Automated Transformation of Unstructured Cardiovascular Diagnostic Reports into Structured Datasets Using Sequentially Deployed Large Language Models

Sumukh Vasisht Shankar MS[1], Lovedeep S Dhingra MBBS[1], Arya Aminorroaya MD, MPH[1], Philip Adejumo BS[1], Girish N Nadkarni MD MPH[2], Hua Xu PhD[5], Cynthia Brandt MD, MPH[5], Evangelos K Oikonomou MD, DPhil[1], Aline F Pedroso PhD[1], Rohan Khera MD, MS[1,3,4,5]

[1] Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA
[2] The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[3] Section of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA
[4] Center for Outcomes Research and Evaluation (CORE), Yale New Haven Hospital, New Haven, CT, USA
[5] Section of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT

Abstract Word Count: 352 words

Word Count: 2379 words

Figures/Tables: 2 Figures / 6 Tables

Keywords: Natural Language Processing, Large Language Models, Echocardiograph, Machine Learning

Correspondence to:
Rohan Khera, MD, MS
195 Church Street, 6th Floor, New Haven, CT 06510
rohan.khera@yale.edu

ABSTRACT

Background: Rich data in cardiovascular diagnostic testing are often sequestered in unstructured reports, with the necessity of manual abstraction limiting their use in real-time applications in patient care and research.

Methods: We developed a two-step process that sequentially deploys generative and interpretative large language models (LLMs; Llama2 70b and Llama2 13b). Using a Llama2 70b model, we generated varying formats of transthoracic echocardiogram (TTE) reports from 3,000 real-world echo reports with paired structured elements, leveraging temporal changes in reporting formats to define the variations. Subsequently, we fine-tuned Llama2 13b using sequentially larger batches of generated echo reports as inputs, to extract data from free-text narratives across 18 clinically relevant echocardiographic fields. This was set up as a prompt-based supervised training task. We evaluated the fine-tuned Llama2 13b model, HeartDx-LM, on several distinct echocardiographic datasets: (i) reports across the different time periods and formats at Yale New Haven Health System (YNHHS), (ii) the Medical Information Mart for Intensive Care (MIMIC) III dataset, and (iii) the MIMIC IV dataset. We used the accuracy of extracted fields and Cohen's Kappa as the metrics and have publicly released the HeartDX-LM model.

Results: The HeartDX-LM model was trained on randomly selected 2,000 synthetic echo reports with varying formats and paired structured labels, with a wide range of clinical findings. We identified a lower threshold of 500 annotated reports required for fine-tuning Llama2 13b to achieve stable and consistent performance. At YNHHS, the HeartDx-LM model accurately extracted 69,144 out of 70,032 values (98.7%) across 18 clinical fields from unstructured reports in the test set from contemporary records where paired structured data were also available. In

older echo reports where only unstructured reports were available, the model achieved 87.1% accuracy against expert annotations for the same 18 fields for a random sample of 100 reports. Similarly, in expert-annotated external validation sets from MIMIC-IV and MIMIC-III, HeartDx-LM correctly extracted 201 out of 220 available values (91.3%) and 615 out of 707 available values (87.9%), respectively, from 100 randomly chosen and expert annotated echo reports from each set.

Conclusion: We developed a novel method using paired large and moderate-sized LLMs to automate the extraction of unstructured echocardiographic reports into tabular datasets. Our approach represents a scalable strategy that transforms unstructured reports into computable elements that can be leveraged to improve cardiovascular care quality and enable research.

INTRODUCTION

Electronic health records (EHR) offer invaluable insights into optimizing cardiovascular care and driving healthcare research.[1–3] In the EHR, data streams that are most amenable to scalable applications include those available as structured tabular data. Therefore, despite their critical role in defining disease conditions, diagnostic testing such as imaging is often available only as unstructured free-text narratives and remains underutilized in disease phenotyping.[4] This gap underscores the pressing need for novel strategies to transform unstructured data into structured data elements, thus enhancing the impact and scalability of health applications that can leverage these rich data.

Prior work to transform unstructured into structured data has primarily focused on extracting isolated data elements,[5–9] with the need to develop pipelines to extract as more data streams are needed. The emergence of large language models (LLMs) as foundation models for language processing has demonstrated impressive properties for parsing text with limited domain-specific development but are limited by the high computational requirements associated with their deployment.[10] On the other hand, the scarcity of annotated unstructured-structured data limits the development of computationally efficient models. Consequently, there is a critical unmet need for novel approaches capable of efficiently transforming clinical notes into tabular data.

To address this, we propose a domain-specific and computationally efficient approach leveraging sequentially deployed LLMs, where we use a larger open-source model to generate synthetic training examples for fine-tuning a smaller model, which enables the development of a generalizable tool for converting imaging reports to tabular data. We use reports of transthoracic echocardiograms (TTEs) as the use case for this application.

METHODS

The study was reviewed by the Yale Institutional Review Board, which waived the need for informed consent, as it represents a secondary analysis of existing data.

Study Overview

We developed and fine-tuned a lightweight language model, HeartDx-LM, to extract clinically relevant diagnostic information from unstructured TTE reports. We trained the model using different text structures leveraging temporal variations in the free-text narratives of the reports to introduce this variation. The process involved generating synthetic reports using reports in a single format where all information was also available as tabular data. These reports were regenerated into different formats using examples from those formats as prompts to a Llama2 70-billion-parameter model. These synthetically adapted echo reports were then used to fine-tune a moderate-sized Llama2 13-billion-parameter model to extract a comprehensive set of quantitative, semi-quantitative, and qualitative diagnostic information from unstructured clinical reports (**Figure 1**).

Data Sources

We used data from the Yale New Haven Health System (YNHHS) EHR, a large academic health system catering to a diverse population in New Haven County, one of the most representative counties in the US. Since 2016, the free-text imaging reports for TTEs have been linked with structured tabular values for the cardiologist-defined echocardiographic features. The linked structured dataset consisted of clinical and operational labels, of which we selected 18 based on

their broad coverage of key conditions. These included ejection fraction (EF), global longitudinal strain, interventricular septal thickness (IVSd), aortic valve (AV) and mitral valve (MV) structure, and qualitative or quantitative features associated with AV and MV stenosis/regurgitation, including left ventricular outflow tract (LVOT) peak velocity and peak gradient, AV peak velocity and mean gradient, AV area by continuity, and AV area index. A brief overview of the data fields is included in **Table 1**. This dataset of 10,000 reports paired with corresponding structured labels was used for model evaluation (test set).

Structurally distinct TTE reports from MIMIC-III and MIMIC-IV datasets were used for external validation of our digitization approach.[11,12] MIMIC-III comprises deidentified EHR data from over forty thousand patients with a hospitalization that included an intensive care unit stay at the Beth Israel Deaconess Medical Center between 2001 and 2012. The data represents broad EHR fields spanning demographics, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality. MIMIC-IV is an updated version of the MIMIC-III database, which incorporates data up to 2019 and includes hospitalizations with emergency department visits. The current study leveraged echocardiographic reports from both MIMIC-III and MIMIC-IV. Representative examples of various report types from the different sources are included in **Supplemental Table S1**.

Model Development: Overall approach

We designed a two-step approach to convert unstructured TTE reports to structured data (digitize) using LLMs. All TTE reports in the YNHHS dataset post-2016 had corresponding clinician-annotated tabular data, which provided us with a training set without the need for manual annotation. However, reports from before this time (pre-2016) were only available in a

free-text format without corresponding tabular data. Moreover, the data had several variations in free text reporting across echocardiographers. From a randomly selected 3,000 pre-2016 YNHHS reports, we observed 5 different reporting formats (**Supplemental Table 2**). We randomly chose one report from each unique reporting format to encode this variation in our training set.

The utilization of Llama for finetuning, as opposed to alternative LLMs, was driven by considerations of its parameter efficiency, domain-specific architecture, and applicability to the medical text processing domain. By prioritizing factors such as model performance, accessibility, and computational power required to finetune the model, we aimed to optimize the efficiency and effectiveness of the finetuning process. This also included the ability to quantize Llama models into a 4-bit configuration for reduction in model size and memory usage.[13–15]

First-stage development: Finetuning Llama2 70b

In the initial phase, we fine-tuned the Llama2 70-billion-parameter LLM to generate TTE reports from the structured data in the post-2016 reports with syntactical characteristics - including the five formatting variations - of the pre-2016 reports. We trained the model on a dataset of 3,000 paired examples from the post-2016 dataset to create unstructured data that faithfully represents the format of pre-2016 reports for subsequent fine-tuning and testing (Prompt template – **Supplemental Table 2**).

The task of restructuring reports from post-2016 format to pre-2016 format involved fine-tuning the pre-trained Llama2-70b model to recognize and replicate the syntactical and structural elements of various pre-2016 reporting formats. The model was trained on a curated dataset containing examples of both post-2016 and pre-2016 reports. The dataset was carefully prepared, ensuring the pre-2016 reports represented multiple versions and styles to comprehensively

expose the model to various data formats. The model was fine-tuned over two epochs with a batch size of 4, using the Adam optimizer with a learning rate of 10^{-5}. The choice of hyperparameters, including the learning rate and epochs, was based on commonly adopted practices for fine-tuning large language models.[16,17] The fine-tuning process was monitored in a validation set, with early stopping of fine-tuning when validation loss did not improve for 5 consecutive evaluation steps to prevent overfitting. This approach was implemented to ensure that the model generalizes well to unseen data while maintaining high accuracy on the training set.

