EchoQA: A Large Collection of Instruction Tuning Data for Echocardiogram Reports

Lama Moukheiber*
Massachusetts Institute of Technology

Mira Moukheiber*
Massachusetts Institute of Technology

Dana Moukheiber
Massachusetts Institute of Technology

Jae-Woo Ju
Seoul National University

Hyung-Chul Lee
Seoul National University

Abstract

We introduce a novel question-answering (QA) dataset using echocardiogram reports sourced from the Medical Information Mart for Intensive Care data. This dataset is specifically designed to enhance QA systems in cardiology, consisting of 771,244 QA pairs addressing a wide array of cardiac abnormalities and their severity. We compare various large language models (LLMs), including both open-source general models and biomedical-specific models, alongside state-of-the-art closed-source models for zero-shot evaluation. Our results show that fine-tuning LLMs improves performance across various QA metrics, highlighting the validity and value of our dataset. Further, we conduct fine-grained fairness audits to assess the bias-performance trade-off of LLMs across marginalized populations. Our objective is to propel the field forward by establishing a benchmark framework for developing LLM AI agents that support clinicians in their daily workflow within the cardiology space. The dataset aims to support the advancement of natural language models for use in diagnostic decision support systems, aiming to increase efficiency in cardiology care.

1 Introduction

Echocardiography is the most prevalent noninvasive technique for assessing heart function and detecting heart diseases. It plays a critical role in clinical cardiology, consistently guiding decision-making processes [1]. Echocardiography is essential for diagnosing diseases, stratifying risks, and evaluating treatment efficacy. The diagnostic reports generated from these tests provide rich clinical data, vital for diagnosing and managing various cardiac conditions [2]. The growing demand for diagnostic echocardiograms makes it difficult to manage and interpret the increasing volume of data, and utilizing AI-powered algorithms can reduce clinician workload.

The advent of large language models (LLMs) holds the potential to transform the field of cardiology. LLMs have been utilized across various natural language processing tasks, such as question-answering (QA), text summarization, and language translation, often in zero- and few-shot scenarios [3, 4]. In-context learning (ICL) enables the models to tackle new tasks with only a few task demonstrations, like in three-shot prompting, without the need to update model parameters [4]. Moreover, converting natural language understanding and generation tasks into instructional inputs enhances LLMs' ability

*Equal contribution.

SafeGenAI Workshop @ 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

to follow domain-specific instructions and improve downstream task performance [5, 6]. Open-source models like Llama [7] and Mistral [8] have shown great potential. Leveraging these models with high-quality instruction-following samples from echocardiogram diagnostic reports is important to streamlining healthcare workflows. Such systems would efficiently handle the complexity of data, reduce clinicians' cognitive load, and facilitate faster decision-making.

A significant challenge has been developing large language models (LLMs) that are trained and evaluated on real echocardiogram reports with ground-truth answers, rather than relying on synthetic data or data from medical licensing exams [9, 10]. This limitation has hindered progress of AI in the cardiology space. However, with the recent advancements in instruction-following capabilities of LLMs, there is an opportunity to bridge the gap between raw data and actionable medical insights and assist with clinical reasoning and knowledge recall. A high-quality question-answering dataset using echocardiogram reports could serve as a benchmark and allow the training of specialized cardiology models, reducing healthcare providers' cognitive load while streamlining workflows to increase efficiency.

Addressing algorithmic bias is crucial in healthcare before model deployment. Utilizing protective attributes, such as race, gender and age can improve the effectiveness of fairness audits in healthcare algorithms [11, 12]. Beyond these common attributes, analyzing social factors such as disability and lifestyle behaviors could assist healthcare providers identify and mitigate disparities in algorithmic outcomes. These living conditions offer valuable insights for addressing gaps in medical care. Such non-clinical factors, commonly known as social determinants of health (SDOH), are vital for conducting thorough audits of algorithmic biases [13]. Ensuring that algorithms account for the broader context of patients' lives promotes equitable healthcare, especially for marginalized groups [14]. Incorporating these data into fairness audits also aids in complying with regulations like Section 1557 of the Affordable Care Act, which mandates that healthcare providers and payers ensure their algorithms do not discriminate [15].

