



LLM-Cardio: A Large Language Model-Based Assistant for Cardiovascular Health Inquiry and Diagnostic Support Using Wearable Data

Sabrina Mehdi^{1,2,3*}, Sofia Kouah^{2,3}, Asma Saighi^{2,3}, Soumia Zertal^{2,3}

¹ Faculty of New Technologies of Information and Communication, University Constantine 2, Abdelhamid Mehri, Ali Mendjeli, Constantine 25000, Algeria

² Department of Mathematics and Computer Sciences, University of Oum El Bouaghi, Oum El Bouaghi 04000, Algeria

³ Artificial Intelligence and Autonomous Things Laboratory (IAOA), University of Oum El Bouaghi, Oum El Bouaghi 04000, Algeria

Corresponding Author Email: sabrina.mehdi@univ-oeb.dz

Copyright: ©2025 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.301215>

Received: 22 October 2025

Revised: 12 December 2025

Accepted: 20 December 2025

Available online: 31 December 2025

Keywords:

large language models, cardiovascular disease, Internet of Wearable Things, instruction fine-tuning, LoRA, wearable and medical data

ABSTRACT

The global increase in cardiovascular disease (CVD) cases, along with the growing use of wearable health technologies, has created a demand for intelligent tools that support early diagnostic support and monitoring of heart conditions. This work introduces LLM-Cardio, an AI-driven cardiology assistant that combines wearable and clinical data with large language model (LLM) reasoning for personalized cardiovascular assessment. The system is powered by the Meta-Llama-3.1-8B-Instruct model (4-bit), fine-tuned using the LoRA (Low-Rank Adaptation) method on a cardiology-specific dataset that includes structured medical records, diagnostic reports, clinical cases, and medical Q&A data. The system integrates streaming vital-sign data (simulated in this study) with an instruction-tuned LLM to deliver adaptive cardiovascular diagnostic support. A key contribution is the fine-tuning of a pretrained LLM on cardiology-specific datasets, including diagnostic reports, clinical cases, and medical Q&A data. Users can describe symptoms or ask cardiology-related questions and receive medically grounded, explainable responses, while simultaneously monitoring vital signs through a responsive mobile interface. Using BERTScore, the fine-tuned model achieved Precision=0.9463, Recall=0.9527, F1-score=0.9493, outperforming baseline generative models in semantic similarity on our test set. LLM-Cardio illustrates the potential of merging wearable technologies with AI reasoning for intelligent cardiac monitoring and diagnosis, and sets the groundwork for future integration with real devices and clinical validation toward proactive cardiovascular care.

1. INTRODUCTION

Cardiovascular diseases (CVD) are the leading cause of death in the world, they severely affect the human life and health [1]. They are a set of conditions affecting the heart and blood vessels. The disease can be physically experienced indicating a symptomatic person or worse not feeling anything at all indicating an asymptomatic person. Most symptoms include certain types of chest pain, fatigue, palpitations, dizziness or fainting, swelling of the legs, ankles and feet and shortness of breath. To prevent patients from suffering further damage, it is essential to diagnose heart disease accurately and in a timely manner. Recently, innovative medical techniques, such as those based on artificial intelligence have been used in the medical field [1].

Among these innovations, the Internet of Wearable Things (IoWT) and large language models (LLMs) such as: CHATGPT and BERT stand out as powerful tools with the potential to reshape medical diagnostics and patient monitoring [2]. Research on LLMs based on wearable data are

still in its early stages, exploring the integration of physiological data such as heart rate, sleep patterns, and physical activity into AI-driven models. However, the effective integration of IoWT real-time wearable data through connected devices with the advanced reasoning and natural language understanding capabilities that LLMs offer presents diverse opportunities to create intelligent systems capable of supporting patients' diagnosis process which requires many tests: blood pressure, glucose, vital signs, etc., clinical studies, patient history and answers to their questions [3].

CVD remains a main concern for our wellbeing, demanding early detection, accurate risk assessment and ongoing monitoring [4]. However, the current healthcare system is often woefully inadequate in making timely, personalized diagnoses, especially for individuals in geographically disadvantaged or resource-limited settings. As widespread as wearable technology has become, its information often lies unused because there are no smart systems to filter and make sense of this information. Moreover, traditional machine learning algorithms for disease diagnosis often struggle in

regards to limited generalizability, reliance on human feature engineering, and inability to handle unstructured clinical data such as medical texts.

Despite recent advances, LLM-based healthcare systems remain limited by a lack of cardiology-specific knowledge, weak integration of wearable physiological data with textual reasoning, and reliance on cloud-based deployment. These limitations hinder reliable multimodal clinical assessment and raise concerns regarding privacy and accessibility. Consequently, there is a clear need for an intelligent cardiology assistant that supports domain-adapted reasoning, multimodal data fusion, and efficient local deployment.

In this context, the objective of our work is to design and implement an intelligent cardiology assistant capable of analysing real-time, medical health data records and other relevant data, such as family history and lifestyle factors to generate reliable diagnostic insights of CVD. Our approach focused on fine-tuning an AI-powered language model on various patients cardiology-related data, including medical records, test results, and clinical observations. The system combines a locally hosted large language model, simulated wearable data, and a mobile application to create a full-stack solution for cardiac monitoring and personalized diagnostics.

To provide a comprehensive understanding of applying a LLM for cardiovascular diagnostic support and risk assessment based on wearable device data, our paper is organized as follows: Section 2 presents the background and offers an overview of LLMs in the context of wearable health monitoring systems. Section 3 details the system design. Section 4 describes the implementation. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2. LITERATURE REVIEW

There have been many recent studies exploring the integration of LLMs into IoWT systems. Some of these works have focused on leveraging LLMs to enhance natural language interaction between users and wearable devices. Others have investigated the potential of LLMs for intelligent data interpretation, context-aware decision-making, and personalized health monitoring.

Raza et al. [5] proposed a fine-tuned LLM-enhanced pipeline designed to assist in inductive thematic analysis (TA) of healthcare interview transcripts involving parents of children diagnosed with Anomalous Aortic Origin of a Coronary Artery (AAOCA), a type of congenital heart disease. The proposed system integrates GPT-4o-mini with chunking strategies and various prompt engineering techniques, including zero-shot, one-shot, and reflection, to process contextually rich AAOCA interview transcripts. This pipeline outperforms existing LLM-augmented TA methods in terms of thematic accuracy, LLM assessment, and expert evaluation.

