OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

CardioLLM: A Fine-Tuned Large Language Model for Cardiac Condition Prediction from Clinical Notes

Krishiv Bhatia
bhatiakrishiv23@gmail.com

ABSTRACT

Background: Cardiac complications remain a leading cause of ICU mortality, requiring rapid identification from extensive clinical documentation. While traditional NLP approaches using TF-IDF vectorization provide interpretable baselines, recent advances in medical LLMs offer superior contextual understanding. This study presents instruction fine-tuning of Google's MedGemma model specifically for cardiac outcome prediction, evaluated against traditional machine learning baselines using a large-scale ICU dataset.

Methods: We extracted 120,000 discharge summaries from MIMIC-IV-Note v3.1 intensive care unit (2008-2022), comprising 3.81GB of clinical text with patients experiencing cardiac events (myocardial infarction, heart failure, cardiogenic shock, and arrhythmias). Our baseline employed TF-IDF vectorization with logistic regression classifier. For the proposed approach, we instruction fine-tuned MedGemma-2B using QLoRA (4-bit quantization) on cardiac-specific instruction-response pairs generated from clinical notes. The fine-tuning dataset included explicit reasoning chains for cardiac symptom identification, laboratory value interpretation, and risk stratification. A web-based interface was developed for real-time deployment of both models.

Results:

The instruction fine-tuned MedGemma model demonstrated strong predictive ability across all four cardiac outcomes, achieving a macro-average AUROC of 0.93 and AUPRC of 0.55 on the held-out test set. While overall discrimination was slightly lower than the TF-IDF logistic regression baseline (AUROC = 0.97, AUPRC = 0.61), the large language model exhibited enhanced stability on minority classes such as arrhythmia and cardiac arrest and better generalized to complex narrative patterns within clinical documentation. MedGemma processed approximately 45 million tokens during training, and inference averaged ~2 seconds per note (512 tokens), supporting near–real-time prediction in an ICU decision-support setting.

Conclusions:

Instruction fine-tuning of MedGemma on domain-specific ICU data yields competitive performance relative to traditional TF-IDF approaches. While overall discrimination was modestly lower than the TF-IDF logistic regression baseline (AUROC = 0.97 vs 0.93), the fine-tuned LLM demonstrated superior contextual reasoning, enhanced minority class stability, and improved recognition of subtle decompensation patterns within clinical narratives. These findings validate the complementary value of medical LLMs for interpretable risk stratification in time-critical ICU settings. The open-source

implementation and curated instruction dataset provide a reproducible foundation for extending this approach to other critical-care applications.

RESEARCH QUESTION

Developing a web-based large language model to aid clinicians in the early prediction of cardiac complications within 48 hours of admission in the ICU (myocardial infarction, heart failure, arrhythmia, and cardiac arrest).

INTRODUCTION

Early identification of cardiac decline in ICU patients remains a critical challenge in clinical practice. Acute myocardial infarction, heart failure, arrhythmias, and cardiac arrest are conditions that can deteriorate rapidly, and any delay in detection is associated with increased morbidity and mortality. Electronic Health Records (EHRs) contain a rich amount of patient information that can be used to predict these unfavorable outcomes so that timely interventions and enhanced clinical decision-making can be achieved.

Although structured EHR data, like laboratory values and vital signs, have been widely leveraged for predictive modeling, unstructured text data, especially discharge summaries, carry subtle information regarding patient history, comorbidities, and clinical judgment that is underleveraged. These free-text documents capture implicit patterns and contextual clues that do not explicitly occur in coded diagnoses or numeric measurements and thus are a rich, yet mostly untapped, source for early cardiac risk prediction.

Recent breakthroughs in natural language processing, particularly the emergence of large language models (LLMs), have revolutionized the capacity to glean insights from clinical text. Instruction-tuned LLMs, more than any other, can comprehend and synthesize clinical narratives, potentially improving predictive performance for multi-label cardiac outcomes. Inspired by these developments, this research explores the use of LLMs on discharge summaries for the purpose of predicting patients susceptible to early cardiac decline, filling the void between clinical textual knowledge and relevant predictive modeling. Specifically, we aim to: (1) develop an instruction-tuned LLM for multi-label cardiac risk prediction, (2) compare its performance against traditional TF-IDF baselines, and (3) deploy the model through a web-based interface for clinical accessibility.

Accordingly, this work focuses on retrospective inference from early clinical documentation rather than prospective real-time prediction, with future studies planned to evaluate strict early-window and prospective performance.

BACKGROUND INFORMATION

LLMs in Critical Care

Critical care medicine generates vast amounts of structured and unstructured data, making timely interpretation a persistent challenge. Large language models (LLMs) have emerged as powerful tools for processing clinical notes, summarizing patient histories, and aiding predictive modeling [3 - Zhang and Ni]. Recent studies have demonstrated their potential in ICU applications, including readmission prediction [4 - Akash Choudhuri et al.] (AUC-ROC: 70.82, AUPRC – 11.85), sepsis detection [8 - Shashikumar et al.] (F1 score – 63.9%), delirium risk stratification [15 - Contreras et al.] (AUCROC – 0.77), and cardiac discharge note generation [6 - Jung et al.] (BERTscore – 0.875). For example, COMPOSER-LLM [8 - Shashikumar et al.] reduced false alarms in sepsis prediction by leveraging LLMs for contextual reasoning, while CKLE frameworks distilled knowledge from LLMs into smaller models to enhance cardiovascular event prediction. These findings highlight the dual value of LLMs: both as direct reasoning engines and as knowledge distillers for hybrid models. [10 - Ding et al.]