Second-stage development: Finetuning Llama2 13b

After the initial step, we used the restructured reports created with the Llama2 70-billion-parameter LLM to train a Llama2 13-billion-parameter LLM. We trained the model on a subset of 2,000 regenerated TTE reports, each paired with clinician-annotated tabular data. This approach enabled the model to learn from TTE reports that vary in formats while still being able to use corresponding clinician-annotated tabular data as the gold standard. This led to our model, HeartDx-LM, which is tailored to extracting structured fields from free-text narratives of TTE reports across the selected 18 clinical variables without requiring the large computational infrastructure needed for the 70-billion-parameter model. HeartDx-LM was trained to discern and extract critical information (Prompt table – **Supplemental Table 4**). We have made the model publicly available on HuggingFace at https://huggingface.co/CarDSLab/HeartDX-LM.

We also digitized the TTE reports using a non-finetuned Llama2-13b model (zero-shot Llama) to compare its performance with its finetuned counterparts.

Evaluation

We conducted a comprehensive evaluation of the model's performance across four distinct datasets – post-2016 YHNNS TTE reports (internal held-out test set), pre-2016 YNHHS reports, MIMIC-III TTE reports, and MIMIC-IV TTE reports.

Firstly, we employed a held-out set comprising 10,000 post-2016 YNHHS reports sourced from the YNHHS EHR. These reports were accompanied by their corresponding structured fields, allowing for direct comparison and assessment of the model's proficiency in extracting structured data from contemporary clinical narratives.

In addition to the post-2016 YNHHS dataset, we also examined the model's performance on 100 pre-2016 YNHHS reports and 100 reports each from the MIMIC-III and MIMIC-IV datasets. The pre-2016 YNHHS dataset used for model evaluation were a distinct set from the one used to develop synthetic examples and had clinical labels manually extracted by three clinical experts. The TTE reports from MIMIC-III were obtained from the EchoNotes Structured Database, which also includes echocardiogram reports from the intensive care unit.[18,19] In the MIMIC-IV dataset, reports were retrieved from discharge summaries that contained TTE report summaries. This structured echocardiogram database included key measures of cardiac structure and function, such as ejection fraction (EF), aortic valve (AV) and mitral valve (MV) structure, and qualitative or quantitative features associated with AV and MV stenosis/regurgitation. The other structured fields of interest like interventricular septal thickness (IVSd), left ventricular outflow tract (LVOT) peak velocity and peak gradient, AV peak velocity and mean gradient, AV area by continuity, and AV area index were derived through manual annotation by three clinical experts, who collaboratively established an annotation scheme delineating the criteria for extracting values for each of the 18 clinical variables. Each report was evaluated based on its

constituent sentences, and the clinicians' annotations were aggregated to create a gold standard for evaluating the model's performance. This ensured the accuracy and reliability of the ground truth.

In addition to evaluating the performance of our fine-tuned models, we also employed the Llama2-13b model without fine-tuning as a comparator to assess its capability in extracting structured data from clinical narratives without prior training on our datasets. This allowed us to benchmark the performance of our approach against the out-of-box (or zero-shot) performance of the Llama2-13b model.

Determining Optimal Training Data Volume for Model Fine-tuning

We investigated the impact of training data volume on model performance by fine-tuning multiple iterations of the Llama2-13b model with progressively increasing numbers of training reports (100, 200, 500, 1,000, and 2,000). We evaluated the model's performance on the held-out post-2016 YNHHS dataset for each iteration. This comprehensive evaluation framework aimed to elucidate the model's capabilities and provide insights into the optimal training data volume required for achieving robust performance in extracting structured information from diverse clinical narratives. We used a statistical benchmark of 95% accuracy to define robust performance.

Statistical Analysis

We assessed HeartDx-LM's performance using the accuracy of extracted values for both continuous and categorical variables. We reported the overall extraction accuracies and accuracies for individual clinical variables compared against the ground truth annotations.

Extraction accuracy was defined as the percentage of values correctly extracted by the model, with incorrect and failed extractions rates also reported. Additionally, Cohen's kappa statistic was used to evaluate the agreement between the model's extractions and the ground truth for both categorical and continuous variables. Specifically, continuous variables were categorized into discrete classes for the Kappa analysis. Each continuous variable in the training dataset was labeled as either 1 for available values or 0 for missing values. For the digitized dataset, the continuous variables were labeled based on their comparison with the original dataset: a label of 1 for values that were available in both the training and digitized datasets and correctly extracted, a label of 0 for values that were missing in both datasets (correctly identified as missing), a label of 2 for values available in the training dataset but extracted incorrectly in the digitized dataset, and a label of 3 for values available in the training dataset but missing in the digitized dataset. We calculated Cohen's Kappa for each continuous variable independently, measuring the agreement between the original and digitized labels across these four categories. The Kappa statistic was computed using the formula $\kappa = (P_o - P_e)/(1 - P_e)$, where P_o is the observed agreement between the two datasets and P_e is the expected agreement by chance. For multiclass Kappa, the observed and expected agreements considered all four categories to provide a comprehensive measure of agreement. By categorizing the continuous variables and then applying Cohen's Kappa, we ensured that the agreement between the original and digitized datasets was evaluated robustly, accounting for both correct and incorrect extractions as well as missing data.

The Cohen's Kappa statistic metric accounts for the possibility of agreement occurring by chance, providing a more robust measure of the model's reliability.[20] A kappa value closer to 1

indicates a high level of agreement, while a value closer to 0 suggests agreement is no better than chance.

RESULTS

Study Population

There were 8,612 unique patients with 10,000 post-2016 YNHHS reports in the test set, with a median age of 73.0 (IQR, 62.0 – 85.0) years, including 5,013 (50.1%) women, 694 (6.9%) non-Hispanic Blacks, 88 (0.9%) Hispanics, and 52 (0.5%) of Asian race. The range of distribution of clinical features (across both development and validation cohorts) are provided in **Table 1**.

Zero-shot model performance

The zero-shot Llama2-13b model generated fragmented, inconsistent, or irrelevant responses, resulting in incomplete and inaccurate extractions. This resulted in 0% extraction accuracy across all the 18 clinical variables. An example of zero-shot model prompt and response is shown in **Supplemental Table 5**.

Model performance in the held-out test set (post-2016 YNHHS dataset)

The HeartDx-LM model extracted 69,144 out of 70,032 values, yielding an accuracy rate of 98.7% and Cohen's Kappa value of 0.99. The accuracy rate was consistent across both continuous (45712/46387 - 98.5%) and categorical (23432/23645 - 99.1%) variables. The model incorrectly extracted 480 (0.7%) values and did not extract 408 (0.6%) values. The inaccurate values were most frequent for EF (97/9143), followed by AV peak velocity (96/8429), LVOT peak velocity (74/8065), and IVSd (48/8449).

Across continuous variables, the accuracy of the model for key clinical variables like EF, LVOT peak velocity, and AV peak velocity were 97.3% (8,902/9,143), 98.8% (7,969/8,065), and 98.8% (8,331/8,429), respectively. For key categorical variables of AV structure, AV regurgitation, MV regurgitation, and LV wall thickness, the accuracies were 98.6% (1,727/1,752), 99.1% (2,813/2,833), 99.7% (6,440/6,460), and 97.7% (3,938/4,032), respectively (**Table 2**).

Model performance in pre-2016 reports

In the 100 randomly sampled and expert-annotated pre-2016 reports, HeartDx-LM achieved an overall accuracy rate of 87.1% (extracting 909 out of the 1044 values), and Cohen's Kappa value of 0.86 across 18 clinical variables when compared against manually annotated labels. The model incorrectly extracted 11 (1.1%) values and failed to extract 124 (11.9%) values. The inaccurate values were most frequent for AV mean gradient (3/67), followed by MV structure (3/92), MV stenosis (2/25), MV regurgitation (1/93) and LV diastolic function (1/78).

HeartDx-LM maintained a high accuracy across both continuous (407/454 - 89.6%) and categorical (502/590 - 85.1%) variables. Accuracy of the model across key continuous variables, of EF, LVOT peak velocity, and AV peak velocity were 90.5% (86/95), 86.2% (50/58), and 92.2% (83/90), respectively. For key categorical variables of AV structure, AV regurgitation, MV regurgitation, and LV wall thickness, the accuracies were 83.2% (79/95), 95.2% (79/83), 93.5% (87/93), and 56.9% (33/58), respectively (**Table 3**).

External Validation: Model performance in MIMIC-III and MIMIC-IV TTE Reports

In 100 TTE reports from the MIMIC-III dataset, HeartDx-LM also demonstrated high accuracy in extracting structured clinical data. The model successfully extracted 615 out of 707 available values correctly, achieving an overall accuracy rate of 86.9% and Cohen's Kappa of 0.90. This included an accuracy of 72.4% for continuous variables (113/156) and 91.1% for categorical variables (502/551).