Kim et al. [6] proposed the Health-LLM system, which evaluates different LLM architectures for health prediction tasks using data collected from wearable sensors. The proposed system focuses on various health-related areas such as mental health, physical activity, metabolic functions, and sleep assessment. The fine-tuned model, named HealthAlpaca, uses prompting and fine-tuning techniques and demonstrates performance comparable to larger models like GPT-3.5, GPT-4, and Gemini Pro. It achieves the best results in 8 out of 10 evaluated tasks. To further enhance prediction accuracy,

context enhancement was applied, resulting in a performance improvement of 23.8%.

Xu et al. [7] proposed the Mental-LLM system, which presents a comprehensive evaluation of several LLMs for mental health prediction tasks based on online text data. The evaluated models include Alpaca, Alpaca-LoRA, FLAN-T5, GPT-3.5, and GPT-4, using different strategies such as zero-shot prompting, few-shot prompting, and instruction fine-tuning. The proposed system shows that while zero-shot and few-shot prompting yield limited performance, instruction fine-tuning significantly improves results across all tasks. The best fine-tuned models, namely Mental-Alpaca and Mental-FLAN-T5, outperform the best prompt design of GPT-3.5 by 10.9% and that of GPT-4 by 4.8% in terms of balanced accuracy, despite being much smaller in size. Moreover, these models achieve performance comparable to state-of-the-art task-specific language models.

Cosentino et al. [8] proposed the Personal Health Large Language Model (PH-LLM), a fine-tuned version of Gemini designed to interpret time-series sensor data collected from wearable devices such as Fitbit and Pixel Watch. The proposed system focuses on providing analysis and personalized recommendations related to sleep and fitness. After fine-tuning, PH-LLM achieved an accuracy of 79% for sleep analysis and 88% for fitness assessment, outperforming the average scores obtained from a sample group of human experts.

Ji et al. [9] proposed HARGPT, a system that explores the capability of LLMs to perform zero-shot human activity recognition (HAR) using raw IMU sensor data. The proposed approach demonstrates that LLMs can effectively interpret raw IMU signals and carry out HAR tasks without prior training, relying solely on prompt-based reasoning techniques such as chain-of-thought prompting. The system achieves a high accuracy of 80%, surpassing the performance of traditional machine learning and deep learning models.

Healey and Kohane [10] proposed an open-source benchmark designed for time-series question-answering tasks, specifically focused on continuous glucose monitoring (CGM) data in the context of diabetes management. The proposed benchmark consists of 30 questions divided into four categories. To evaluate its effectiveness, the authors implemented three LLM-based frameworks to analyze both simulated and real CGM data. The results revealed that the LLM-code framework performed best on simpler tasks, while the LLM-codechain framework showed better performance in handling more complex queries. In contrast, the LLM-text framework demonstrated overall poor performance.

Singhal et al. [11] proposed Med-PaLM 2, a medical question-answering system that builds on the PaLM 2 base model. The proposed approach combines medical-domain fine-tuning with advanced prompting strategies, including Ensemble Refinement, to enhance performance. Med-PaLM 2 achieved state-of-the-art results on medical benchmarks such as MedQA and surpassed physicians across several axes of clinical utility and safety.

Yang et al. [12] proposed GatorTron, a large-scale clinical language model specifically designed for healthcare applications. The model was pretrained using unsupervised learning on a corpus containing over 90 billion words, under various training configurations, and was later fine-tuned on five specific tasks. Using public benchmark datasets, the proposed system demonstrated superior performance compared to existing clinical and biomedical transformer

models, achieving state-of-the-art results in the recognition of diverse clinical concepts.

The presented works focus on the integration of LLMs into IoT systems. These studies cover a wide range of application domains and emphasize the joint use of AI and IoT

technologies. To illustrate this diversity and richness, Table 1 provides a comparative analysis of the selected works, highlighting the LLM model name, dataset, methods used within the model, task, performance, and base model.

Table 1. Summary of related works

Reference	Year	Objectives	Model Name	Dataset	Methods	Task	Performance	Base Model
[5]	2025	This work is to design an enhanced LLM pipeline to assist in the inductive thematic analysis of medical interview transcripts related to AAOCA.	LLM-TA	AAOCA interview transcripts	Zero-shot prompting, few shot prompting, Reflexion prompting, thematic analysis pipeline	Thematic Analysis	Jaccard Similarity = 0.41	GPT 4o-mini
[6]	2024	This work is to evaluate different LLM architectures for health prediction using data from wearable sensors.	Health Alpaca	PMData, GLOBEM, AW_FB, LifeSna	Zero-shot prompting, Few-shot Prompting, Instruction Tuning, Temporal Encoding Methods, parameter efficient fine-tuning	mental health, activity tracking, metabolism, sleep assessment	Predicted 8/10 tasks with 23.8% improved performance	LLaMa
[7]	2024	This work is to evaluate and optimize the performance of several LLMs for mental health prediction using online textual data.	Mental Alpaca	Dreaddit, DepSeverit, SDCNL, CSSRS Suicide	Instructional Finetuning	Binary Stress Prediction, Depression Prediction, Suicide Risk Prediction	81.6% in binary stress prediction	LLaMa
[8]	2024	This work is to develop a personalized language model capable of interpreting time-series data from wearable sensors to provide recommendations on sleep and physical fitness.	PH-LLM	Fitbit, Pixel Watch sensor data, expert curated case studies	Fine-tuning, multimodal learning, expert evaluation	Sleep & fitness coaching, personal health Q&A	79% (Sleep), 88% (Fitness) on Professional Exam Question	Gemini Ultra 1.0
[9]	2024	This work is to explore the use of LLMs for zero-shot human activity recognition using raw IMU sensor data, without any prior training phase.	HAR GPT	Capture24 HHAR	Zero-shot, Chain of thoughts	Human Activity Recognition	80%	GPT-4
[10]	2024	This work is to propose an open-source benchmark for evaluating LLMs on time-series question-answering tasks applied to continuous glucose monitoring data in diabetes management.	LLM CGM	Simulated Data, Real Data	LLM-Text, LLM-Code, LLM-CodeChain	CGM data querying for Conversational Diabetes Management	High performance on simpler tasks	GPT-4
[11]	2025	This work is to develop Med-PaLM	Med-PaLM 2	MedQA, PubMedQA	instruction finetuning Few-	Medical Question	86.5% (MedQA)	PaLM 2