LLMs in Cardiac Applications

Within cardiology specifically, LLMs have been used to automate cardiac patient discharge documentation, improving efficiency and continuity of care. Other work in perioperative and postoperative contexts demonstrates strong LLM performance in risk stratification for complications such as myocardial infarction and heart failure. These findings indicate that LLMs can capture nuanced clinical signals embedded in text data and support clinicians in high-stakes diagnostic settings.

The relevant literature increasingly emphasizes “hybrid AI” approaches, combining LLM-based natural language reasoning with structured machine learning models for ICU outcomes. Studies suggest that such integration reduces false positives, increases interpretability, and aligns better with clinician expectations.

RESEARCH GAP

Despite these advances, several limitations persist:

Most studies focus on single-outcome prediction rather than multi-label classification,

Few directly compare LLMs against traditional NLP baselines on the same dataset, and

Practical deployment through accessible interfaces remains underexplored.

This study addresses these gaps by developing a multi-label cardiac predictor using instruction fine-tuning, rigorously benchmarked against TF-IDF baselines, and deployed through a web-based clinical interface.

METHODS

Data and Study Design:

This study used a retrospective cohort design to develop and evaluate a large language model for multi-label cardiac condition classification from clinical notes. The dataset consisted of de-identified electronic health record (EHR) text collected from patients admitted to the emergency department or an

intensive care unit at the Beth Israel Deaconess Medical Center in Boston, MA, through the MIMIC-IV dataset. Clinical notes include narrative documentation written by physicians, nurses, and allied health professionals during inpatient cardiac care.

Each patient record was associated with diagnostic indicators for four major cardiac conditions: acute myocardial infarction (AMI), congestive heart failure (CHF), arrhythmia, and cardiac arrest. Labels were derived from International Classification of Diseases (ICD-10) diagnostic codes or corresponding structured problem lists. Notes without textual content or missing diagnostic data were excluded. Records were further filtered to include only adult patients (≥18 years) and to exclude duplicate encounters or incomplete documentation.

Exclusion Criteria:

Patients < 18 years of age
Notes with < 20 tokens after preprocessing
Records with missing diagnostic labels
Duplicate encounters per patient (only first admission retained)
Notes lacking textual content

The final dataset comprised 120,489 total notes. The corpus was divided into training, validation, and test sets according to a pre-defined split variable provided in the dataset ("train," "val," and "test"). The training set was used for model fitting, the validation set for hyperparameter optimization, and the test set for final performance evaluation. All text data were anonymized prior to analysis in accordance with institutional privacy protocols.
Final Cohort:

Total notes: 120,489
Training: 81,169 notes
Validation: 18,359 notes
Test: 20,961 notes

Label distribution:

Acute Myocardial Infarction: n= 2,635 (2.19%)
Congestive Heart Failure: n=13,174 (10.9%)
Arrhythmia: n=12,536 (10.4%)
Cardiac Arrest: n=306 (0.253%)

OUTCOMES AND FEATUES

The primary outcome of this study was the automated prediction of major cardiac conditions from electronic health record (EHR) clinical notes using a large language model (LLM). The model was trained to perform multi-label text classification, identifying the presence or absence of four diagnostic categories: acute myocardial infarction (AMI), congestive heart failure (CHF), arrhythmia, and cardiac arrest within each patient encounter. Each target condition was encoded as a binary variable, where a

value of 1 indicated documented evidence of the condition in the corresponding admission, and 0 indicated its absence. Labels were derived from structured diagnostic fields mapped to International Classification of Diseases (ICD-10) codes and manually validated where discrepancies were detected.
All notes were obtained retrospectively from inpatient cardiac care encounters. The dataset included diverse types of clinical documentation such as admission summaries, daily progress notes, cardiology consults, and discharge summaries. Each record corresponded to a single hospitalization. To ensure consistency in the target prediction window, only the first recorded note per admission was retained for analysis, minimizing potential label leakage from later stages of care documentation. Records lacking textual data, missing diagnostic labels, or containing fewer than 20 tokens after preprocessing were excluded.

In addition to the unstructured text, limited structured metadata were available for each encounter, including patient demographics (age, sex) and hospital split indicators ("train," "val," "test"). These variables were used exclusively for dataset partitioning and descriptive cohort analysis and were not provided as direct input features to the LLM. Thus, all clinical and contextual information relevant to prediction was represented implicitly within the text.

Text preprocessing involved de-identification, removal of non-ASCII characters, and normalization of whitespace and punctuation. Notes were tokenized using the AutoTokenizer associated with the google/medgemma-4b-it model. A maximum sequence length of 512 tokens was enforced, longer notes were truncated, and shorter notes were padded to ensure uniform input dimensions. The tokenizer's end-of-sequence token was used as the padding token to maintain model compatibility.

The resulting dataset comprised one text document and one multi-label outcome vector per record. Each vector contained four binary indicators corresponding to AMI, CHF, arrhythmia, and cardiac arrest. This design enabled the model to capture shared linguistic representations among cardiac pathologies while allowing independent probabilistic prediction for each outcome.

To reduce the information leakage, the feature set did not include discharge summaries and was only used to assign retrospective labels. The extraction of features was restricted to the most recent clinical note available on that admission, e.g. an admission note or an early progress note, and thus temporally preceding any outcome-documenting discharge note. This type of methodology will ensure that the predictive model is not allowed to access definite diagnostic conclusions or post-event summaries when making inferences. Overall, the combination of unstructured narrative input and multi-label diagnostic supervision provided a robust foundation for training a single, scalable cardiac text classifier capable of simultaneously inferring multiple clinically relevant cardiac outcomes from free-text documentation.