There were 12 (1.7%) values inaccurately extracted and 80 (11.3%) failed extractions, mainly in the qualitative labels. The inaccurate values were most frequent for AV structure (4/95), followed by AV regurgitation (2/91), AV stenosis (2/61), and MV regurgitation (1/93). The model failed to extract 80 (11.3%) values in the reports across all 18 variables. External validation in the MIMIC III dataset also demonstrated high accuracy, with the model achieving over 90% accuracy for key continuous and categorical clinical variables (e.g. EF: 97.8% [90/92], AV structure: 94.7% [90/95], MV regurgitation: 97.8% [88/90], and LV wall thickness: 92.7% [76/82]; **Table 4**).

In the MIMIC-IV dataset, the model successfully extracted 201 out of 220 available values, achieving an overall accuracy rate of 91.3% and Cohen's Kappa value of 0.95. This included an accuracy of 97.8% (44/45) for continuous variables and 89.7% (157/175) for categorical variables. The model extracted 2 (0.9%) incorrect values, 1 out of 45 values of EF and 1 out of 42 values MV regurgitation. Additionally, the model failed to extract 17 (7.7%) values present in the reports across the 18 variables. Values of accuracy for specific labels can be found in **Table 5**. The performance of HeartDX-LM across all 4 datasets is summarized in **Figure 3**.

Data Volume for Model Fine-tuning

In our evaluation of the data threshold necessary for model development, we analyzed the accuracy of our models as a function of progressively larger number of reports used for fine-tuning. The Llama2-13b models finetuned using 100 and 200 reports had accuracies of 13.5% and 85.7%, respectively. The accuracies increased to 97.8%, 98.2%, and 98.8 with the use of 500, 1000, and 2000 TTE reports, respectively (**Figure 2**). A minimum of 500 reports were necessary to achieve our pre-specified accuracy benchmark of 95%, with accuracy plateauing beyond this point.

DISCUSSION

We developed and validated HeartDx-LM, an innovative strategy to extract structured clinical data from unstructured clinical reports. This novel strategy leverages the output of an LLM to train a smaller, lightweight model, eliminating the need for high computational capacity in the final deployment. HeartDx-LM demonstrated robust performance in digitizing TTE reports across varying reporting formats from geographically and temporally distinct data sources and was able to successfully extract qualitative and quantitative clinical labels with high accuracy. The model's adaptability and extensibility enable its potential deployment in diverse and low-resource clinical settings and applicability to other diagnostic reports. Furthermore, our research determined the minimum threshold for the number of TTE reports required for fine-tuning models for optimally balanced accuracy and computational resources, providing valuable guidance for future model development.

Prior models to digitize TTE reports predominantly relied on rule-based or keyword-based NLP models.[21–23] For example, early studies have used specific keywords and predefined rules to analyze echocardiography and radiology reports without considering variations in

reporting formats and dynamic changes in clinical parameters.[24–26] Moreover, these methods predominantly focus on extracting a few specific clinical labels, such as low EF, and often fail to capture the full spectrum of clinically relevant labels needed for broader applications, healthcare decision-making, and planning.[27]

In contrast, HeartDx-LM, was engineered to extract multiple qualitative and quantitative clinical labels. This comprehensive extraction capability enhances the model's utility in clinical practice as it can be scaled to similar domains, where diagnostic information is captured in unstructured reports. This innovative two-step approach to digitizing entire reports is an alternative for generating training sets for smaller LLMs, reducing the need for extensive manual annotation and the reliance on high computational power. Since most TTE reports are stored as unstructured text, this approach can significantly expand our dataset for new model training, and enable access to diverse settings, including those with limited technological infrastructure, with potential use for cross-setting electronic clinical quality measures.[2,28–31]

Our study has limitations that deserve consideration. Notably, the performance of our models showed variability across different clinical fields, especially when certain domain-specific terms were reported differently across different datasets. Nonetheless, the overall and field-wise performance was acceptable across all external sites. Additionally, the computational resources required for finetuning LLMs may pose practical constraints in real-world healthcare settings. However, the deployment of the lightweight finetuned model, that we have also publicly released on HuggingFace, does not require intensive computational resources and can be used for transformation of unstructured reports into tabular datasets. Finally, while our study underscores the potential in using LLMs for the automated extraction of structured clinical information from unstructured narratives in EHR, future research should prioritize enhancing the interpretability of

LLM-based models. This can be achieved by delving into the contextual analysis of clinical notes and refining the model's ability to discern subtle nuances in medical language to further optimize the performance and generalizability of LLM-based approaches.

CONCLUSION

We developed a novel method using paired large and moderate-sized LLMs to automate the extraction of unstructured echocardiographic reports into tabular datasets. Our approach represents a scalable strategy that transforms unstructured reports into computable elements that can be leveraged to improve cardiovascular care quality and enable research.

ACKNOWLEDGEMENTS

Funding

Dr. Khera was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (under awards R01AG089981, R01HL167858, and K23HL153775) and the Doris Duke Charitable Foundation (under award 2022060). Dr. Oikonomou was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award F32HL170592). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflict of Interest

Dr. Khera is an Associate Editor of JAMA and is a co-founder of Ensight-AI. Dr. Khera receives support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under awards R01HL167858 and K23HL153775) and the Doris Duke Charitable Foundation (under award 2022060). He receives support from the Blavatnik Foundation through the Blavatnik Fund for Innovation at Yale. He also receives research support, through Yale, from Bristol-Myers Squibb, BridgeBio, and Novo Nordisk. In addition to 63/346,610, Dr. Khera is a coinventor of U.S. Pending Patent Applications WO2023230345A1, US20220336048A1, 63/484,426, 63/508,315, 63/580,137, 63/619,241, 63/346,610, 63/562,335 and 18/813,882. Dr. Khera, Dr. Oikonomou and Mr. Vasisht Shankar are co-inventors of the US patent application 63/606,203. Dr. Khera and Dr. Oikonomou are co-founders of Evidence2Health, a precision health platform to improve evidence-based cardiovascular care. Dr. Oikonomou receives support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under

award F32HL170592). He is a co-inventor of the U.S. Patent Applications 18/813,882, 17/720,068, 63/619,241, 63/177,117, 63/580,137, 63/606,203, 63/562,335, US11948230B2 , US20210374951A1. He has been a consultant for Caristo Diagnostics Ltd and Ensight-AI Inc, and has received royalty fees from technology licensed through the University of Oxford.. Mr. Vasisht Shankar works as a data scientist at Evidence2Health (outside the current work). Dr. Nadkarni is a founder of Renalytix, Pensieve, and Verici and provides consultancy services to AstraZeneca, Reata, Renalytix, and Pensieve. He also has equity in Renalytix, Pensieve, and Verici.

Data Availability

The dataset cannot be made publicly available because they are electronic health records. Sharing this data externally without proper consent could compromise patient privacy and would violate the Institutional Review Board's approval for the study.

Code Availability

The code for the study is available from the authors upon request.