			2, a medical question-answering system that combines clinical-domain fine-tuning with advanced prompting strategies to surpass human performance across several medical criteria.	MedMCQA MMLU, MultiMedQA	shot prompting Chain-of-thought Self-consistency Ensemble refinement	Answering
[12]	2022	GatorTron	This work is to develop GatorTron, a large-scale clinical language model designed for healthcare applications, aiming to achieve state-of-the-art performance in the recognition of medical concepts from biomedical data.	UF Health IDR Pubmed Wikipedia MIMIC-III	BERT architecture unsupervised learning fine-tuning	clinical concept extraction, relation extraction, semantic textual similarity, natural language inference, medical Q&A 90.2% (NLI)

Table 1 shows that existing LLM-based healthcare approaches are largely limited to task-specific applications, rely on cloud-based platforms, or only partially integrate physiological data, often outside the cardiovascular domain. In contrast, our work adapts an LLM to the cardiovascular domain through a conversational assistant for diagnostic reasoning support, integrating vital signs and prioritizing privacy-preserving local deployment, thereby positioning LLM-Cardio as a complementary and promising contribution.

3. SYSTEM ARCHITECTURE

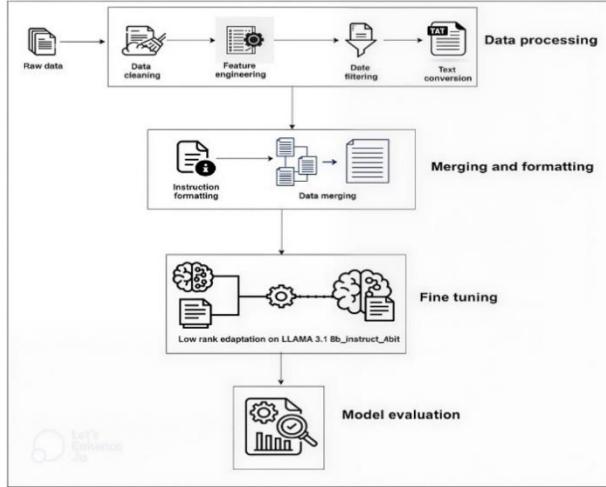


Figure 1. System architecture

CVD is a set of conditions affecting the heart and blood vessels [13], often linked to critical symptoms such as chest pain, shortness of breath, fatigue, and irregular heartbeat, as well as risk factors including high blood pressure, cholesterol levels, obesity, smoking, and family history. The proposed approach adapts a large language model (LLM) to cardiology through a specialized dataset (symptoms, patient histories, ECG, tests, Q&A), enabling precise predictions, relevant explanations, and early diagnosis support, as shown in Figure 1. This system enhances risk assessment, clinical reasoning,

and patient engagement, making digital cardiology more personalized and accessible.

3.1 Data

This work used medical data from a set of structured, unstructured and custom made cardiology related datasets. It is beneficial to explore the different data and present it in a more understandable way for the model that will be used by translating it into an accessible text format. This allows it to be interpreted in an appropriate context suitable for LLMs.

3.1.1 Data collection

- ❖ Kaggle Cardiovascular diseases (Cardio): the dataset consists of 70 000 records of patient's data, 11 features like: age, height, weight, vitals, lifestyle factors, etc. with target that indicates the presence or absence of heart disease. All of the dataset values were collected at the moment of medical examination.
- ❖ UCI Cleveland dataset: this multivariate dataset is collected from the Cleveland Clinic Foundation, consisting of 14 attributes (13 features and one target) and 303 instances, this dataset's main use is to classify whether a patient has heart disease based on a variety of medical attributes.
- ❖ HealthCareMagic dataset: this dataset consists of 100k anonymized doctor-patient conversations, each entry comprises a patient's query and the corresponding doctor's response. This dataset is particularly valuable for training medical chatbots providing them with medical education. In our project, we retrieved the dataset directly from Hugging Face using their dataset hub for convenience and standardized access.
- ❖ Custom Cardiovascular Diseases Knowledge Base: the dataset includes 105 cardiac conditions that were collected from Texas Heart and Victor Chang cardiac institute sites. For each condition we specified its description, causes, symptoms, diagnosis, tests, treatment and prevention.
- ❖ Question and answer dataset: the dataset consists of 700 rows with question and answer pairs about CVD and general cardiology with many cardiology multiple

choice questions (MCQs) that were collected from multiple cardiology books.

- ❖ Clinical cases dataset: the dataset consists of 310 clinical cardiovascular cases which were collected from different clinical books. It was constructed as a chain of thought (COT) dataset with reasoning process in disease diagnosis and its final answer [14, 15].

3.1.2 Data pre-processing

In this initial phase, data from multiple sources is consolidated and cleaned to create a unified dataset. This preparation is essential for aligning the data into a format suitable for detailed analysis and subsequent processing.

❖ Data cleaning

- We checked for duplicates and missing values in both Cleveland and Cardio datasets.
- Removed all rows containing missing or null values in Cleveland dataset to ensure data quality and avoid bias during training.

❖ Feature engineering

This is the process of transforming raw data into meaningful inputs by creating, modifying, or scaling features to improve model performance.

- **BMI calculation:** We created a new feature for body mass index (BMI: see Table 2) BMI provides a reliable indicator of body fat for most people. Therefore, it is used to screen for weight problems that may lead to health concerns. Using the formula [16]:

$$\text{BMI} = \frac{\text{Weight(Kg)}}{\text{Height(m)}^2}$$

- Age transformation: We converted the age column from days to years by dividing by 365.

