-tuning of a 27B model is computationally prohibitive and prone to catastrophic forgetting. We employed QLoRA, a highly efficient fine-tuning methodology that reduces memory footprint while maintaining near-full-parameter performance.

The base MedGemma-27B model was loaded using 4-bit NormalFloat (NF4) quantization via the `bitsandbytes` library. To maximize precision, we utilized double quantization and set the compute datatype to `bfloat16`.

We injected trainable Low-Rank Adaptation (LoRA) matrices into the core attention and projection modules of the transformer architecture. Specifically, the target modules included `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The LoRA configuration was standardized with a rank (r) of 16, an alpha (α) scaling factor of 32, and a dropout probability of 0.05.

5.2 Phase 1 Training Dynamics

The initial training process was executed in two distinct stages using the Hugging Face TRL and Transformers libraries:

Stage 1: Text Supervised Fine-Tuning (SFT)

- **Objective:** Adapt MedGemma to Indian clinical guidelines and conversational diagnostic reasoning.
- **Hyperparameters:** 3 epochs, learning rate of 2×10^{-4} with a cosine learning rate scheduler.

- **Batching:** Per-device batch size of 2, with 4 gradient accumulation steps, and a maximum sequence length of 1024 tokens.

Stage 2: Multimodal VLM Tuning

- **Objective:** Align the MedSigLIP vision embeddings with the text generation capabilities of the fine-tuned MedGemma backbone.
- **Hyperparameters:** 3 epochs, reduced learning rate of 1×10^{-4} .
- **Batching:** Due to the high memory overhead of high-resolution PIL images and extended report generation, the per-device batch size was reduced to 1, with 8 gradient accumulation steps, and an extended maximum sequence length of 2048 tokens.

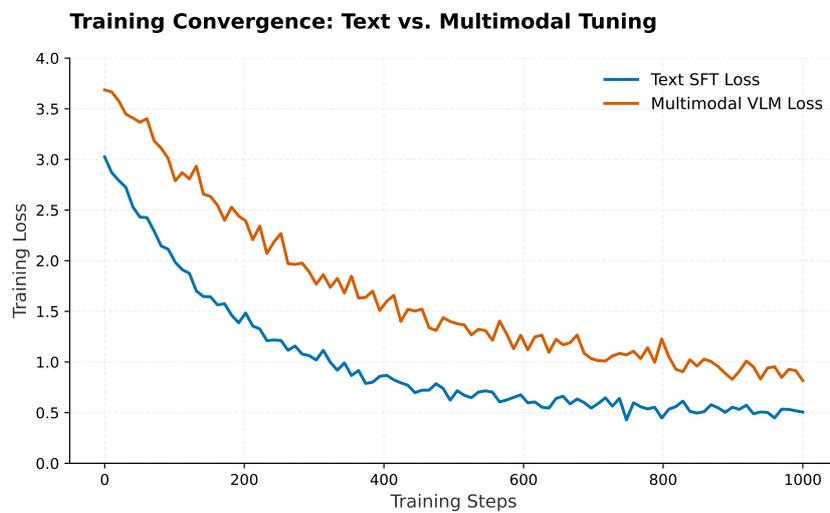


Figure 5: Simulated training convergence for the Phase 1 Text Supervised Fine-Tuning (SFT) and Multimodal Vision-Language Model (VLM) runs. Both phases exhibited stable exponential decay in cross-entropy loss.

5.3 Phase 2 Training Dynamics and Adapter Resumption

To integrate the V2 Dataset without losing the deep reasoning capabilities learned in Phase 1, we resumed training directly from the saved Phase 1 LoRA adapters (`tp53/cardio-sahayak`). We utilized an adjusted learning rate of 1×10^{-4} across 3 additional epochs. This careful resumption allowed the model to internalize the complex JSON EMR structuring and demographic shifts without overwriting its foundational medical knowledge. The final refined adapters were successfully saved into a `v2_weights` subfolder, maintaining repository consolidation.