Based on the challenges aforementioned, our work makes the following three contributions:

- *Development of EchoQA:* We present EchoQA, the largest open-access, real patient question-answering dataset for echocardiography, meticulously developed by expert clinicians. Our aim is to propel the medical field by creating a foundation for training LLM-based AI agents that will assist cardiologists in their daily workflows. EchoQA offers a robust dataset that offers researchers and practitioners to test and compare different machine learning approaches for disease diagnosis and management.
- *Zero-shot, Few-shot and Instruction Fine-Tuning Evaluations:* Leveraging the EchoQA dataset, we validate its utility by fine-tuning a variety of LLMs, encompassing both general-purpose and medical-domain models, and comparing their performance to zero-shot setups. Additionally, for comparison we conduct zero-shot and 3-shot evaluations on commercial LLMs. Furthermore, we release the best-performing echocardiogram model, *Echo-Mistral*, making it accessible to the wider research community.
- *Fairness Audits on Social Determinants of Health:* To investigate algorithmic bias, we use social determinants of health to enable more fine-grained audits of algorithmic fairness. These evaluations provide critical insights into potential disparities often overlooked in LLM studies, ensuring a more equitable and inclusive approach to medical artificial intelligence.

2 Related Work

Medical question answering datasets. Medical question-answering benchmark datasets have been developed to address different aspects of medical information retrieval and understanding. Examples include datasets designed for medical licensing exams and conceptual medical knowledge, such as MedQA, JAMA Clinical Challenge, MedBullet, and MMLU Clinical Topics [9, 16, 17]. Additionally, literature-based QA datasets, such as PubMedQA, consist of biomedical research questions derived from PubMed abstracts [18]. On the other hand, datasets like HealthSearchQA, LiveQA, and MedicationQA provide insights into medical information needs from a consumer perspective [19–21]. More specifically, MedicationQA addresses questions related to medications and their uses, aiding in pharmaceutical information retrieval [21]. QA datasets utilizing real-world medical data from electronic health records include emrQA, which consists of factual questions with answers

derived from discharge summary reports in the i2b2 dataset [22]. Similarly, RadQA consists if radiology-related questions commonly encountered in clinical practice, using data extracted from radiology reports in the MIMIC database [23]. However, none of these datasets include questions a echocardiogram reports, which differ significantly in semantic content and vocabulary. Table 2 provides a summary of the medical QA datasets described above.

Table 1: Overview of medical Question-Answering (QA) datasets by domain, QA type, and use of real-world data. QA types include Multiple Choice (predefined answers) and Long-form (free-text responses).

Dataset	**Domain**	**QA Type**	**Real-world medical data**
MedQA [9]	Medical Board Exams (USMLE)	Multiple choice	
JAMA Clinical Challenge [17]	Exam for clinical cases (JAMACC)	Multiple choice	
MedBullet [17]	Medical Board Exams (USMLE)	Multiple choice	
MMLU Clinical Topics [9]	Medicine and biology-related topics	Multiple choice	
PubMedQA [18]	Literature-based (PubMed abstracts)	Multiple choice	
HealthSearchQA [19]	Consumer searched questions	Long-form	✓
LiveQA [20]	Consumer health	Long-form	✓
MedicationQA [21]	Consumer questions about medications	Long-form	✓
emrQA [22]	Discharge reports (i2b2 data)	Long-form	✓
RadQA [23]	Radiology reports (MIMIC data)	Long-form	✓
EchoQA (Ours)	Echocardiography reports (MIMIC data)	Long-form	✓

LLM and echocardiography. There is limited research on the use of large language models (LLMs) specifically within cardiology. One work introduced EchoGPT, a fine-tuned Llama-2 model [7] employing Quantized Low-Rank Adaptation (QLoRA) to assist with echocardiography report summarization and initial drafting of reports for clinician review, effectively streamlining the reporting workflow [24]. Further, prior studies indicate that general-purpose LLMs, such as ChatGPT, struggle with echocardiography board review questions, highlighting the need for specialized training to enhance performance in cardiology applications [25].