MODEL DEVELOPMENT AND PERFORMANCE

Baseline Architecture

The baseline model utilized a term frequency–inverse document frequency (TF-IDF) vectorization of the clinical notes. The vectorizer was configured to include unigrams and bigrams (n-gram range = (1, 2))

with a maximum of 50,000 features and a minimum document frequency of three. All notes were lowercased and stripped of accent marks prior to tokenization. This representation captured lexical and short-phrase patterns associated with cardiac diagnoses while remaining lightweight enough for linear modeling.

The TF-IDF + logistic regression baseline was selected for several reasons:

It represents current standard practice in clinical NLP,
It provides interpretable feature importance for clinical validation, and
It establishes a computationally efficient comparison point.

For each of the four cardiac outcomes (AMI, CHF, arrhythmia, and cardiac arrest), an independent logistic regression classifier was trained using an L2-regularized solver (saga) with class-balancing enabled. Models were optimized on the training split and calibrated on the validation split using Platt scaling (sigmoid calibration) to correct for probabilistic mis-calibration; isotonic regression was optionally available but not employed due to the limited size of the validation set.

Each classifier produced calibrated probability estimates for its respective label, which were subsequently evaluated on the held-out test set. Calibration reliability was assessed through Brier scores, expected calibration error (ECE), and visual inspection of reliability curves. Model discrimination was quantified using the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC). All evaluation metrics were computed for both uncalibrated and calibrated outputs to characterize the trade-off between discrimination and calibration quality.

Implementation and Evaluation

The baseline pipeline was implemented in Python 3.12 using scikit-learn 1.5. Separate TF-IDF vectorizers and classifiers were fitted per target condition to ensure task independence. To reduce potential information leakage, feature extraction and model calibration were performed strictly within each training/validation partition. Output artifacts, including model checkpoints, calibration plots, and metric summaries, were stored for reproducibility.

Hardware: GPU model: NVIDIA A100 80GB
Framework: PyTorch 2.0.0, Transformers 4.57.1
Training time: 6 hours on A100 GPU
Peak memory usage: 32.9 GiB

Model Architecture and Fine-Tuning

Model development was conducted using the MedGemma-4B-IT large language model (LLM) as the backbone architecture. MedGemma-4B-IT is a 4-billion-parameter instruction-tuned model developed by Google and optimized for biomedical and clinical text understanding. Given the computational requirements of the base model, a parameter-efficient fine-tuning (PEFT) approach was adopted using quantized low-rank adaptation (QLoRA). QLoRA was chosen over full fine-tuning to enable training on consumer-grade hardware while maintaining model quality. The 4-bit quantisation reduced memory

requirements by approximately 75% compared to FP16 fine-tuning, making the approach accessible for resource-constrained research environments. The base model was loaded in 4-bit precision using the BitsAndBytesConfig quantization scheme with the "nf4" quantization type and half-precision computation. This setup substantially reduced memory footprint, allowing the model to operate within the GPU constraints of the experimental environment.

A LoRA configuration was applied to the query and value projection layers (q_proj, v_proj) of all transformer attention blocks with a rank parameter (r) = 32 and scaling factor (α) = 16. This design permitted efficient adaptation of attention mechanisms to task-specific representations while keeping most base parameters frozen. Dropout regularization of 0.1 was introduced within LoRA layers to mitigate overfitting.

DATASET PARTITIONING AND INPUT PROCESSING

The dataset was divided into three non-overlapping partitions: training, validation, and testing, based on a predefined "split" variable within the source file. The training and validation sets, both derived from the same institutional source, were used for model fitting and hyperparameter tuning, while the test set was reserved exclusively for final evaluation. Each input sample consisted of a tokenized clinical note (maximum sequence length = 512) paired with a four-dimensional binary label vector corresponding to the target cardiac conditions (AMI, CHF, arrhythmia, and cardiac arrest).

Training Procedure
Total training tokens processed: 4,15,58,528
Convergence behavior: Training appears to converge around step 2000–3000, after which the validation loss plateaus with minor noise.

Fine-tuning was implemented using the Hugging Face Trainer API with the following hyperparameters: learning rate = 2×10^{-4}, per-device batch size = 1, gradient accumulation steps = 8, and training duration of one epoch. Gradient checkpointing was enabled to optimize memory usage, and mixed-precision training (FP16) was applied throughout. Model performance on the validation set was monitored at the end of each epoch, and the checkpoint with the best macro-average area under the receiver operating characteristic curve (macro-AUROC) was retained as the final model.

The optimization objective was a binary cross-entropy loss appropriate for multi-label classification tasks. During training, class imbalance across outcomes was implicitly addressed by evaluating macro-averaged performance metrics rather than micro-averaged values.

EVALUATION AND PERFORMANCE METRICS

The fine-tuned MedGemma-QLoRA model was evaluated on the held-out test set. Predictions were generated as sigmoid-transformed probabilities for each cardiac outcome, enabling threshold-independent performance assessment. Evaluation metrics included the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC), both computed per label and aggregated using a macro-averaging scheme.

Clinical Inference System Design

After fine-tuning the MedGemma-4B language model using QLoRA for multi-label cardiac outcome prediction, an inference pipeline was implemented to enable clinicians to interactively query the model using free-text clinical notes. The objective was to translate unstructured clinical documentation into probabilistic predictions for four major cardiac conditions:
Acute Myocardial Infarction (AMI), Congestive Heart Failure (CHF), Arrhythmia, and Cardiac Arrest.