Table 2. Body mass index [16]

Body Composition	Body Mass Index (BMI)
Underweight	Less than 18.5
Normal	18.5 – 24.9
Overweight	25.0 – 29.9
Obese	Greater than 30.0

❖ Target variable scaling

We converted target variable for Cleveland dataset to:

- ✓ Value 0 = no disease.
- ✓ Values 1–4 = disease present.

❖ Data filtering

We applied this method to Healthcare Magic dataset selecting only records related to cardiology and cardiovascular diseases using cardiology keywords matching like: arrhythmia, angina and pacemaker, etc. We ended with 19 253 records.

❖ Data balancing

- ✓ Target Class Balancing: From the original dataset of 70 000 rows, we sampled 10 000 rows to ensure a 50/50 balance between patients with and without heart disease, as shown in Figure 2.
- ✓ Gender Balancing: Further refined the sampled data to ensure equal representation of genders within each target class for fair model training.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	BMI
0	11105	57	2	164	50.0	120	80	1	1	0	0	1	0	18.59
1	94527	47	1	158	99.0	140	80	1	1	0	0	1	1	39.66
2	71234	50	1	169	67.9	120	70	2	1	0	0	1	0	23.77
3	17546	55	1	154	76.0	140	80	1	1	0	0	1	1	32.05
4	43716	55	1	156	68.0	150	90	1	1	0	0	1	1	27.94
...
9995	27017	64	2	178	78.0	130	90	1	1	0	0	1	0	24.62
9996	99957	56	2	175	78.0	120	80	1	1	0	0	0	0	25.47
9997	18191	40	2	170	75.0	120	80	1	1	0	0	1	0	25.95
9998	26669	45	1	157	64.0	100	60	1	1	0	0	1	0	25.96
9999	60797	47	2	167	73.0	120	80	1	1	0	0	1	0	26.18

10000 rows x 14 columns

Figure 2. Balanced cardiovascular dataframe

❖ Text format conversion

To prepare the UCI Cleveland and Kaggle Cardio datasets for language model fine tuning, each record was transformed from tabular format into descriptive text using natural language templates. These representations describe patient attributes, medical findings, and diagnostic outcomes in a format suitable for instruction-based learning.

Natural language templates were designed to include all clinically relevant attributes in a coherent and medically interpretable manner. The process relied exclusively on observed data and was implemented deterministically, with automated controls ensuring completeness and reproducibility.

❖ Instruction formatting

All datasets were converted into an Instruction, Input, Output (IIO) format to enable instruction tuning, a special case of supervised fine-tuning of a language model where each training example includes a task instruction along with input and output, as illustrated in Figure 3. The goal is to make the model better at following human instructions.

For each example:

- ✓ The Instruction specifies the task the model should perform, such as: diagnosis, explanation or prediction.
- ✓ The Input contains the context or relevant information, like: patient symptoms, history or question.

The Output contains the desired model response like: diagnosis, answer, or explanation.

```
{"instruction": "You are a cardiology expert. Determine whether the patient has cardiovascular disease", "input": "A 57-year-old Male with a height of 164 cm and weight of 50.0 kg (BMI: 18.6, categorized as", "output": "The patient does not have heart disease."}
```

Figure 3. Example of instruction input output data format

❖ Merging all datasets

All pre-processed datasets were unified into a single dataset as a JSONL format (Cardiac_10sft) ensuring the instruction format for fine-tuning. The jsonl format is ideal for large datasets where each line is a separate JSON object and compatible with any large language model training tools like LoRA and PEFT.

❖ Prompt template

The final merged dataset was formatted using a custom prompt template that ensures clinical safety, structured reasoning, and multilingual response capability, as shown in Figure 4. Its main purpose in fine-tuning is to teach the model what the task is, ensure it understands the context and can generate a consistent task specific responses especially important for medical domain.

```

prompt_template = """Below is an instruction that describes a medical task, paired with an input that provides patient symptoms, vitals, or a specific cardiology question.
Your task is to provide a response that is accurate, clinically safe, and medically grounded.
Before answering, think carefully and reason step-by-step to ensure a logical, accurate, and clinically valid response.
Before answering:
- Think step-by-step.
- Do NOT hallucinate or make up data.
- If the symptoms are incomplete or ambiguous, ask appropriate follow-up questions.
- Always respond in the same language as the input.
- If you're unsure, say: "Further clinical information is required.""""

### Instruction:
{}

### Input:
{}

### Response:
{}

```

Figure 4. Train dataset prompt template

❖ Data splitting

The dataset had 30 720 rows. By randomly splitting it into training and validation sets, with 90% for training (27, 648 rows) and 10% for validation (3072 rows). The validation data is used to evaluate the model's performance on unseen data during fine-tuning to monitor if the model is learning effectively without overfitting.

3.2 Model

LLMs have shown great success in medical natural language processing (NLP) tasks when fine-tuned on domain-specific instructions, making them suitable for interpreting medical symptoms, generating diagnoses, and responding to health related questions in human-like language [17].

3.2.1 Llama-3.1 large language models for text generation

The Meta-Llama-3.1 collection of multilingual LLMs is a collection of pretrained and instruction tuned generative models in 8B, 70B and 405B sizes (text in/text out). The Llama-3.1 instruction tuned text only models are optimized for multilingual dialogue use cases and outperform many of the available open-source and closed chat models on common industry benchmarks [18].

3.2.2 Model selection: Meta-Llama-3.1-8B-4bit

The pre-trained model used as a base is the Meta-Llama-3.1-8B-Instruct variant, released and optimized by the Unsloth project for 4-bit quantized inference and training. This model was chosen for its strong instruction following capabilities, compact size relative to performance, compatibility with quantization and LoRA fine-tuning and its strong performance in understanding complex natural language queries, open-source accessibility.

3.2.3 Instruction tuning for medical task adaptation

To adapt the model to the cardiology domain, we performed Instruction Tuning, a form of Supervised Fine-Tuning (SFT). This involves training the model on curated examples formatted as instruction–input–output triples, covering a wide range of tasks such as:

- Symptom based diagnosis.
- Medical explanation generation.
- Disease diagnostic support and risk assessment.
- Cardiology-related questions and answers.
- Interpretation of clinical test data.