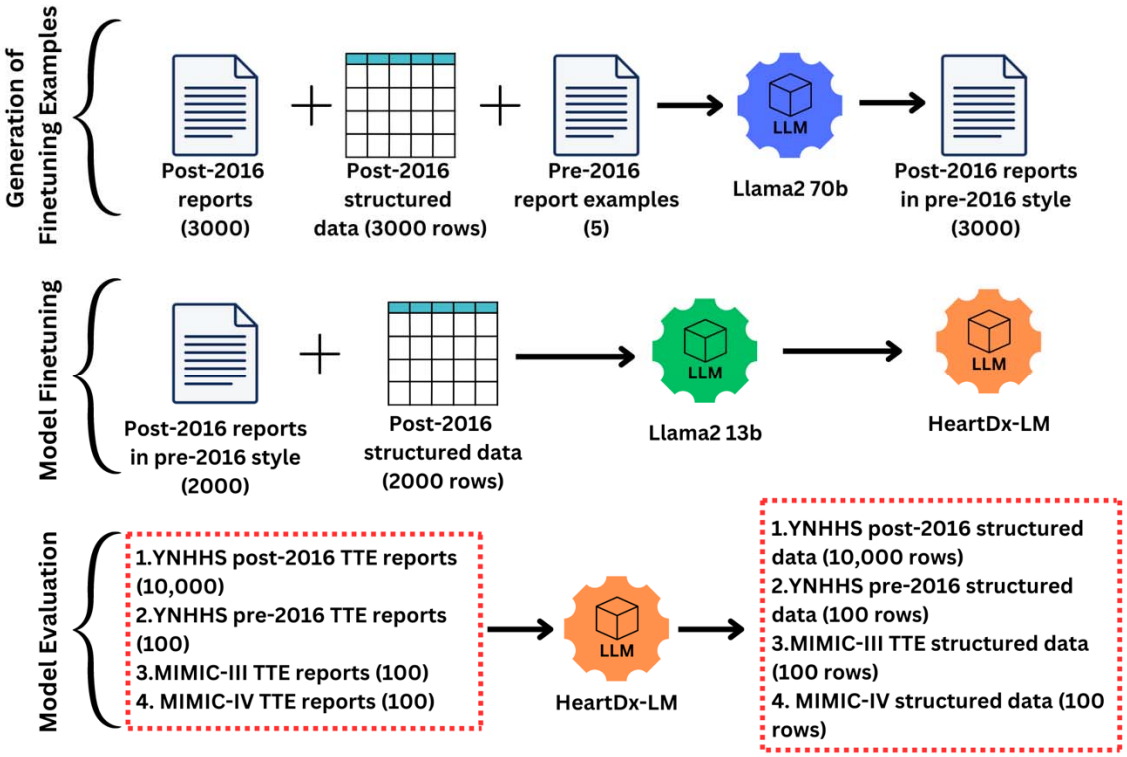
REFERENCES

1. Sangha, V. *et al.* Detection of Left Ventricular Systolic Dysfunction From Electrocardiographic Images. *Circulation* **148**, 765–777 (2023).
2. Khunte, A. *et al.* Detection of left ventricular systolic dysfunction from single-lead electrocardiography adapted for portable and wearable devices. *NPJ Digit Med* **6**, 124 (2023).
3. Dhingra, L. S., Shen, M., Mangla, A. & Khera, R. Cardiovascular Care Innovation through Data-Driven Discoveries in the Electronic Health Record. *Am. J. Cardiol.* **203**, 136–148 (2023).
4. Consultant, H. I. T. Why unstructured data holds the key to intelligent healthcare systems [Internet]. *Atlanta (GA): HIT Consultant* (2015).
5. Hashir, M. & Sawhney, R. Towards unstructured mortality prediction with free-text clinical notes. *J. Biomed. Inform.* **108**, 103489 (2020).
6. Aghajani Nargesi, A. *et al.* Abstract 16207: Deep learning-based natural language processing of discharge summaries for automated identification of heart failure with reduced ejection fraction. *Circulation* (2023) doi:10.1161/circ.148.suppl_1.16207.
7. Brown, J. R. *et al.* Information Extraction From Electronic Health Records to Predict Readmission Following Acute Myocardial Infarction: Does Natural Language Processing Using Clinical Notes Improve Prediction of Readmission? *J. Am. Heart Assoc.* **11**, e024198 (2022).
8. Wang, Y. *et al.* Clinical information extraction applications: A literature review. *J. Biomed. Inform.* **77**, 34–49 (2018).
9. Kreimeyer, K. *et al.* Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J. Biomed. Inform.* **73**, 14–29 (2017).
10. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and Adoption of Large Language Models in Medicine. *JAMA* **330**, 866–869 (2023).
11. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016).
12. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* **10**, 1 (2023).
13. Li, B., Chen, J. & Zhu, J. Memory Efficient Optimizers with 4-bit States. *Adv. Neural Inf. Process. Syst.* **abs/2309.01507**, (2023).
14. Kodali, R. K., Upreti, Y. P. & Boppana, L. A Quantization Approach for the Reduced Size of Large Language Models. in *2024 16th International Conference on Knowledge and Smart Technology (KST)* 144–148 (IEEE, 2024).
15. Kim, J. *et al.* Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *Adv. Neural Inf. Process. Syst.* **abs/2305.14152**, (2023).
16. Ma, X., Fang, G. & Wang, X. LLM-Pruner: On the structural pruning of large language models. *Adv. Neural Inf. Process. Syst.* **abs/2305.11627**, (2023).
17. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. *Adv. Neural Inf. Process. Syst.* **abs/2305.14314**, (2023).
18. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215-20 (2000).

19. Kwak, G. H. *et al.* EchoNotes Structured Database derived from MIMIC-III (ECHO-NOTE2NUM). PhysioNet https://doi.org/10.13026/XHRZ-HT59 (2024).
20. Bland, M. Cohen's kappa. University of York Department of Health Sciences. Preprint at (2008).
21. Gehrmann, S. *et al.* Comparing Rule-Based and Deep Learning Models for Patient Phenotyping. *arXiv [cs.CL]* (2017).
22. Mykowiecka, A., Marciniak, M. & Kupść, A. Rule-based information extraction from patients' clinical data. *J. Biomed. Inform.* **42**, 923–936 (2009).
23. Silverman, G. M. *et al.* NLP Methods for Extraction of Symptoms from Unstructured Data for Use in Prognostic COVID-19 Analytic Models. *jair* **72**, 429–474 (2021).
24. Buckley, J. M. *et al.* The feasibility of using natural language processing to extract clinical information from breast pathology reports. *J. Pathol. Inform.* **3**, 23 (2012).
25. Casey, A. *et al.* A systematic review of natural language processing applied to radiology reports. *BMC Med. Inform. Decis. Mak.* **21**, 179 (2021).
26. Viani, N. *et al.* Information extraction from Italian medical reports: An ontology-driven approach. *Int. J. Med. Inform.* **111**, 140–148 (2018).
27. Nargesi, A. A. *et al.* Automated identification of heart failure with reduced ejection fraction using deep learning-based natural language processing. *medRxiv* (2023) doi:10.1101/2023.09.10.23295315.
28. Oikonomou, E. K., Holste, G., Nadkarni, G., Wang, Z. & Khera, R. Cross-modal validation of an artificial intelligence video-based approach for the automated risk stratification of aortic stenosis. *J. Am. Coll. Cardiol.* (2024) doi:10.1016/s0735-1097(24)03408-9.
29. Oikonomou, E. K. *et al.* A Multimodal Video-Based AI Biomarker for Aortic Stenosis Development and Progression. *JAMA Cardiol* (2024) doi:10.1001/jamacardio.2024.0595.
30. Oikonomou, E. K. *et al.* Real-world evaluation of an algorithmic machine-learning-guided testing approach in stable chest pain: a multinational, multicohort study. *Eur Heart J Digit Health* **5**, 303–313 (2024).
31. Mph, A. A. M. D. *et al.* Deep learning-enabled detection of aortic stenosis from noisy single lead electrocardiograms. *medRxiv* (2023) doi:10.1101/2023.09.29.23296310.

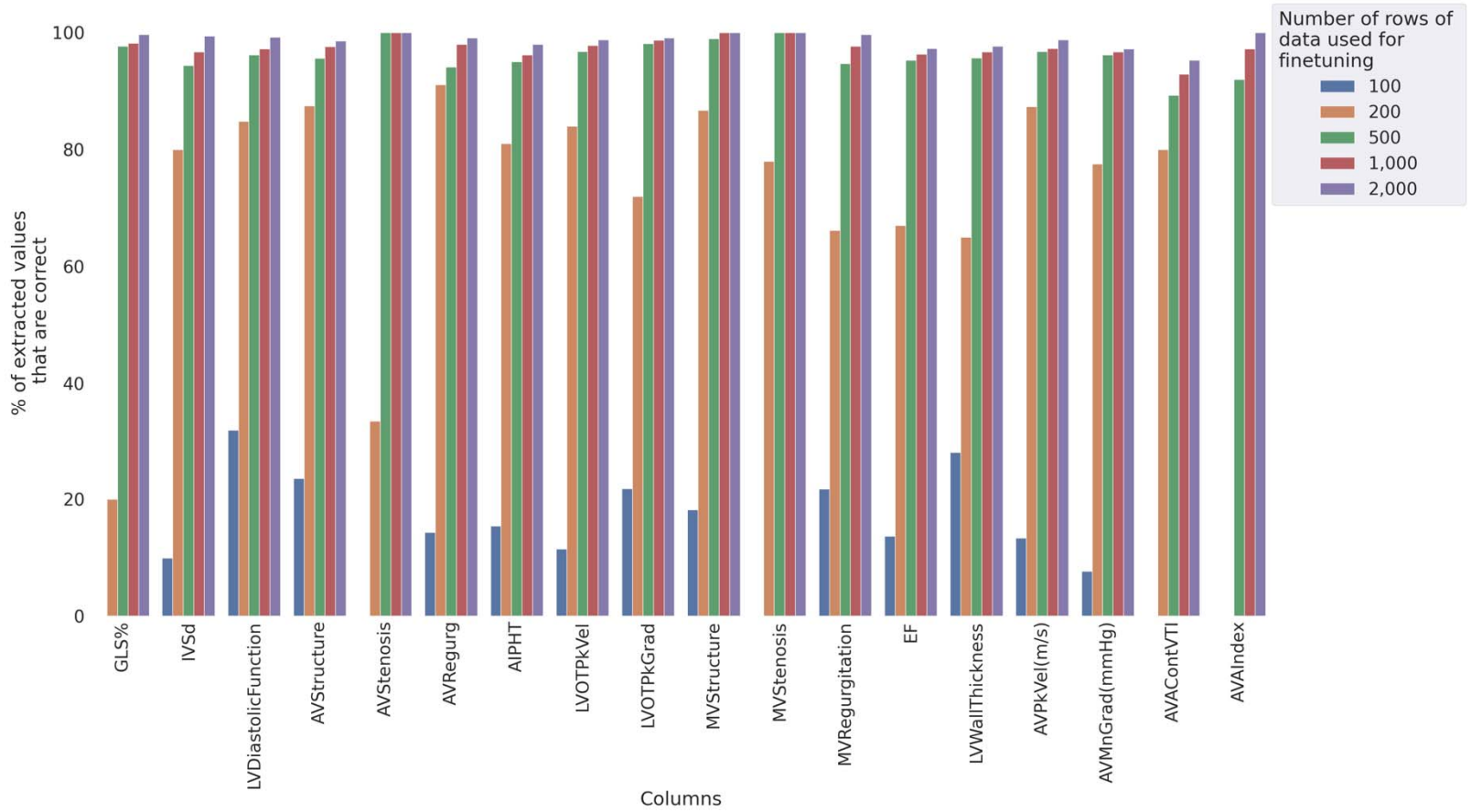
FIGURES

Figure 1: Model Development Approach and Study Design.