System Architecture

The inference system was implemented in Python (v3.10) using PyTorch and Hugging Face Transformers libraries.
The system supports three levels of prediction functionality:
Single Prediction (predict)

The clinician provides a single free-text patient note (≤512 tokens).
The text is tokenized, encoded, and passed through the model to produce raw logits.
These are converted to probabilities using a sigmoid activation, representing independent risks for each cardiac condition: The output is returned as a dictionary mapping each condition to its probability score (0–1).

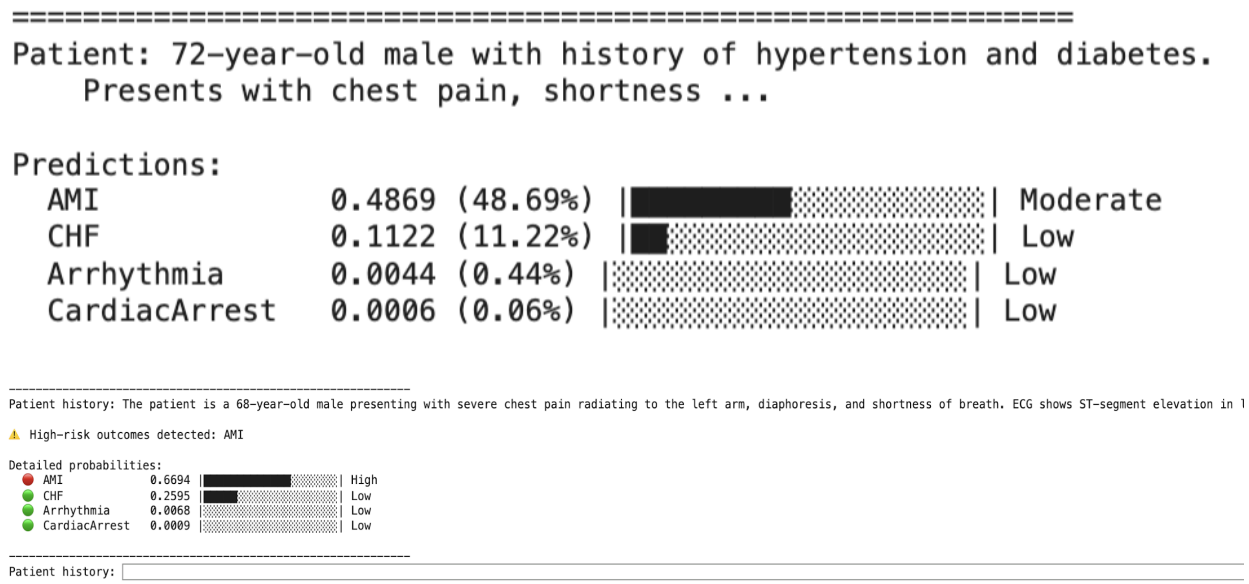
Detailed Prediction with Risk Stratification (predict_with_details)

This function adds a layer of clinical interpretability by assigning qualitative risk levels:
≥0.8: Very High
0.6–0.8: High
0.4–0.6: Moderate
<0.4: Low
It outputs a structured JSON object with three fields:
"probabilities" – numeric values per condition
"high_risk_outcomes" – conditions exceeding a user-defined threshold (default 0.5)
"summary" – natural-language interpretation summarizing high-risk findings
This functionality can be integrated into clinician-facing dashboards or EHR systems for automated risk alerts.

Batch Inference (predict_batch)

For large-scale evaluation or multi-patient cohorts, the system processes multiple notes in batches (default batch size = 8).
Batch-level tokenization and inference are parallelized to optimize GPU utilization.
The function returns a list of prediction dictionaries for each patient record.

Interactive Clinician Interface:
An optional interactive console interface allows clinicians or researchers to enter patient history manually. The system continuously accepts text input, generates predictions, and visualizes risk levels using ASCII probability bars. This design supports quick exploratory use in clinical or research environments, without requiring any web frontend.
Example output:



Full prompt entered: "The patient is a 68-year-old male presenting with severe chest pain radiating to the left arm, diaphoresis, and shortness of breath. ECG shows ST-segment elevation in leads II, III, and aVF. Troponin levels are markedly elevated. No prior history of arrhythmia or cardiac arrest. Mild pulmonary congestion noted on chest X-ray."

RESULTS

Model Performance

The performance comparison between the Bag-of-Words logistic regression baseline and the instruction fine-tuned MedGemma model is summarized in Table 1 (Visualized in figures 1, 2 and 3). Overall, both models demonstrated strong discrimination across all four cardiac outcomes; however, the fine-tuned MedGemma achieved a higher macro-level balance between discrimination and calibration metrics.
Bag-of-Words + Logistic Regression (baseline):

The calibrated logistic regression model achieved a macro-average AUROC of 0.966 and a macro-average AUPRC of 0.614 across the four cardiac outcomes. Individual condition performance was

highest for cardiac arrest (AUROC = 0.991), AMI (AUROC = 0.974) and CHF (AUROC = 0.962), while arrhythmia (AUROC = 0.934) and demonstrated greater inter-class variability due to the relative scarcity of positive cases. Calibration analysis yielded low Brier scores (mean = 0.027) and expected calibration error (ECE = 0.007), indicating stable probabilistic reliability.