Fairness audits. While progress has been made in addressing algorithmic fairness in healthcare, most studies have focused primarily on biases related to protected attributes such as age, gender, and race [11, 26, 27]. Recent research emphasizes the need to examine biases from more multidimensional perspectives, particularly by analyzing the social processes that contribute to these biases. Evaluating fairness through the lens of intersectional social identities provides a deeper understanding of the socially constructed nature of attributes like race and gender. Incorporating social determinants offers valuable insights into the processes driving disparities in machine learning models [14]. Furthermore, conducting bias audits centered on these factors is more practical, as they are not just social constructs but modifiable aspects [28, 29]. We leverage social determinants of health to conduct fine-grained audits of algorithmic fairness on biomedical and closed source LLMs. With that, we hope to ensure that the models are equitable and account for the broader context of individuals' lives.

3 Experimental Setup

3.1 Dataset

We curate a comprehensive question-answering dataset sourced from the Medical Information Mart for Intensive Care (MIMIC-IV) database, which is a fully de-identified clinical dataset comprising over 80,000 echocardiogram reports collected at Beth Israel Deaconess Medical Center between 2012-2019 [30]. The echocardiogram reports include details on specific heart structures, such as the left atrium, right atrium/interatrial septum, left ventricle, right ventricle, mitral valve, aortic valve, and tricuspid valve. Each patient's echocardiography report is processed to extract unique sentences for each heart structure.

Following [31], clinical experts identify abnormalities in these heart structures, and assign levels ranging from -3 to 3 based on the standardized diagnostic criteria established by the American Society of Echocardiography indicating the category and severity level of the heart structure abnormality.

[32–34]. A category of -3 indicates that the study is inadequate for evaluating the cardiac abnormality. A category of 0 applies when the study is adequate for evaluating cardiac function but reveals no

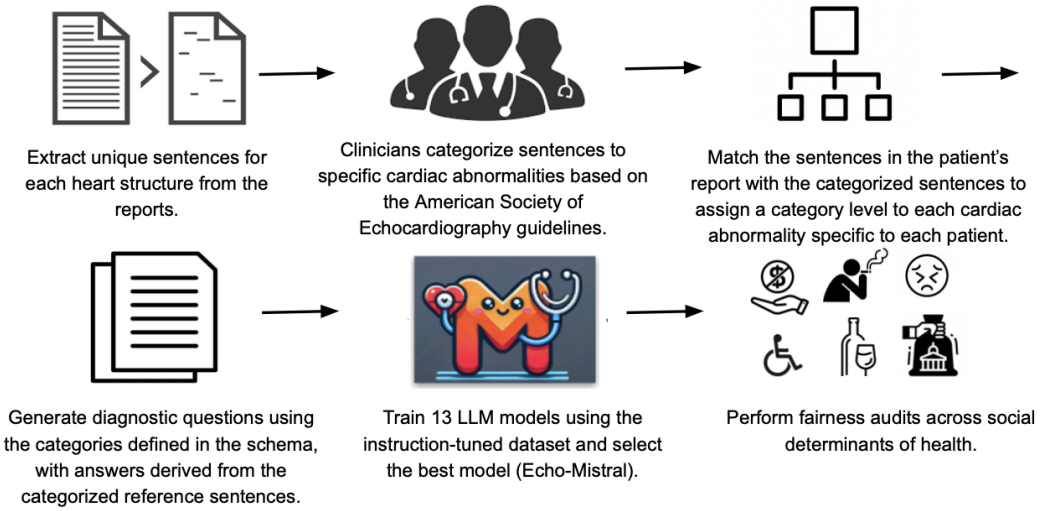

Figure 1: Workflow of the methodology.

abnormalities. Sentences describing abnormal function without specifying severity are assigned a category of -2. Severity levels are categorized as 1 for mild, 2 for moderate, and 3 for severe. For specific features, such as left ventricular cavity size, left ventricular systolic function, and right ventricular cavity size, a category of -1 indicates hyperdynamic left ventricular systolic function or a small cavity size for the left or right ventricle. The sentences in the patient's notes are then matched with the categorized reference sentences to assign the corresponding category levels to each cardiac abnormality. When multiple sentences for the same structure match different severity levels—mild, moderate, or severe—the highest category level is retained to prioritize the most clinically significant finding. In cases where conflicting category levels are identified for the same structure and sentence (e.g., two matching sentences with differing levels), a placeholder value of -50 is assigned to indicate ambiguity or disagreement in the categorization. This process ensures consistency while accounting for uncertainties in sentence matching.