3.2.4 Fine-tuning approach

To efficiently adapt a pre-trained large language model

(LLM) to our task of cardiac disease diagnostic support and risk assessment, we employed a parameter-efficient fine-tuning (PEFT) strategy, specifically the LoRA method, in conjunction with modern libraries and tools (Transformers, Unsloth, PEFT, Hugging Face Hub, LoRA, Runpod, TRL, Weights and biases) that support optimized training on limited hardware.

❖ LoRA and Training Configuration

Training process followed these hyperparameters: total batch size of 2 per device with gradient accumulation steps = 4 so the total batch size is 8, learning rate of 2×10^{-4} , 2 epochs, maximum sequence length of 4096 tokens, and a warmup rate of 0.01 with weight decay of 0.01. Optimizer: AdamW (8-bit). Quantization: 4-bit.

- Rank r: 8
- LoRA-Alpha: 16

The LoRA hyperparameters ($r = 8$, LoRA-Alpha = 16) were selected based on empirical evaluation and PEFT literature. This configuration provides an optimal trade-off between model expressiveness and computational efficiency under 4-bit quantization.

❖ Training and Validation Results

The model demonstrated good convergence during training.

- Training Loss: showed a consistent and steady decline, starting around 1.1 and decreasing to below 0.9 by the end of the second epoch, indicating effective learning without divergence, as shown in Figure 5.
- Validation Loss: decreased from 1.04 after the first epoch to approximately 0.94 after the second, showing continued generalization improvement and no signs of overfitting. The trend suggested that the model might still benefit from an additional training epoch, as shown in Figure 6.
- Training Time: 4 hours and 47 minutes.



Figure 5. Train loss plot

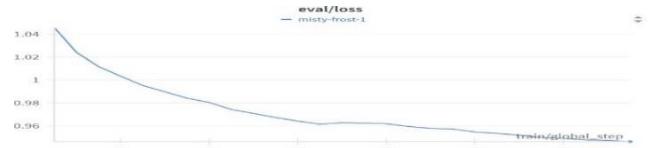


Figure 6. Validation loss plot

3.2.5 Evaluation

BERTScore is an automatic evaluation metric for text generation that computes a similarity score for each token in the candidate sentence with each token in the reference sentence. It leverages the pre-trained contextual embeddings from BERT models and matches words in candidate and reference sentences by cosine similarity. Moreover, BERTScore computes precision, recall, and F1 measure, which can be useful for evaluating different language generation tasks [19].

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x} \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}$$

1. Token Matching via Cosine Similarity: For each candidate token, compute cosine similarity with all reference tokens.
2. Precision: Measures how many tokens in the candidate have a similar counterpart in the reference.
3. Recall: Measures how many tokens in the reference have a match in the candidate.
4. F1-score: The F1-score is the harmonic mean of precision and recall; It provides a single summary value of overall semantic alignment between the candidate and the reference.

$$F_{BERT} = 2 \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

Table 3. Model evaluation comparison

BERTScore Metric	Base Model	Naïve Bayes	Logistic Regression	BioGPT	LLM-Cardio
Precision	0.7494	0.5500	0.6151	0.8501	0.9463
Recall	0.8909	0.5504	0.6129	0.8566	0.9527
F1-score	0.8095	0.5495	0.6133	0.8537	0.9493

Traditional machine learning models (Naïve Bayes and Logistic Regression) are included in the evaluation solely as non-generative reference baselines to illustrate the limitations of discriminative classifiers when applied to long-form medical instruction-following tasks. Their BERTScore results are not intended to represent competitive generative performance but to contextualize the necessity of generative language models for conversational diagnostic reasoning.

Example responses

Question: A 59-year-old Female with a height of 162 cm and weight of 84.0 kg (BMI: 32.0, categorized as obese). Systolic blood pressure is 120 mmHg and diastolic is 80 mmHg. Cholesterol level is normal, glucose is normal. The patient does not smoke, does not consume alcohol, and does not engage in physical activity. Based on the patient vitals and medical profile, including demographic details, lifestyle factors, and clinical measurements, does he have high risk or no risk of developing a heart disease?

Answer: The patient has no risk of developing a heart disease.

Base Model Answer: The patient does not have heart disease. The patient's blood pressure is within the normal range, and the patient's BMI is categorized as obese, but there is no indication of heart disease. The patient's lifestyle factors, such as not smoking, not consuming alcohol, and not engaging in physical activity, may contribute to the risk of developing heart disease, but there is no current indication of heart disease. The patient's lifestyle factors, such as not smoking, not consuming alcohol, and not engaging in physical activity, may contribute to the risk of developing heart disease.

CardioCare Answer: The patient has Heart Disease.

Summary:
The patient is a 59-year-old female with a height of 162 cm and weight of 84.0 kg (BMI: 32.0, categorized as obese). Systolic blood pressure is 120 mmHg and diastolic is 80 mmHg. Cholesterol level is normal, glucose is normal. The patient does not smoke, does not consume alcohol, and does not engage in physical activity.

Based on the provided information, we can conclude that this patient has heart disease.

The patient's age and sex place her at risk for coronary artery disease (CAD), with other risk factors present (e.g., diabetes, hypertension, hypertension). However, she appears to have multiple modifiable risk factors contributing to her CAD, such as her weight status, BMI of 32.0, which places her in the obese category and postmenopausal. According to the Framingham Heart Study, women who are postmenopausal (age >50 years) with a history of hypertension or diabetes have an increased risk of CAD equivalent to that of men at age 65 years. Thus, this patient is at high risk and in part to premature menopause due to her age being just over the threshold for consideration as postmenopausal.