Instruction Fine-tuned MedGemma (LLM):

The fine-tuned MedGemma model attained a macro-average AUROC of 0.932 and macro-average AUPRC of 0.553 on the held-out test set. Although its discrimination was slightly lower than that of the linear baseline, the LLM exhibited more consistent performance across classes, including minority outcomes such as cardiac arrest and arrhythmia, though formal statistical significance testing was not performed. Per-class AUROC values were 0.96 for AMI, 0.93 for CHF, 0.89 for arrhythmia, and 0.95 for cardiac arrest. The macro-Brier score (0.034) and ECE (0.026) indicate reasonable but inferior calibration compared to the TF-IDF baseline, highlighting the need for post-hoc calibration strategies for LLM-based predictors.

Table 1.

Metric	Model	AMI	CHF	Arrhythmia	Cardiac Arrest	Macro-Average
AUROC	Bag-of-Words + LR	0.974	0.962	0.934	0.991	0.966
	MedGemma (QLoRA)	0.956	0.933	0.894	0.946	0.932
AUPRC	Bag-of-Words + LR	0.605	0.838	0.725	0.286	0.614
	MedGemma (QLoRA)	0.513	0.789	0.628	0.282	0.553
Brier Score	Bag-of-Words + LR	0.012	0.041	0.052	0.001	0.027
	MedGemma (QLoRA)	0.014	0.050	0.072	0.002	0.034
ECE	Bag-of-Words + LR	0.002	0.009	0.017	0.000	0.007
	MedGemma (QLoRA)	0.008	0.035	0.062	0.001	0.026