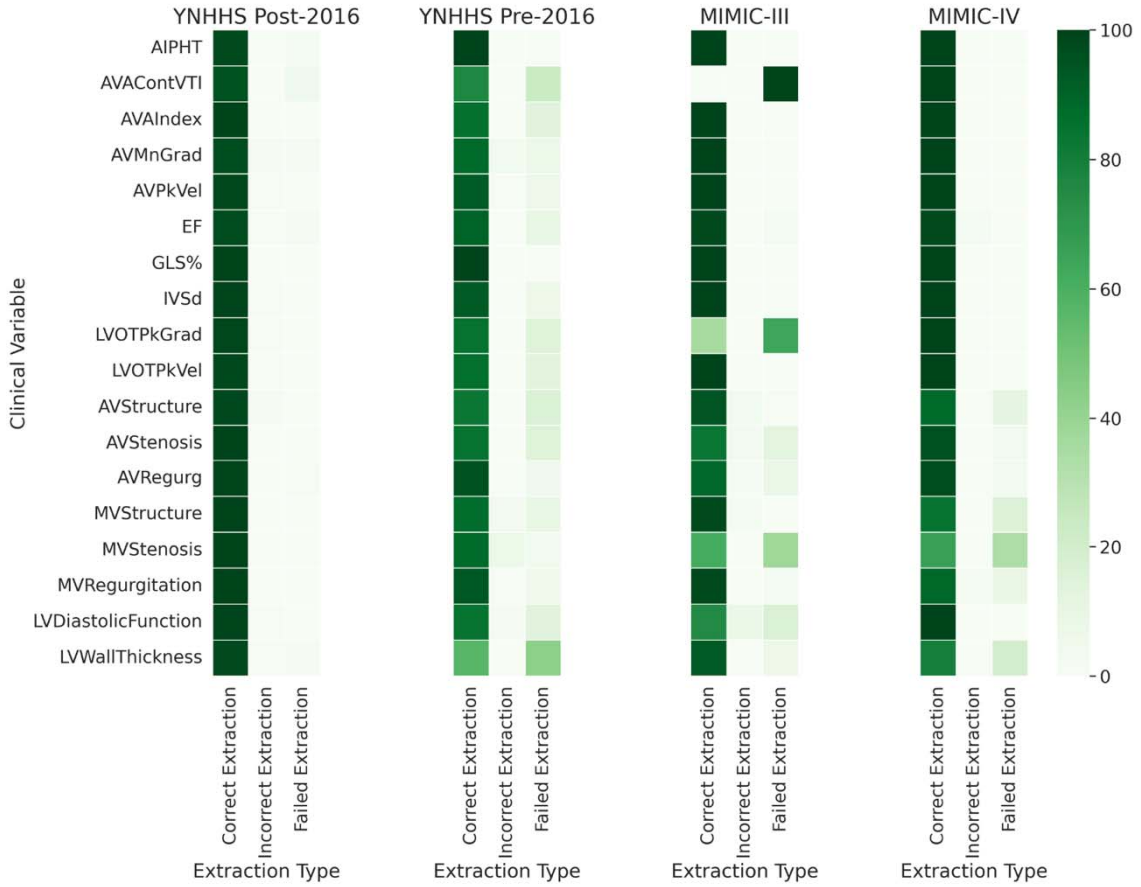
Abbreviations: YNHHS, Yale New Haven Health System; MIMIC, Medical Information Mart for Intensive Care

Figure 2. Performance of models fine-tuned with varying number of paired unstructured reports and structured tables for tabulation of clinical variables from unstructured reports.



Abbreviations: AVA Cont VTI, Aortic Valve Area Calculated by Velocity Time Integral ; AVA Index, Aortic Valve Area Index; AV Mn Grad, Aortic Valve Mean Gradient; AIPHT, Aortic Insufficiency Pressure Half-Time; AV Pk Vel, Aortic Valve Peak Velocity; AV Regurgitation, Aortic Valve Regurgitation; AV Stenosis, Aortic Valve Stenosis; AV Structure, Aortic Valve Structure; GLS, Global Longitudinal Strain; IVSd, Interventricular Septum Thickness; ; LV Diastolic Function, Left Ventricular Diastolic Function; LVOT Pk Grad, Left Ventricular Outflow Tract Peak Gradient; LVOT Pk Vel, Left Ventricular Outflow Tract Peak Velocity; LV Wall Thickness, Left Ventricular Wall Thickness; MV Regurgitation, Mitral Valve Regurgitation; MV Stenosis, Mitral Valve Stenosis; MV Structure, Mitral Valve Structure.

Figure 3. Accuracy of HeartDX-LM for Label Extraction Across the Four Datasets.



Abbreviations: AVA Cont VTI, Aortic Valve Area Calculated by Velocity Time Integral ; AVA Index, Aortic Valve Area Index; AV Mn Grad, Aortic Valve Mean Gradient; AIPHT, Aortic Insufficiency Pressure Half-Time; AV Pk Vel, Aortic Valve Peak Velocity; AV Regurgitation, Aortic Valve Regurgitation; AV Stenosis, Aortic Valve Stenosis; AV Structure, Aortic Valve Structure; GLS, Global Longitudinal Strain; IVSd, Interventricular Septum Thickness; ; LV Diastolic Function, Left Ventricular Diastolic Function; LVOT Pk Grad, Left Ventricular Outflow Tract Peak Gradient; LVOT Pk Vel, Left Ventricular Outflow Tract Peak Velocity; LV Wall Thickness, Left Ventricular Wall Thickness; MV Regurgitation, Mitral Valve Regurgitation; MV Stenosis, Mitral Valve Stenosis; MV Structure, Mitral Valve Structure.

TABLES

Table 1: Data Summarization of the train, test, and validation datasets

Clinical Domain	Data Type	YNHHS post-2016 dataset		YNHHS pre-2016 dataset		MIMIC-III dataset		MIMIC-IV dataset	
		Data Available	Median [IQR] / %	Data Available	Median [IQR]/ %	Data Available	Median [IQR]/ %	Data Available	Median [IQR]/ %
Aortic Insufficiency Pressure Half-Time (AIPHT)	Continuous	1,200/10,000	508.00 [417.00, 611.00]	0/100	-	0/100	-	0/100	-
Aortic Valve Area Calculated by Velocity Time Integral	Continuous	1,032/10,000	1.42 [0.95, 1.83]	13/100	2.50 [1.45, 3.1]	7/100	2.10 [1.73, 2.55]	0/100	-
Aortic Valve Area Index	Continuous	27/10,000	0.58 [0.44, 0.86]	14/100	0.8 [0.42, 1.6]	0/100	-	0/100	-
Aortic Valve Mean Gradient	Continuous	1,278/10,000	12.00 [8.00, 22.00]	67/100	14.00 [10.00, 27.00]	3/100	17.00 [14.50, 19.50]	0/100	-
Aortic Valve Peak Velocity	Continuous	8,429/10,000	1.44 [1.22, 1.77]	90/100	1.5 [1.2, 1.9]	0/100	-	0/100	-
Ejection Fraction	Continuous	9,143/10,000	61.00 [53.00, 66.00]	95/100	62.00 [60.00, 65.00]	92/100	55.00 [50.00, 60.00]	45/100	58.00 [50.00, 63.00]
Global Longitudinal Strain (GLS%)	Continuous	384/10,000	-17.0 [-19.0, -15.0]	0/100	-	0/100	-	0/100	-
Interventricular	Continuous	8,449/10,000	0.98 [0.86,	97/100	1.0	1/100	1.0 [-]	0/100	-

Septum Thickness (IVSd)			1.13]		[0.80, 1.20]				
Left Ventricular Outflow Tract Peak Gradient	Continuous in YNHHS reports and Categorical in MIMIC-III reports	7,940/10,000	4.00 [3.00, 6.00]	20/100	3.69 [2.67, 6.38]	53/100	53%	0/100	-
Left Ventricular Outflow Tract Peak Velocity	Continuous	8,065/10,000	1.02 [0.87, 1.18]	58/100	1.00 [0.50, 1.30]	0/100	-	0/100	-
Aortic Valve Structure	Categorical								
Normal		1,607/10,000	16.07%	39/100	39%	39/100	39%	20	20%
Bicuspid		64/10,000	0.64%	12/100	12%	-	-	-	-
Tricuspid		81/10,000	0.81%	28/100	28%	-	-	-	-
Mildly Thickened		-	-	-	-	45/100	45%	3	3%
Moderately Thickened		-	-	-	-	7/100	7%	1	1%
Severely Thickened		-	-	-	-	4/100	4%	1	1%
Aortic Valve Stenosis	Categorical								
No		-	-	31/100	31%	40/100	40%	19	19%
Mild		238/10,000	2.38%	15/100	15%	8/100	8%	5	5%
Mild-Mod		18/10,000	0.18%	1/100	1%	4/100	4%	-	-
Moderate		67/10,000	0.67%	5/100	5%	3/100	3%	-	-
Mod-Sev		42/10,000	0.42%	3/100	3%	1/100	1%	-	-
Severe		187/10,000	1.87%	11/100	11%	5/100	5%	-	-
Aortic Valve Regurgitation	Categorical								
No		-	-	-	-	57/100	57%	19	19%