As illustrated in Figure 2, using these categories, diagnostic questions, such as "Is the study adequate to assess left ventricular systolic dysfunction?" are generated. The answers to these questions are derived directly from the categorized reference sentences, resulting in more than 700,000 question-answer pairs, with the categories depicted in Table 2. Evaluating cardiac parameters from diagnostic reports is crucial in clinical practice, as these parameters provide essential information about the heart's structure and function, which is vital for diagnosing and managing various cardiovascular diseases. The dataset will be hosted on a PhysioNet, an NIH-funded health data repository [35].

3.2 Model Inference & Training

We evaluate model performance across three experimental setups: zero-shot, three-shot, and fine-tuning setups. These experiments are conducted on open-source biomedical models, open-source general models, and closed-source models. We employ supervised fine-tuning (SFT) to fine-tune our models using a diverse selection from recent open-source and biomedical domain-specific large language models (LLMs) to validate the value of the training data compared to zero-shot or three-shot performance. From the open-source category, we utilize Llama-3-8B [7], Mistral-7B [8], Phi-3-mini [36], and Zeyphr-7B [37], and Falcon-7B [38] and compare fine-tuning with zero-shot performance. For the biomedical open-source models, we incorporate BioMistral-7B [39], M42-health [40], PMC-LLaMa-13B [41], and Meditron-7B [42] aiming to leverage their deep understanding of biomedical terminology and context derived from medical abstracts and texts and compare fine-tuning with zero-shot performance. Additionally, we use the propriety models, including Amazon-titan [43], Claude [44], Cohere [45], GPT-4o [46] models for zero-shot and three-shot evaluation. These models offer advanced capabilities in text generation and comprehension tasks without further fine-tuning. The closed-source models are run on Azure OpenAI[47] and Amazon Bedrock [48]to ensure HIPAA compliance and patient privacy.

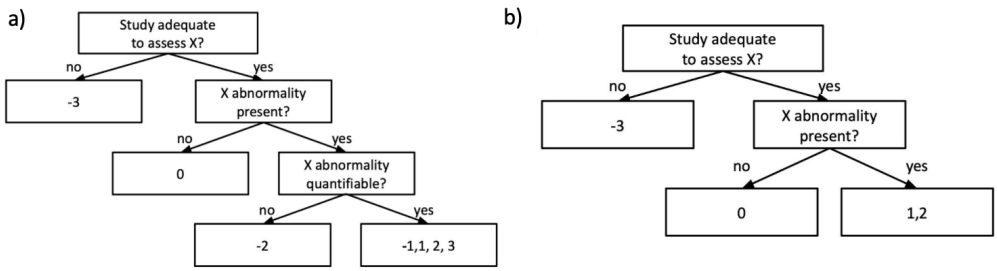

Figure 2: Categorization of cardiac abnormalities. X represents a specific cardiac abnormality. a) The schema includes the following cardiac abnormalities: right atrial pressure; tricuspid valve regurgitation, tricuspid valve stenosis, and pulmonary hypertension; right ventricular systolic function, right ventricular cavity, and right ventricular wall; left atrial cavity; mitral valve regurgitation and mitral valve stenosis; left ventricular systolic function, left ventricular cavity, left ventricular wall, left ventricular diastolic function, left ventricular outflow tract obstruction, and left regional wall motion abnormality; and aortic valve regurgitation and aortic valve stenosis. b) The schema includes other right ventricular and atrial abnormalities: right ventricular pressure overload and right ventricular volume overload; and right atrial enlargement.

Table 2: Cardiac abnormalities found in the echocardiogram reports.

Cardiac Abnormalities	**Number of QA's**
Right atrial abnormalities	
Right atrial enlargement	45,254
Right atrial pressure	2,371
Tricuspid valve abnormalities	
Tricuspid valve regurgitation	13,332
Tricuspid valve stenosis	19,509
Pulmonary hypertension	21,376
Right ventricular abnormalities	
Right ventricular systolic function	74,236
Right ventricular cavity	71,971
Right ventricular volume overload	5,075
Right ventricular pressure overload	5,065
Right ventricular wall	7,316
Left atrial abnormalities	
Left atrium cavity	14,425
Mitral valve abnormalities	
Mitral valve stenosis	38,044
Mitral valve regurgitation	53,205
Left ventricular abnormalities	
Left ventricular systolic function	64,305
Left ventricular cavity	64,354
Left ventricular wall	64,295
Left ventricular diastolic function	5,769
Left ventricular outflow tract obstruction	40,697
Left regional wall motion abnormality	39,310
Aortic valve abnormalities	
Aortic valve stenosis	61,451
Aortic valve regurgitation	59,884
Total	**771,244**