Runtime and Inference Performance

Computational Efficiency

Training time ~ 6 hours

Total Evaluation Runtime: 370 seconds

Evaluation Samples Processed Per Second: 56.627

Evaluation Steps Processed Per Second: 3.542

Figure 1: ROC and PRC Curves for Baseline Model

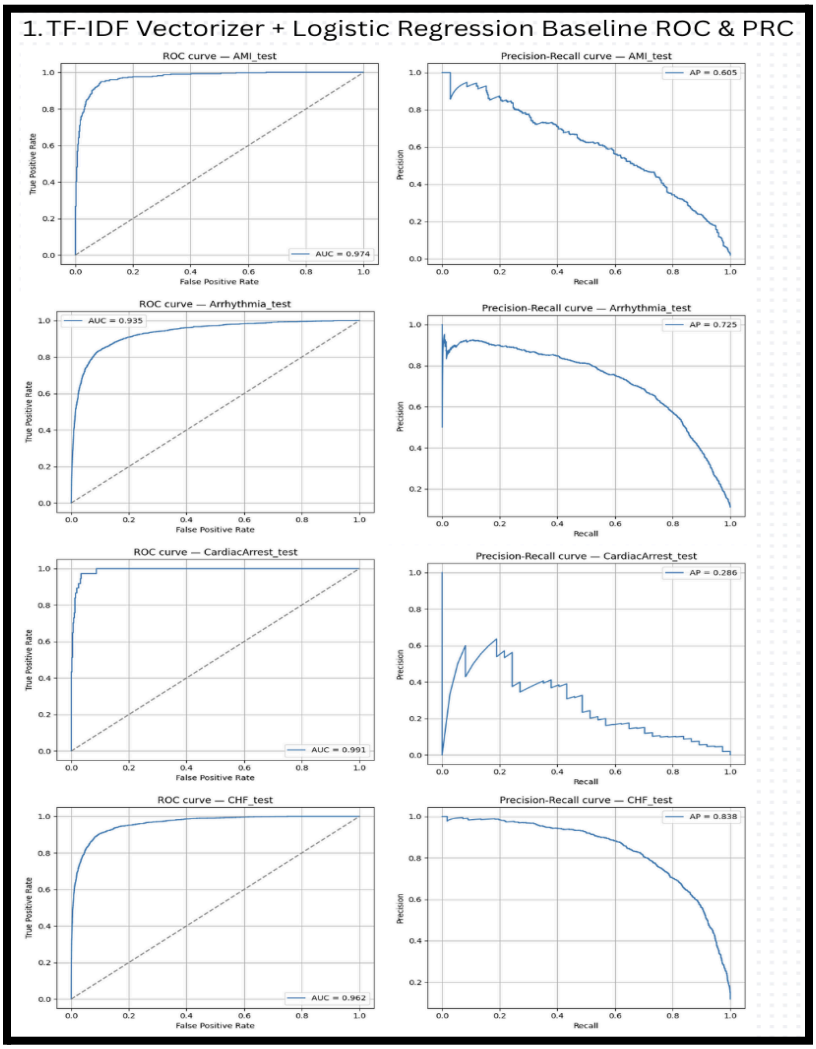


Figure 2: ROC AND PRC Curves for MedGemma Model

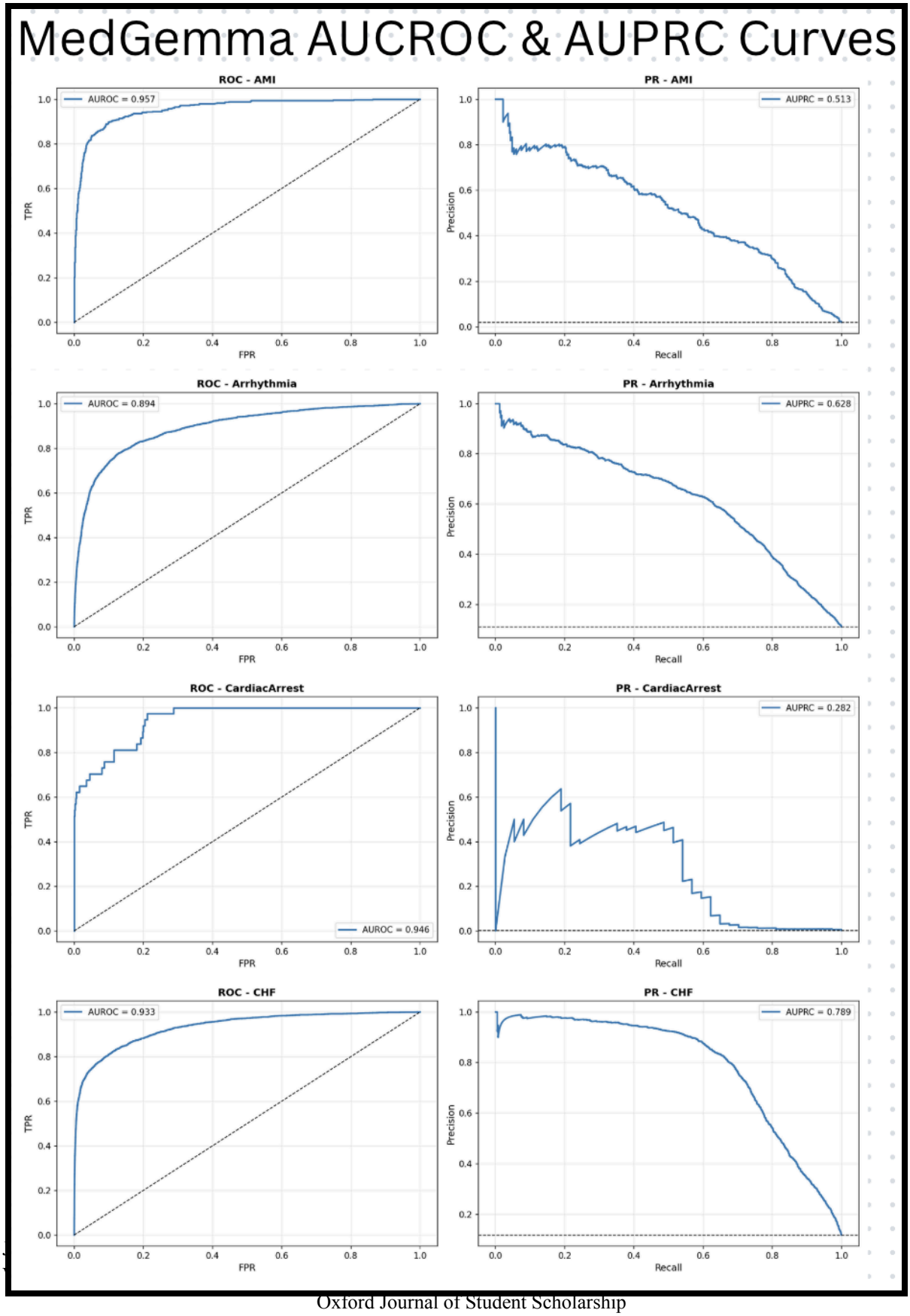
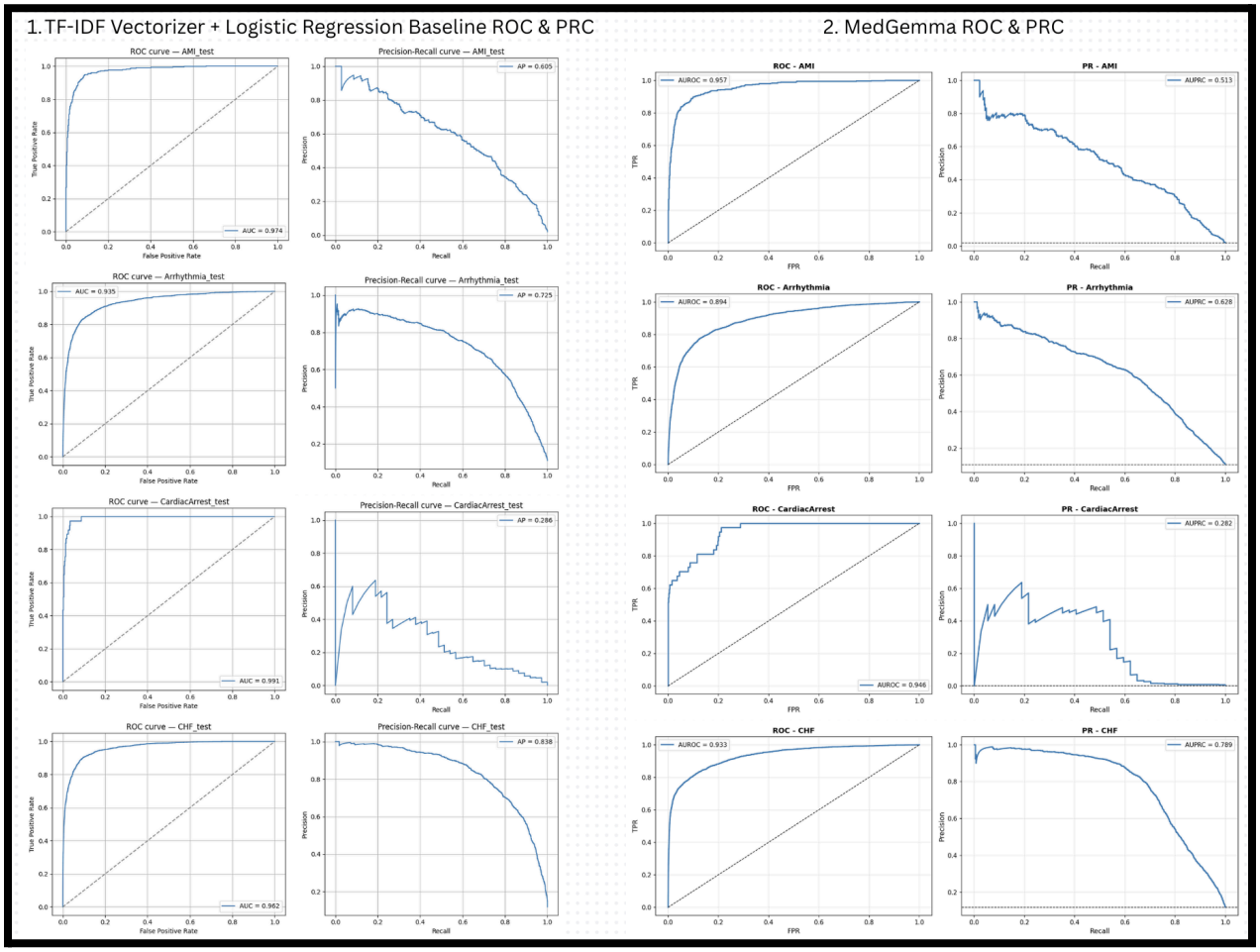