Mild		1,630/10,000	16.30%	37/100	37%	14/100	14%	9	9%
Trace		786/10,000	7.86%	12/100	12%	13/100	13%	5	5%
Mild-Mod		209/10,000	2.09%	8/100	8%	3/100	3%	-	-
Moderate		177/10,000	1.77%	14/100	14%	1/100	1%	1	1%
Mod-Sev		17/10,000	0.17%	12/100	12%	2/100	2%	-	-
Severe		14/10,000	0.14%	-	-	1/100	1%	-	-
Mitral Valve Structure	Categorical								
Normal		1,180/10,000	11.80%	59/100	59%	19/100	19%	19	19%
Thickened		355/10,000	3.55%	33/100	33%	62/100	62%	10	10%
Myxomatous		83/10,000	0.83%	-	-	-	-	-	-
Tethered		27/10,000	0.27%	-	-	-	-	-	-
Rheumatic		11/10,000	0.11%	-	-	-	-	-	-
Not Well Seen		-	-	-	-	6/100	6%	3	3%
Mitral Valve Stenosis	Categorical								
No		-	-	-	-	13/100	13%	-	-
Mild		169/10,000	1.69%	19/100	19%	4/100	4%	2	2%
Moderate		31/10,000	0.31%	3/100	3%	1/100	1%	1	1%
Trace		18/10,000	0.18%	1/100	1%	1/100	1%	-	-
Severe		13/10,000	0.13%	2/100	2%	-	-	-	-
Mild- Mod		5/10,000	0.05%	-	-	2/100	2%	-	-
Mod- Sev		4/10,000	0.04%	-	-	-	-	-	-
Mitral Valve Regurgitation	Categorical								
No		-	-	-	-	19/100	19%	14	14%
Mild		3,079/10,000	30.79%	55/100	55%	20/100	20%	13	13%
Trace (Trivial)		2,110/10,000	21.10%	16/100	16%	32/100	32%	8	8%
Moderate		551/10,000	5.51%	5/100	5%	9/100	9%	5	5%
Mild-Mod		380/10,000	3.80%	2/100	2%	7/100	7%	1	1%
Mod-Sev		173/10,000	1.73%	1/100	1%	-	-	-	-
Severe		157/10,000	1.57%	3/100	3%	2/100	2%	2	2%

Trace		10/10,000	0.10%	11/100	11%	1/100	1%	-	-
Left Ventricular Diastolic Function	Categorical								
Mild		3,147/10,000	31.47%	11/100	11%	4/100	4%	3	3%
Normal		2,032/10,000	20.32%	57/100	57%	17/100	17%	1	1%
Moderate		821/10,000	8.21%	9/100	9%	2/100	2%	-	-
Severe		121/10,000	1.21%	1/100	1%	1/100	1%	-	-
Left Ventricular Wall Thickness	Categorical								
Mildly Increased		2,349/10,000	23.49%	33/100	33%	37/100	37%	9	9%
Normal		1,314/10,000	13.14%	18/100	18%	44/100	44%	1	1%
Moderately Increased		303/10,000	3.03%	6/100	6%	1/100	1%	-	-
Severely Increased		58/10,000	0.58%	1/100	1%	-	-	-	-
Decreased		8/10,000	0.08%	-	-	-	-	-	-

Table 2: Model Performance Evaluation of HeartDX-LM on the held-out test set.

Metrics Column	Available Values	Correct: N (%) (Extracted value matches original value)	Incorrect: N (%) (Extracted value differs from original value)	Failed: N (%) (Value not extracted)	Cohen's Kappa
All variables	70,032	69,144 (98.7)	480 (0.7)	408 (0.6)	0.99
AIPHT	1,200	1,173 (98)	0 (0)	27 (2)	0.98
AVA Cont VTI	1,032	981 (95.3)	0 (0)	51 (4.7)	0.97
AVA Index	27	27 (100)	0 (0)	0 (0)	1.0
AV Mn Grad	1,728	1,680 (97.2)	25 (1.4)	23 (1.4)	0.98
AV Pk Vel	8,429	8,331 (98.8)	96 (1.1)	2 (0.1)	0.96
Ejection Fraction	9,143	8,902 (97.3)	97 (1.1)	144 (1.6)	0.87
GLS%	384	383 (99.7)	0 (0)	1 (0.3)	0.99
IVSd	8,449	8,401 (99.4)	48 (0.6)	0 (0)	0.98
LVOT Pk Grad	7,940	7,868 (99.1)	43 (0.6)	29 (0.3)	0.98
LVOT Pk Vel	8,065	7,969 (98.8)	74 (0.9)	22 (0.3)	0.97
AV Structure	1,752	1,727 (98.6)	25 (1.7)	0 (0)	0.99
AV Stenosis	552	552 (100)	0 (0)	0 (0)	1.0
AV Regurgitation	2,833	2,813 (99.1)	0 (0)	20 (0.9)	0.99
MV Structure	1,656	1,656 (100)	0 (0)	0 (0)	1.0
MV Stenosis	240	240 (100)	0 (0)	0 (0)	1.0
MV Regurgitation	6,460	6,440 (99.7)	0 (0)	20 (0.3)	0.99
LV Diastolic Function	6,121	6,074 (99.2)	47 (0.8)	0 (0)	0.99
LV Wall Thickness	4,032	3,938 (97.7)	25 (0.6)	69 (1.7)	0.98

Abbreviations: AVA Cont VTI, Aortic Valve Area Calculated by Velocity Time Integral ; AVA Index, Aortic Valve Area Index; AV Mn Grad, Aortic Valve Mean Gradient; AIPHT, Aortic Insufficiency Pressure Half-Time; AV Pk Vel, Aortic Valve Peak Velocity; AV Regurgitation, Aortic Valve Regurgitation; AV Stenosis, Aortic Valve Stenosis; AV Structure, Aortic Valve Structure; GLS, Global Longitudinal Strain; IVSd, Interventricular Septum Thickness; ; LV Diastolic Function, Left Ventricular Diastolic Function; LVOT Pk Grad, Left Ventricular Outflow Tract Peak Gradient; LVOT Pk Vel, Left Ventricular Outflow Tract Peak Velocity; LV Wall Thickness, Left Ventricular Wall Thickness; MV Regurgitation, Mitral Valve Regurgitation; MV Stenosis, Mitral Valve Stenosis; MV Structure, Mitral Valve Structure.

Table 3: Model Performance Evaluation of HeartDX-LM on Pre-2016 Reports.

Metrics Column	Available Values	Correct: N (%) (Extracted value matches original value)	Incorrect: N (%) (Extracted value differs from original value)	Failed: N (%) (Value not extracted)	Cohen's Kappa
All variables	1044	909 (87.1)	11 (1.1)	124 (11.9)	0.86
AIPHT	0	-	-	-	-
AVA Cont VTI	13	10 (76.9)	0 (0)	3 (23.1)	0.87
AVA Index	14	12 (85.7)	0 (0)	2 (14.3)	0.92
AV Mn Grad	67	59 (88)	3 (4.4)	5 (74.6)	0.84
AV Pk Vel	90	83 (92.2)	1 (1.1)	6 (6.6)	0.71
Ejection Fraction	95	86 (90.5)	0 (0)	9 (9.5)	0.50
GLS%	0	-	-	-	-
IVSd	97	90 (92.8)	0 (0)	7 (7.2)	0.44
LVOT Pk Grad	20	17 (85)	0 (0)	3 (15)	0.91
LVOT Pk Vel	58	50 (86.2)	0 (0)	8 (13.8)	0.85
AV Structure	95	79 (83.2)	0 (0)	16 (16.8)	0.35
AV Stenosis	66	56 (84.8)	0 (0)	10 (15.2)	0.81
AV Regurgitation	83	79 (95.2)	0 (0)	4 (4.8)	0.87
MV Structure	92	80 (86.9)	3 (3.2)	9 (9.8)	0.53
MV Stenosis	25	22 (88)	2 (8)	1 (4)	0.92
MV Regurgitation	93	87 (93.5)	1 (1.1)	5 (5.4)	0.67
LV Diastolic Function	78	66 (84.6)	1 (1.3)	11 (14.1)	0.72
LV Wall Thickness	58	33 (56.9)	0 (0)	25 (43.1)	0.60

Abbreviations: AVA Cont VTI, Aortic Valve Area Calculated by Velocity Time Integral ; AVA Index, Aortic Valve Area Index; AV Mn Grad, Aortic Valve Mean Gradient; AIPHT, Aortic Insufficiency Pressure Half-Time; AV Pk Vel, Aortic Valve Peak Velocity; AV Regurgitation, Aortic Valve Regurgitation; AV Stenosis, Aortic Valve Stenosis; AV Structure, Aortic Valve Structure; GLS, Global Longitudinal Strain; IVSd, Interventricular Septum Thickness; ; LV Diastolic Function, Left Ventricular Diastolic Function; LVOT Pk Grad, Left Ventricular Outflow Tract Peak Gradient; LVOT Pk Vel, Left Ventricular Outflow Tract Peak Velocity; LV Wall Thickness, Left Ventricular Wall Thickness; MV Regurgitation, Mitral Valve Regurgitation; MV Stenosis, Mitral Valve Stenosis; MV Structure, Mitral Valve Structure.