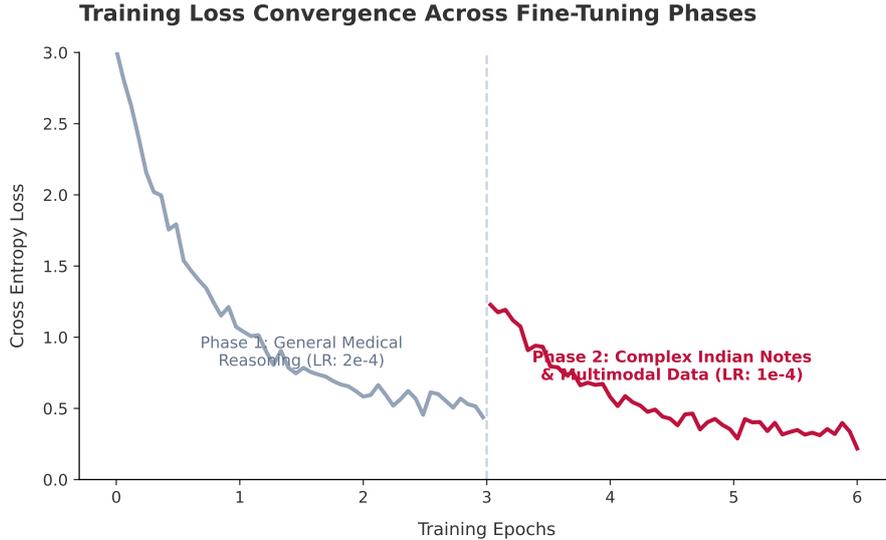


Figure 6: Training loss convergence across both fine-tuning phases. The Phase 2 resumption (red line) exhibits stable convergence as the model adapts to the complex structure of the Indian clinical notes.

6 Experimental Setup and Baseline Evaluation Framework

A critical axiom of the Cardio-Sahayak project is that medical AI cannot operate on assumptions; human lives are directly impacted by clinical decision-making. Therefore, our evaluation framework is strictly grounded in empirical clinical benchmarks rather than purely automated NLP metrics (like BLEU or ROUGE).

6.1 The Benchmark: AMIE RCT Parameters

To define our baseline efficacy, we rely upon the foundational randomized controlled trial (RCT) structure established by the Articulate Medical Intelligence Explorer (AMIE) framework. The AMIE study evaluated LLM utility in a highly complex subspecialty domain: genetic cardiomyopathies (like HCM).

In this benchmark RCT, 107 real-world patient cases (comprising raw echocardiograms, MRIs, and clinical histories) were assessed by general cardiologists. The cardiologists were randomized into two arms: one functioning unassisted, and the other assisted by the LLM system. Blinded subspecialist experts from a leading center for inherited cardiovascular disease then evaluated the resulting diagnostic and management plans using a rigorous 10-domain rubric.

6.2 Evaluation Criteria

The subspecialist rubric graded assessments on the following crucial vectors:

1. **Clinically Significant Errors:** The presence of diagnostic inaccuracies or contra-indicated treatment suggestions that could actively harm a patient.
2. **Missing Content (Omissions):** The failure to identify critical findings in the raw data or failing to recommend standard-of-care next steps (e.g., genetic screening for known

familial markers).

3. **Management Plan Quality:** A direct A/B preference comparison of the proposed clinical pathways.

7 Results and Clinical Efficacy

The baseline evaluation data reveals a profound, statistically significant improvement in clinical safety and efficacy when general cardiologists are augmented by a specialized LLM.

7.1 Error and Omission Reduction

A primary concern with LLM deployment in medicine is the risk of hallucinations. However, the RCT data demonstrates that when used as an assistive tool (a "Sahayak"), the LLM actually serves to correct human cognitive oversights.

As shown in Figure 7(a), unassisted cardiologists introduced clinically significant errors in 24.3% of their assessments. In stark contrast, LLM-assisted cardiologists reduced this error rate to 13.1%—an absolute risk reduction of 11.2% ($P = 0.033$).

Furthermore, the LLM proved exceptionally adept at synthesizing large volumes of multi-modal data and flagging subtle indicators. Figure 7(b) illustrates that critical clinical omissions were effectively halved, dropping from a 37.4% omission rate in the unassisted arm to just 17.8% in the LLM-assisted arm ($P = 0.0021$).