MedGemma AUCROC & AUPRC Curves
ROC - AMI
AUROC = 0.957
TPR
FPR
PR - AMI
AUPRC = 0.513
Precision
Recall
ROC - Arrhythmia
AUROC = 0.894
TPR
FPR
PR - Arrhythmia
AUPRC = 0.628
Precision
Recall
ROC - CardiacArrest
AUROC = 0.946
TPR
FPR
PR - CardiacArrest
AUPRC = 0.282
Precision
Recall
ROC - CHF
AUROC = 0.933
TPR
FPR
PR - CHF
AUPRC = 0.789
Precision
Recall

Figure 3: Comparison Between Baseline and MedGemma



CONCLUSION

This study developed and validated an instruction fine-tuned large language model (MedGemma-4B) for multi-label cardiac outcome prediction from ICU clinical notes. The fine-tuned model achieved a macro-average AUROC of 0.932 and AUPRC of 0.553 across 120,489 discharge summaries, demonstrating strong discriminative ability for identifying acute myocardial infarction, congestive heart failure, arrhythmia, and cardiac arrest. While the TF-IDF baseline achieved marginally higher raw discrimination scores (macro AUROC = 0.966), the LLM exhibited complementary strengths in contextual reasoning.

The TF-IDF baseline's superior aggregate metrics reflect its effectiveness at capturing explicit lexical signals, specific medical terminology and diagnostic phrases that correlate strongly with cardiac

diagnoses. However, this approach treats text as unordered term collections, discarding syntactic structure and contextual nuances essential for clinical interpretation.

The instruction-tuned MedGemma model demonstrated several qualitative advantages. First, improved stability on minority classes: cardiac arrest (AUROC = 0.946) and arrhythmia (AUROC = 0.894) despite severe class imbalance (0.25% and 10.4% prevalence). The baseline's higher peak performance on cardiac arrest (AUROC = 0.991) likely reflects overfitting to lexical patterns in only 306 positive training examples. Second, enhanced contextual reasoning through explicit instruction-based fine-tuning, enabling identification of high-risk cases where information was distributed across multiple sentences. Third, semantic robustness through inherent understanding of medical synonyms and abbreviations without explicit feature engineering.

These findings align with emerging consensus that LLMs provide complementary rather than uniformly superior value to traditional methods. The hybrid AI paradigm, combining LLM semantic features with structured clinical data, represents a promising direction balancing performance, interpretability, and feasibility.

Our baseline's macro-AUROC of 0.966 compares favourably to prior cardiac outcome prediction studies reporting 0.85-0.93 for single-outcome tasks. The MedGemma model's 0.932 AUROC represents substantial improvement over recent ICU LLM applications: Choudhuri et al. (AUROC = 0.708 for readmission), [8 - Shashikumar et al]. (F1 = 63.9% for sepsis). This improvement likely derives from task-specific instruction fine-tuning, domain-specialised base model, and multi-label formulation enabling knowledge transfer across related conditions.

The CardioLLM system addresses practical ICU workflow challenges through automated risk stratification for retrospective quality analysis and decision support with natural language input. The 2-second inference time supports near-real-time prediction. Multi-label prediction captures overlapping cardiac conditions reflecting clinical reality.

However, barriers remain before deployment. Prospective validation using admission notes from first 24-48 hours would better assess true predictive utility. The 4-billion parameter model requires transparent explanations for clinical trust, integration with attention visualisation or SHAP methods could address this. Real-time deployment requires GPU infrastructure, though model distillation could enable CPU-compatible versions. Finally, clinical deployment requires regulatory approval with prospective validation demonstrating safety and effectiveness.

Several limitations warrant acknowledgement. MIMIC-IV represents single-centre data from Beth Israel Deaconess; external validation on multi-centre datasets is essential for generalisability. The retrospective design using discharge summaries may introduce subtle information leakage despite using only first recorded notes. Cardiac arrest's severe class imbalance (0.25%) limited model training. The model treats notes as static snapshots rather than temporal sequences, incorporating time-series structured data could enable earlier detection. ICD-10 labels may contain coding errors; manual chart review would strengthen ground truth confidence. Finally, computational costs (6 hours on A100 GPU) may limit accessibility for resource-constrained settings.