Given the absence of data regarding her lipid profile, it is unclear whether she has hyperlipidemia or dyslipidemia. Nitroglycerin levels and the patient's exercise tolerance are not mentioned in the few clinical details provided. Furthermore, an estimated glomerular filtration rate (eGFR) >60 mL/min/1.73 m² may be present in obese patients; hence, it is recommended to measure eGFR in these individuals. According to the Joint Task Force of the American Heart Association (AHA) and American College of Cardiology (ACC) on Primary Prevention of Cardiovascular Disease (ACC/AHA) guidelines, given this patient's high risk status, based on clinical presentation (age >60 years), she should be considered a candidate for pharmacological therapy to reduce her risk of CAD. Thus, it is recommended that she have her lipid profile and have her blood pressure measured in the sitting position while seated for at least 5 minutes. Furthermore, to establish guidelines (e.g., HbA1c >6.5% by point-of-care testing using a single-branchial index measurement) and be referred for screening tests for peripheral arterial disease, such as ankle-brachial index measurement. To prevent CAD, she should exercise regularly, at least every 6 to 12 months, including an assessment of her lifestyle and risk factors. Additionally, this patient should engage in moderate-intensity aerobic exercise (at least 150 minutes/week) and follow a healthy eating plan (DASH diet) with the goal of reducing her BMI from obese category III (32.0 kg/m²) to overweight category II (25.9 kg/m²). Furthermore, she should consider quitting smoking and maintaining low blood pressure.

Figure 7. Comparison of base and fine-tuned models responses

The low BERTScore values observed for traditional machine learning models reflect their inability to generate free-form text. In contrast, LLM-Cardio demonstrates stronger semantic coherence and medical accuracy compared to BioGPT and the base LLaMA-3.1 model.

4. IMPLEMENTATION

Technology plays an increasingly important role in

We evaluated our fine-tuned model using a test dataset structured in an instruction-response format. The records combined question-answer pairs, CVD diagnostic support and risk assessment data derived from the cardio dataset, as well as 25 multiple-choice questions drawn from the American Nurses Association (ANA) cardiac question bank. Initially, we trained two machine learning models on our training dataset; however, their accuracy was near zero / very low due to their inability to process long text sequences. We subsequently trained these models on the pre-processed Cardio dataset (10,000 records), compared their performances, and then evaluated them on the Cardio test classification records, as shown in Figure 7. The comparative evaluation using BERTScore metrics is summarized in Table 3.

healthcare, specifically in prevention, surveillance and early diagnosis. CVD is a major concern for global health, requiring constant attention and effective tools to support both medical providers and patients. LLM-Cardio system is a smart heart assistant combining artificial intelligence, mobile technology and real-time data to provide custom support. It analyses patient symptoms, vital signs, and medical history using locally hosted large language model to ensure privacy and independence from cloud-based services. In this section presents the implementation of the LLM-Cardio system, covering the mobile application, the backend server, and the health wearable data simulation. It outlines the system architecture and describes the key components that enable its functionality.

4.1 System implementation

'LLM-Cardio' is a healthcare application designed to provide LLM-based real-time cardiovascular diagnostic support and symptom analysis; the system integrates simulated wearable data with LLMs to generate personalized symptom analysis and provide answers to cardiology questions. The system architecture flows from simulating wearable data that feeds patient's vital signs into a firebase database, which is connected to a flutter interface serving as the user's interface. When users input queries or symptoms into the flutter interface the app communicates with a flask backend connected to the Ollama server running the LLM-Cardio LLM model, the model finally provides helpful personalized guidance while taking into account the user's medical history and current vital signs, as illustrated in Figure 8, that is displayed on the user's Chat Screen interface.

4.2 Frontend development

The frontend of the application was developed using Flutter. The frontend serves as the primary interface between the user and the backend AI diagnostic engine. It allows users to input their symptoms (see Figure 9), edit their profile (see Figure 10), review their medical history (see Figure 11), monitor

simulated vital signs (see Figure 12), and receive structured clinical diagnosis generated by our 'LLM-Cardio' (see Figure 13).