Table 4: Model Performance Evaluation of HeartDX-LM on MIMIC-III TTE Reports.

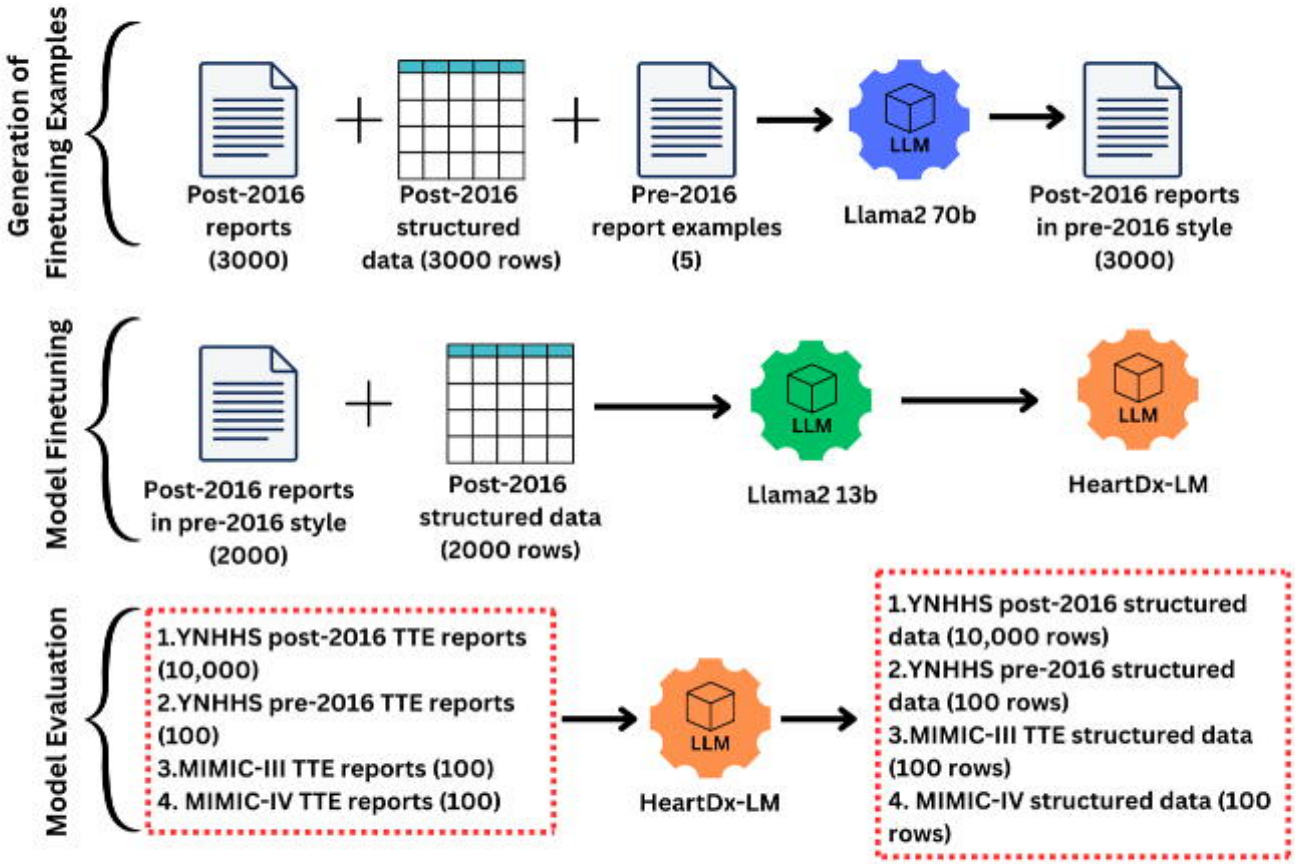
Metrics Column	Available Values	Correct: N (%) (Extracted value matches original value)	Incorrect: N (%) (Extracted value differs from original value)	Failed: N (%) (Value not extracted)	Cohen's Kappa
All variables	707	615 (86.9)	12 (1.7)	80 (11.3)	0.90
AIPHT	0	-	-	-	-
AVA Cont VTI	7	0 (0)	0 (0)	7 (100)	0.48
AVA Index	0	-	-	-	-
AV Mn Grad	3	3 (100)	0 (0)	0 (0)	1.0
AV Pk Vel	0	-	-	-	-
Ejection Fraction	92	90 (97.8)	0 (0)	2 (2.2)	0.88
GLS%	0	-	-	-	
IVSd	1	1 (100)	0 (0)	0 (0)	1.0
LVOT Pk Grad	53	19 (35.8)	0 (0)	34 (64.2)	0.50
LVOT Pk Vel	0	-	-	-	-
AV Structure	95	90 (94.7)	4 (4.2)	1 (1.1)	0.65
AV Stenosis	61	51 (83.6)	2 (3.3)	8 (13.1)	0.81
AV Regurgitation	91	81 (89)	2 (2.2)	8 (8.8)	0.61
MV Structure	87	85 (97.7)	2 (2.3)	0 (0)	0.92
MV Stenosis	21	13 (61.9)	0 (0)	8 (38.1)	0.77
MV Regurgitation	90	88 (97.8)	0 (0)	2 (2.2)	0.90
LV Diastolic Function	24	18 (75)	2 (8.3)	4 (16.7)	0.84
LV Wall Thickness	82	76 (92.7)	0 (0)	6 (7.3)	0.83

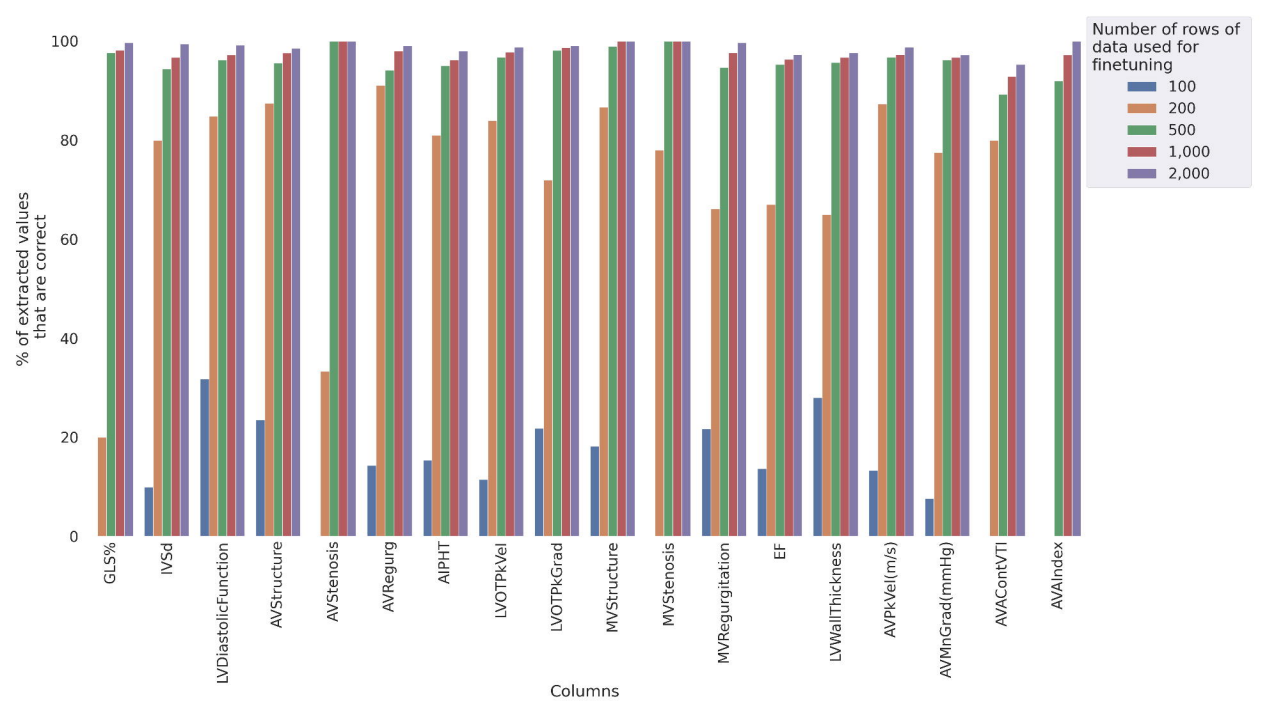
Abbreviations: AVA Cont VTI, Aortic Valve Area Calculated by Velocity Time Integral ; AVA Index, Aortic Valve Area Index; AV Mn Grad, Aortic Valve Mean Gradient; AIPHT, Aortic Insufficiency Pressure Half-Time; AV Pk Vel, Aortic Valve Peak Velocity; AV Regurgitation, Aortic Valve Regurgitation; AV Stenosis, Aortic Valve Stenosis; AV Structure, Aortic Valve Structure; GLS, Global Longitudinal Strain; IVSd, Interventricular Septum Thickness; ; LV Diastolic Function, Left Ventricular Diastolic Function; LVOT Pk Grad, Left Ventricular Outflow Tract Peak Gradient; LVOT Pk Vel, Left Ventricular Outflow Tract Peak Velocity; LV Wall Thickness, Left Ventricular Wall Thickness; MV Regurgitation, Mitral Valve Regurgitation; MV Stenosis, Mitral Valve Stenosis; MV Structure, Mitral Valve Structure.