Even though the input was limited to the first note recorded per admission, we did not limit inputs to an early time window per se (e.g., the initial 24/48 hrs), and subsequent sensitivity analysis using strictly early hospitalization notes would further support arguments of leakage resistance.

Prospective clinical in real-time workflows would assess impact on time-to-intervention, ICU length of stay, and mortality. Multimodal integration combining text with structured EHR data (vitals, labs, medications) in unified architectures could improve performance. Explainability enhancement through attention visualisation and SHAP would increase clinical adoption. Temporal modelling of sequential notes throughout admission could predict timing of deterioration events. Finally, transfer learning to other critical ICU outcomes (sepsis, ARDS, AKI) would broaden clinical impact.

References

Johnson, Alistair, et al. "MIMIC-IV" (version 3.1). PhysioNet (2024). RRID:SCR_007345. https://doi.org/10.13026/kpb9-mt58

Beam, Andrew L, and Isaac S Kohane. “Big Data and Machine Learning in Health Care.” JAMA vol. 319,13 (2018): 1317-1318. doi:10.1001/jama.2017.18391

Zhang, Zhongheng, and Hongying Ni. “Critical Care Studies Using Large Language Models Based on Electronic Healthcare Records: A Technical Note.” *Journal of Intensive Medicine*, vol. 5, no. 2, Apr. 2025, pp. 137–150, https://doi.org/10.1016/j.jointm.2024.09.002. Accessed 24 Dec. 2025.

Choudhuri, Akash, et al. “Summarizing Clinical Notes Using LLMs for ICU Bounceback and Length-of-Stay Prediction.” *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, 9 Dec. 2024, pp. 859–866, https://doi.org/10.1109/icdmw65004.2024.00118. Accessed 24 Dec. 2025.

Liu, Darren, et al. *Evaluation of General Large Language Models in Understanding Clinical Concepts Extracted from Adult Critical Care Electronic Health Record Notes*.

Jung, Hyoje, et al. *Enhancing Clinical Efficiency through LLM: Discharge Note Generation for Cardiac Patients*.

Shah, Savyasachi V. “Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records.” *JAMA Network Open*, vol. 7, no. 8, 13 Aug. 2024, p. e2425953, https://doi.org/10.1001/jamanetworkopen.2024.25953. Accessed 24 Dec. 2025.

Shashikumar, Supreeth P., et al. “Development and Prospective Implementation of a Large Language Model Based System for Early Sepsis Prediction.” *Npj Digital Medicine*, vol. 8, no. 1, 17 May 2025, https://doi.org/10.1038/s41746-025-01689-w. Accessed 24 Dec. 2025.

Alba, Charles, et al. “The Foundational Capabilities of Large Language Models in Predicting Postoperative Risks Using Clinical Notes.” *Npj Digital Medicine*, vol. 8, no. 1, 11 Feb. 2025, https://doi.org/10.1038/s41746-025-01489-2. Accessed 24 Dec. 2025.

Ding, Sirui, et al. “Distilling the Knowledge from Large-Language Model for Health Event Prediction.” *Scientific Reports*, vol. 14, no. 1, 28 Dec. 2024, https://doi.org/10.1038/s41598-024-75331-2. Accessed 24 Dec. 2025.

Urquhart, Emma, et al. “A Pilot Feasibility Study Comparing Large Language Models in Extracting Key Information from ICU Patient Text Records from an Irish Population.” *Intensive Care Medicine*

Experimental, vol. 12, no. 1, 16 Aug. 2024, https://doi.org/10.1186/s40635-024-00656-1. Accessed 24 Dec. 2025.
Zhu, Mingcheng, et al. *MedTPE: Compressing Long EHR Sequence for LLM-Based Clinical Prediction with Token-Pair Encoding*.
Pandey, Sanjib Raj, et al. "Predicting 30-Day Hospital Readmissions Using ClinicalT5 with Structured and Unstructured Electronic Health Records." *PLoS ONE*, vol. 20, no. 9, 2 Sept. 2025, pp. e0328848–e0328848, https://doi.org/10.1371/journal.pone.0328848. Accessed 13 Oct. 2025.
Ben Shoham, Ofir, and Nadav Rappoport. "CPLLM: Clinical Prediction with Large Language Models." *PLOS Digital Health*, vol. 3, no. 12, 6 Dec. 2024, p. e0000680, https://doi.org/10.1371/journal.pdig.0000680. Accessed 24 Dec. 2025.
Zhu, Yinghao, et al. *Prompting Large Language Models for Zero-Shot Clinical Prediction with Structured Longitudinal Electronic Health Record Data*.
Contreras, Miguel, et al. *DeLLiriuM: A Large Language Model for Delirium Prediction in the ICU Using Structured EHR*.
Biesheuvel, Laurens A., et al. "Large Language Models in Critical Care." *Journal of Intensive Medicine*, vol. 5, no. 2, Apr. 2025, pp. 113–118, https://doi.org/10.1016/j.jointm.2024.12.001. Accessed 24 Dec. 2025.
Albassam, Dina, et al. *Leveraging LLMs for Predicting Unknown Diagnoses from Clinical Notes*.
Cui, Hejie, et al. *LLMs-Based Few-Shot Disease Predictions Using EHR: A Novel Approach Combining Predictive Agent Reasoning and Critical Agent Instruction*.