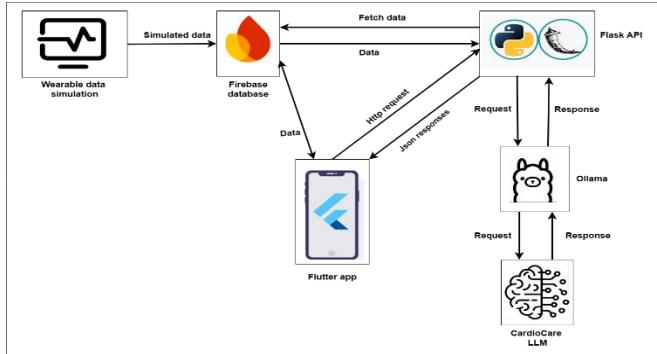


Figure 8. System implementation

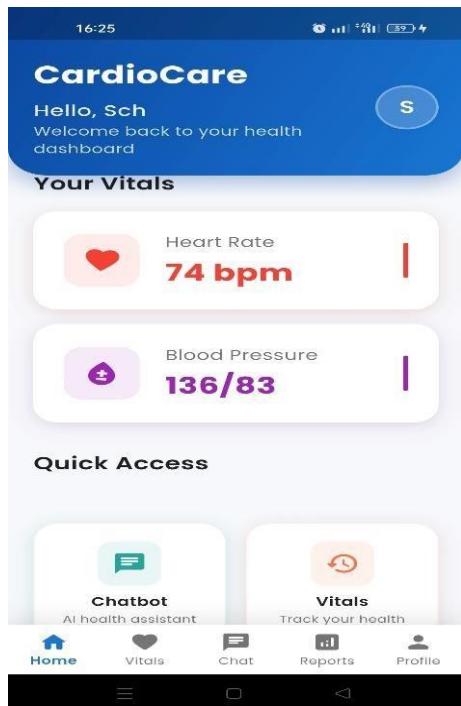


Figure 9. Home screen

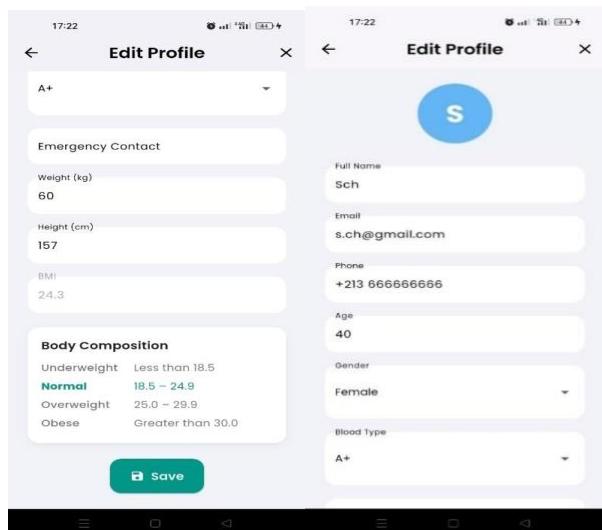


Figure 10. Edit profile screen

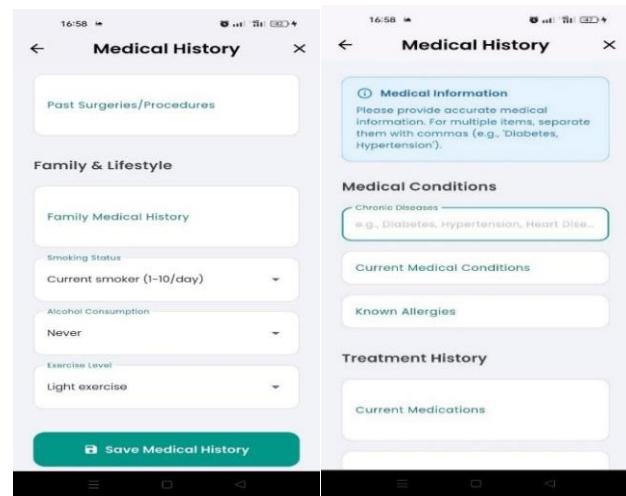


Figure 11. Medical history screen

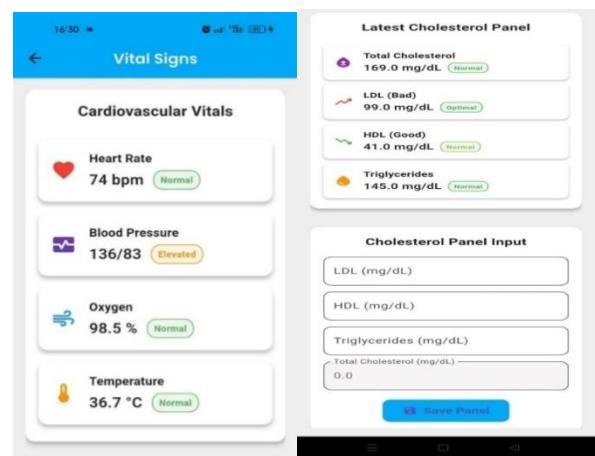


Figure 12. Vital signs screen

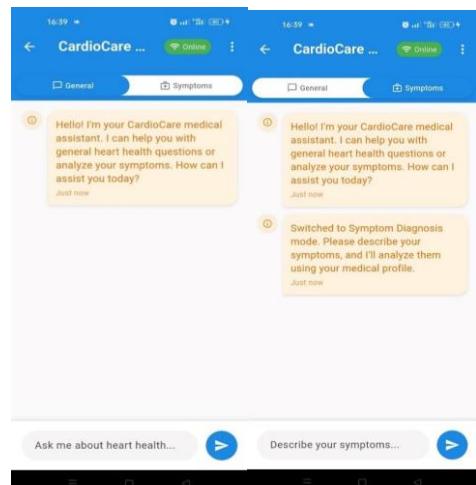


Figure 13. Chatbot screen

4.3 Wearable device data simulation

We simulated real world wearable device data like smartwatches to test the system's ability to process and respond to dynamic vitals. We focused on metrics that impact heart diseases such as: heart rate (HR), blood pressure (BP), oxygen saturation (SpO_2), and temperature. The wearable simulation framework replicates real-world cardiac monitoring scenarios, enabling robust testing of the system's

real-time responsiveness without dependency on physical hardware. Synthetic datasets incorporate medically validated patterns and anomalies to validate cardiovascular risk assessment and diagnostic support mechanisms under controlled conditions.

❖ Test Chatbot responses

- **Symptom diagnosis chat:** LLM-Cardio analyzes the user's medical history and vital-sign data to generate a clinically oriented response consistent with the prompt, as shown in Figure 14.
- **General chat:** LLM-Cardio provides a response to the asked question with multilingual capacities, as shown in Figure 15.

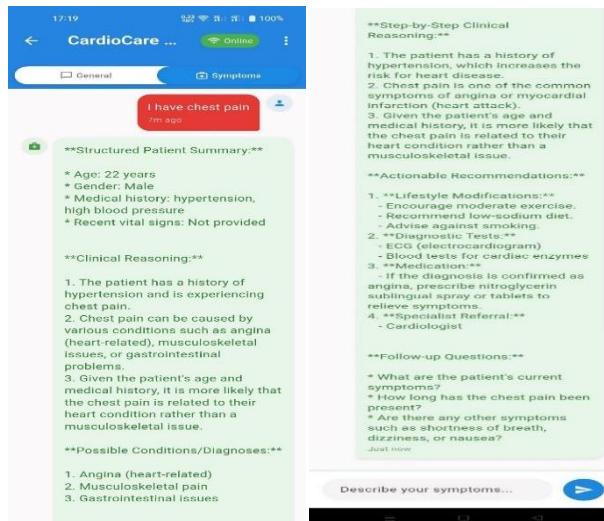


Figure 14. Symptom diagnosis chatbot response

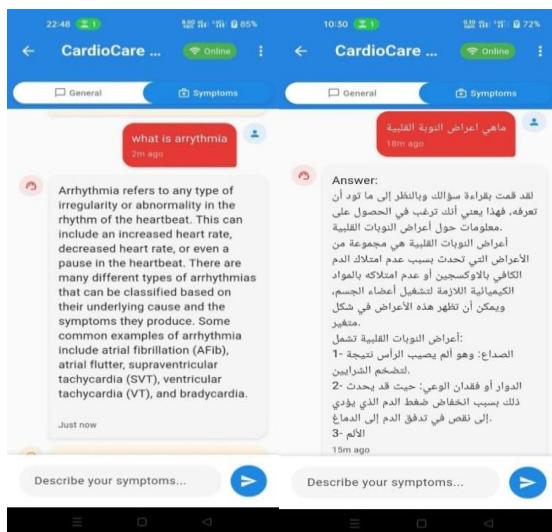


Figure 15. General Chat response

5. CONCLUSION

CVD remains one of the leading causes of death globally, with early diagnostic support and monitoring systems still presenting major challenges [20], particularly in remote or underserved areas. Our work aimed to develop an intelligent cardiology assistant that leverages real-time wearable data and integrates it with a locally hosted model to create a private,