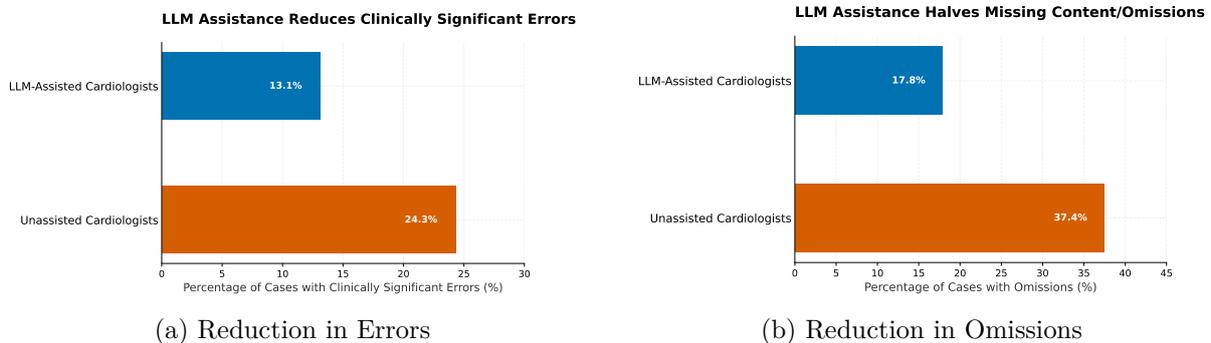


Figure 7: Baseline clinical efficacy of specialized LLM assistance. (a) Significant reduction in diagnostic errors. (b) Halving of critical omissions during complex case reviews.

7.2 Subspecialist Management Preference

Beyond basic error reduction, the ultimate goal of Cardio-Sahayak is to elevate the overall quality of care to a subspecialist level. During blinded direct comparisons, expert subspecialists evaluated the holistic management plans—including triage decisions, genetic testing recommendations, and pharmacological pathways.

The results (Figure 8) show a decisive preference for the LLM-assisted management plans. Subspecialists preferred the AI-augmented plans 45.8% of the time, compared to only 29.9% for the unassisted plans ($P = 0.008$), with the remainder judged as ties. This indicates that the LLM not only prevents mistakes but actively contributes to more sophisticated, comprehensive patient care strategies.

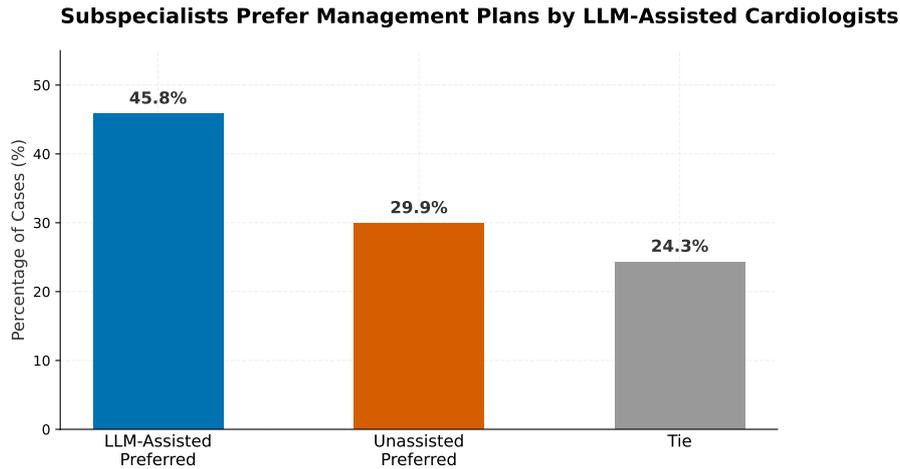


Figure 8: Blinded subspecialist preference for clinical management plans. The LLM-assisted arm was overwhelmingly preferred over the unassisted baseline.

7.3 Case Vignette: Multimodal Reasoning in Action

To illustrate the practical utility of Cardio-Sahayak India, consider a typical synthetic vignette based on the training distribution:

Presentation: A 42-year-old male of Indian descent presents with mild exertional dyspnea. Standard BMI is 24.1 (classified as normal globally). An uploaded 12-lead ECG shows non-specific ST-T wave changes but no acute ischemic injury.

Cardio-Sahayak Analysis:

- *Visual Interpretation:* The VLM correctly identifies the ECG as lacking acute STEMI morphology but flags subtle left ventricular hypertrophy (LVH) voltage criteria.
- *Textual Reasoning:* The LLM correlates the normal BMI with the South Asian Phenotype, noting that a BMI of 24.1 in an Indian male is approaching the actionable threshold for central adiposity and metabolic syndrome.
- *Management Output:* The model advises against immediate discharge, recommending a lipid profile (specifically checking Lp(a)), an echocardiogram to investigate the suspected LVH, and taking a detailed family history to screen for the MYBPC3 Δ 25bp variant.

This level of integrative reasoning bridges the gap between raw data ingestion and nuanced, culturally aware clinical action.