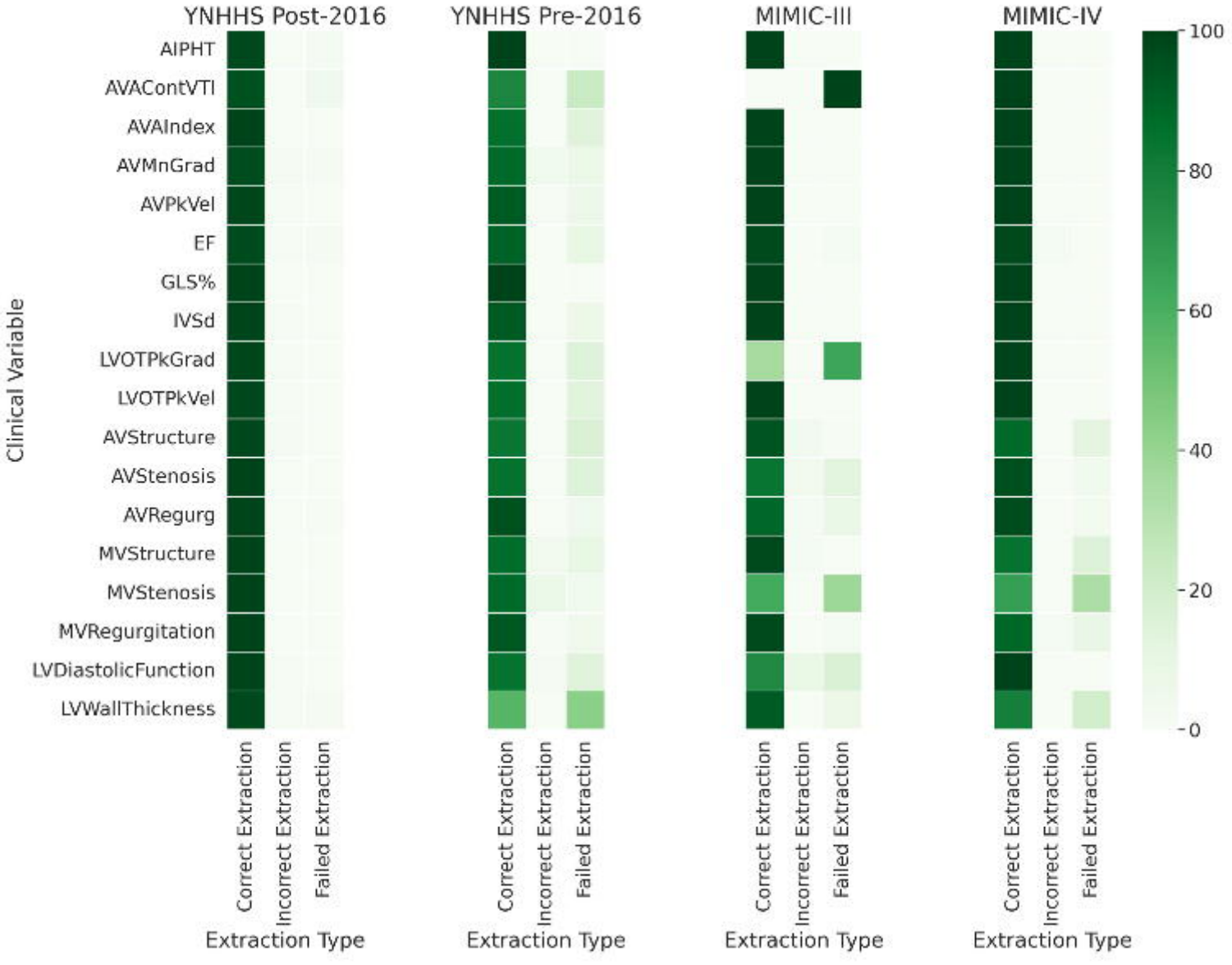
Table 5: Model Performance Evaluation of HeartDX-LM on MIMIC-IV Reports.

Metrics Column	Available Values	Correct: N (%) (Extracted value matches original value)	Incorrect: N (%) (Extracted value differs from original value)	Failed: N (%) (Value not extracted)	Cohen's Kappa
All variables	220	201 (91.3)	2 (0.9)	17 (7.7)	0.95
AIPHT	0	-	-	-	-
AVA Cont VTI	0	-	-	-	-
AVA Index	0	-	-	-	-
AV Mn Grad	0	-	-	-	-
AV Pk Vel	0	-	-	-	-
Ejection Fraction	45	44 (97.8)	1 (2.2)	0 (0)	0.98
GLS%	0	-	-	-	-
IVSd	0	-	-	-	-
LVOT Pk Grad	0	-	-	-	-
LVOT Pk Vel	0	-	-	-	-
AV Structure	25	22 (88)	0 (0)	3 (12)	0.92
AV Stenosis	24	23 (95.8)	0 (0)	1 (4.2)	0.97
AV Regurgitation	34	33 (97)	0 (0)	1 (3)	0.98
MV Structure	32	27 (84.4)	0 (0)	5 (15.6)	0.89
MV Stenosis	3	2 (66.7)	0 (0)	1 (33.3)	0.83
MV Regurgitation	43	38 (88.4)	1 (1.1)	4 (9.3)	0.90
LV Diastolic Function	4	4 (100)	0 (0)	0 (0)	1.0
LV Wall Thickness	10	8 (80)	0 (0)	2 (20)	0.89

Abbreviations: AVA Cont VTI, Aortic Valve Area Calculated by Velocity Time Integral ; AVA Index, Aortic Valve Area Index; AV Mn Grad, Aortic Valve Mean Gradient; AIPHT, Aortic Insufficiency Pressure Half-Time; AV Pk Vel, Aortic Valve Peak Velocity; AV Regurgitation, Aortic Valve Regurgitation; AV Stenosis, Aortic Valve Stenosis; AV Structure, Aortic Valve Structure; GLS, Global Longitudinal Strain; IVSd, Interventricular Septum Thickness; ; LV Diastolic Function, Left Ventricular Diastolic Function; LVOT Pk Grad, Left Ventricular Outflow Tract Peak Gradient; LVOT Pk Vel, Left Ventricular Outflow Tract Peak Velocity; LV Wall Thickness, Left Ventricular Wall Thickness; MV Regurgitation, Mitral Valve Regurgitation; MV Stenosis, Mitral Valve Stenosis; MV Structure, Mitral Valve Structure.


Generation of Finetuning Examples
Post-2016 reports (3000)
Post-2016 structured data (3000 rows)
Pre-2016 report examples (5)
LLM
Llama2 70b
Post-2016 reports in pre-2016 style (3000)
Model Finetuning
Post-2016 reports in pre-2016 style (2000)
Post-2016 structured data (2000 rows)
LLM
Llama2 13b
LLM
HeartDx-LM
Model Evaluation
1.YNHHS post-2016 TTE reports (10,000)
2.YNHHS pre-2016 TTE reports (100)
3.MIMIC-III TTE reports (100)
4. MIMIC-IV TTE reports (100)
LLM
HeartDx-LM
1.YNHHS post-2016 structured data (10,000 rows)
2.YNHHS pre-2016 structured data (100 rows)
3.MIMIC-III TTE structured data (100 rows)
4. MIMIC-IV structured data (100 rows)


Number of rows of data used for finetuning
100
200
500
1,000
2,000
% of extracted values that are correct
0
20
40
60
80
100
GLS%
IVSd
LVDiastolicFunction
AVStructure
AVStenosis
AVRegurg
AIPHT
LVOTPkVel
LVOTPkGrad
MVStructure
MVStenosis
MVRegurgitation
EF
LVWallThickness
AVPkVel(m/s)
AVMnGrad(mmHg)
AVAContVTI
AVAIndex
Columns


YNHHS Post-2016
YNHHS Pre-2016
MIMIC-III
MIMIC-IV
Clinical Variable
AIPHT
AVAContVTI
AVAIndex
AVMnGrad
AVPkVel
EF
GLS%
IVSd
LVOTPkGrad
LVOTPkVel
AVStructure
AVStenosis
AVRegurg
MVStructure
MVStenosis
MVRegurgitation
LVDiastolicFunction
LVWallThickness
Correct Extraction
Incorrect Extraction
Failed Extraction
Extraction Type
100
80
60
40
20
0