scalable, and smart diagnostic system. The pre-trained LLaMA-3.1 8B model was fine-tuned using instruction tuning on a custom cardiology-specific dataset, developed with ethical considerations and an accompanying disclaimer. The resulting system, LLM-Cardio, was evaluated using the BERTScore metric to assess the quality of its generated diagnostic responses. As a result, our fine-tuned model outperformed traditional machine learning models such as Naïve Bayes and logistic regression in cardiovascular disease diagnostic support and risk assessment, due to their limited capacity to handle long-form medical text. Compared with BioGPT and the base LLaMA model, the fine-tuned LLM-Cardio excelled in cardiovascular disease diagnosis generation, achieving a precision of 0.9463 and thereby demonstrating its effectiveness in producing accurate and relevant cardiology diagnostics.

Despite these promising results, our work remains at the proof-of-concept stage due to the use of simulated data, the absence of clinical validation, and limited multimodality. Looking ahead, future work will focus on integrating real-world wearable data, conducting diagnostic evaluations in collaboration with healthcare professionals, and extending the system toward full multimodality, including medical image processing, in order to enhance its clinical robustness. With these enhancements, LLM-Cardio has the potential to become a valuable everyday health assistant, particularly in remote or underserved areas.

REFERENCES

- [1] Gaziano, T. A. (2022). Cardiovascular diseases worldwide. In Public Health Approach to Cardiovascular Disease Prevention & Management.
- [2] Ferrara, E. (2024). Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: A survey of early trends, datasets, and challenges. Sensors, 24(15): 5045. <https://doi.org/10.3390/s24155045>
- [3] Zong, M.Y., Hekmati, A., Guastalla, M., Li, Y.Y., Krishnamachari, B. (2025). Integrating large language models with internet of things: Applications. Discover Internet of Things, 5: 2. <https://doi.org/10.1007/s43926-024-00083-4>
- [4] Ullah, M., Hamayun, S., Wahab, A., Khan, S.U., et al. (2023). Smart technologies used as smart tools in the management of cardiovascular disease and their future perspective. Current Problems in Cardiology, 48(11): 101922. <https://doi.org/10.1016/j.cpcardiol.2023.101922>
- [5] Raza, M.Z., Xu, J.W., Lim, T., Boddy, L., Mery, C.M., Well, A., Ding, Y. (2025). LLM-TA: An LLM-enhanced thematic analysis pipeline for transcripts from parents of children with congenital heart disease. arXiv preprint arXiv:2502.01620. <https://doi.org/10.48550/arXiv.2502.01620>
- [6] Kim, Y., Xu, X.H., McDuff, D., Breazeal, C., Park, H.W. (2024). Health-LLM: Large language models for health prediction via wearable sensor data. arXiv preprint arXiv:2401.06866. <https://doi.org/10.48550/arXiv.2401.06866>
- [7] Xu, X.H., Yao, B.S., Dong, Y.Z., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A.K., Wang, D.K. (2024). Mental-LLM: Leveraging large language models

for mental health prediction via online text data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(1): 1-32. <https://doi.org/10.1145/3643540>

[8] Cosentino, J., Belyaeva, A., Liu, X., Furlotte, N.A., et al. (2024). Towards a personal health large language model. arXiv preprint arXiv:2406.06474. <https://doi.org/10.48550/arXiv.2406.06474>

[9] Ji, S.J., Zheng, X.Z., Wu, C.S. (2024). HARGPT: Are LLMs zero-shot human activity recognizers? In 2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys), Hong Kong, China, pp. 38-43. <https://doi.org/10.1109/FMSys62467.2024.00011>

[10] Healey, E., Kohane, I. (2024). LLM-CGM: A benchmark for large language model-enabled querying of continuous glucose monitoring data for conversational diabetes management. In Biocomputing 2025: Proceedings of the Pacific Symposium, pp. 82-93. https://doi.org/10.1142/9789819807024_0007

[11] Singhal, K., Tu, T., Gottweis, J., Sayres, R., et al. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, 31: 943-950. <https://doi.org/10.1038/s41591-024-03423-7>

[12] Yang, X., Chen, A., PourNejatian, N., Shin, H.C., et al. (2022). GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records. arXiv preprint arXiv:2203.03540. <https://doi.org/10.48550/arXiv.2203.03540>

[13] Thiriet, M. (2019). Cardiovascular disease: An introduction. In *Vasculopathies. Biomathematical and Biomechanical Modeling of the Circulatory and Ventilatory Systems*, pp. 1-90. https://doi.org/10.1007/978-3-319-89315-0_1

[14] Klatt, C.A. (2019). Case Files® Collection on AccessMedicine™. *Journal of Electronic Resources in Medical Libraries*, 16(1): 9-18. <https://doi.org/10.1080/15424065.2019.1596774>

[15] Luthra, A. (2014). 50 Cases in Clinical Cardiology: A Problem Solving Approach. Jaypee Brothers Medical Pub. <https://doi.org/10.5005/jp/books/12141>

[16] Mohajan, D., Mohajan, H.K. (2023). Body mass index (BMI) is a popular anthropometric tool to measure obesity among adults. *Journal of Innovations in Medical Research*, 2(4): 25-33. <https://doi.org/10.56397/JIMR/2023.04.06>

[17] Yang, R., Tan, T.F., Lu, W., Thirunavukarasu, A.J., Ting, D.S.W., Liu, N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4): 255-263. <https://doi.org/10.1002/hcs2.61>

[18] NVIDIA. (2024). Meta / llama-3.1-8b-instruct. https://docs.api.nvidia.com/nim/reference/meta-llama-3_1-8b, accessed on Nov. 28, 2025.

[19] Zhang, T.Y., 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>

[20] Franco, M., Cooper, R.S., Bilal, U., Fuster, V. (2011). Challenges and opportunities for cardiovascular disease prevention. *The American Journal of Medicine*, 124(2): 95-102. <https://doi.org/10.1016/j.amjmed.2010.08.015>