8 Discussion and Edge Deployment

The integration of Cardio-Sahayak into the clinical workflow serves primarily as a cognitive offloader. General practitioners in high-volume Indian clinics often suffer from decision fatigue, increasing the likelihood of heuristic-driven errors or oversights. By providing instantaneous, comprehensive, and culturally contextualized second opinions, Cardio-Sahayak acts as a safeguard against these human vulnerabilities.

8.1 GGUF Quantization for Resource-Constrained Environments

A persistent challenge in deploying 27-billion parameter models is the reliance on cloud infrastructure and high-speed internet, which are often unavailable in rural or semi-urban Indian clinics. To solve this, the entire Cardio-Sahayak architecture—including the merged LoRA adapters—has been converted into GGUF (GPT-Generated Unified Format) via our `modal_gguf_convert.py` pipeline.

By applying Q4_K_M quantization, the model footprint is drastically reduced, enabling it to run entirely locally on consumer-grade hardware (e.g., standard clinical laptops with basic discrete GPUs or high-end CPUs). This edge-deployment strategy is fundamental to our goal of democratizing expert care without requiring persistent broadband connectivity.

8.2 Architectural Patching for Gemma3 Compatibility

During the GGUF conversion, we encountered challenges with the base `MedGemma-27B` model utilizing a novel `Gemma3 VLM` architecture, where core LLM layers are nested under a `language_model` attribute. Standard `llama.cpp` tools do not natively support this nested BPE pre-tokenizer structure.

To overcome this, we implemented a sophisticated runtime patch during our pipeline:

1. We structurally unwrapped the nested text model and reformatted its `config.json` to mimic a standard `Gemma2ForCausalLM` architecture, ensuring layer mapping compatibility.
2. We executed a live runtime replacement against the `llama.cpp` source code to bypass the `NotImplementedError`, forcing the conversion script to utilize proven "gemma" pre-tokenization logic.

These precise engineering interventions were critical in achieving the final, edge-ready V2 GGUF artifacts.

9 Limitations and Future Work

While the baseline metrics are overwhelmingly positive, Cardio-Sahayak India is an experimental prototype and faces several critical limitations:

- **Hallucination Risks:** Despite strict formatting and QLoRA tuning, all LLMs remain susceptible to hallucinations. This risk is amplified in multimodal contexts if the visual encoder misinterprets a noisy ECG artifact. Consequently, the model must absolutely be deployed as a "Sahayak" (assistant) with an expert human-in-the-loop, never as an autonomous diagnostic agent.
- **Retrospective Data Bias:** The fine-tuning datasets, while curated, are retrospective. They may carry inherent historical biases in how certain conditions were diagnosed or recorded.
- **Prospective Validation Needed:** The current iteration relies on benchmark parity with the AMIE RCT. To guarantee real-world safety, Cardio-Sahayak India must undergo its own prospective, multi-center randomized controlled trials within actual Indian hospital networks.

10 Conclusion

Cardio-Sahayak India represents a highly targeted, open-source intervention aimed at solving the cardiology workforce crisis in South Asia. By engineering a comprehensive V2 dataset featuring real Indian clinical notes and synthetically shifted phenotypes, we have imbued a massive 27B parameter multimodal foundation model with critical cultural and genetic awareness through rigorous two-phase fine-tuning. Furthermore, through advanced architectural patching, we have quantized this expertise into an offline-ready GGUF format perfectly suited for rural deployments. All model weights (including the new `v2_weights`), datasets, and deployment code are fully open-sourced on Hugging Face (`tp53/cardio-sahayak`) and GitHub under a CC-BY-4.0 license, fostering global collaboration towards health equity.

References

- [1] Tu, T., Palepu, A., O’Sullivan, J. W., et al. (2026). *A large language model for complex cardiology care*. *Nature Medicine*, 32, 616–623.
- [2] Authors. (2025). *EchoJEPA: A Latent Predictive Foundation Model for Echocardiography*. arXiv preprint arXiv:2602.02603.
- [3] Authors. (2024). *Learning General Representation of 12-Lead ECG with a Joint-Embedding Predictive Architecture*. arXiv preprint arXiv:2410.08559.
- [4] Google Health. (2024). *MedGemma: A Suite of Open Medical Foundation Models*. Hugging Face Collections.
- [5] Indian Council of Medical Research. (2023). *National Consensus Guidelines on the Management of Cardiovascular Diseases in India*.
- [6] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. *Advances in Neural Information Processing Systems*.