Due to computational limitations, we sample 10,000 subjects and divide the data into 70% training, 10% validation, and 20% testing, ensuring that the data of each subject appears in only one group. The training set is used for fine-tuning the model, while the test set is used for inference and evaluation.

Training and evaluation datasets are processed and tokenized, and the model is configured with dynamic token embedding resizing to accommodate task-specific tokens effectively. The prompts used for 0-shot and 3-shot questions answering are shown in Figure 5 and Figure 6 in the Appendix.

For supervised fine-tuning, we train the models for one epoch with a learning rate of 2e-4, using a cosine learning rate schedule with a 1% warm-up ratio to stabilize the initial training phase. Training is performed using the AdamW optimizer, leveraging gradient accumulation steps of 4 to simulate larger batch sizes and optimize memory usage. To enhance training efficiency, we employ Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. This includes setting the scaling factor (LoRA alpha) to 16 to control the influence of task-specific adaptations, applying a 10% dropout rate to prevent overfitting, and using a rank of 64 to enable task-specific adaptation with minimal additional parameters. Additionally, we employ the BitsAndBytes (BnB) quantization technique with Normal Float 4 (NF4) for numerical stability and Brain Floating Point 16-bit (bfloat16) for faster computations. Fine-tuning is conducted on NVIDIA A100 80GB GPUs, utilizing model sharding to efficiently distribute computational resources.

3.3 Model Evaluation & Fairness Audits

We conduct a comprehensive analysis of our model's performance using a suite of well-established metrics. We employ BLEU score to measure the precision of n-grams between the generated and reference answers, providing insights into the accuracy of the text generation [49]. To assess the balance between precision and recall, we utilize the average F1 Score [50]. We apply the ROUGE-1 and ROUGE-2 metrics to evaluate the overlap of unigrams and bigrams, respectively, between the generated and reference answers, thereby gauging the lexical similarity at different granularities [51]. Additionally, we use the ROUGE-L metric to measure the longest common subsequence, indicating the extent to which the generated answer aligns with the reference in terms of sequence matching [52]. Lastly, we utilize the average METEOR Score, which evaluates precision and recall while considering synonyms and stemming [53].

Futhermore, we perform fairness audits by examining social health attributes, as these factors provide insights into the conditions in which individuals live —critical influences on a person's health and well-being. To perform these audits, we utilize census tract-level SDOH data from the MIMIC dataset [54]. Our analysis investigates fairness disparities across subgroups defined by societal attributes, such as whether a patient lives in areas with high unemployment rates, relies heavily on public assistance or food stamps, includes adults who are heavy drinkers or smokers, or reports experiencing mental distress or having a disability. We discretize the dimensions into high, upper middle, lower middle, low groups, based on the quantile of the distribution for each dimension. We select the best-performing model configuration to evaluate bias. For each LLM, bias across various dimensions is assessed using the F1 equality difference metric in [55], which measures the average absolute difference between the f1 of individual social groups and the overall f1 across all groups within the corresponding social category. In particular, for a dimension D and its associated set of demographic groups $\mathcal{G}^D = \{\mathcal{G}_1^D, \mathcal{G}_2^D, \ldots\}$, F1 equality difference $= \frac{1}{|\mathcal{G}^D|} \sum_{\mathcal{G}_i^D \in \mathcal{G}^D} |F1(\mathcal{G}_i^D) - F1(\mathcal{G}^D)|$.

4 Results & Discussion

Table 3 presents the performance metrics across various models, including open-source fine-tuned biomedical models, open-source fine-tuned general models, and closed-source general models.

Fine-tuning on the EchoQA dataset significantly enhances model performance, with fine-tuned open-source models achieving higher scores than their 0-shot settings and, outperforming both closed-source models in 3-shot learning scenarios and biomedical-specific models, validating the value of domain-specific training data. Among open-source models, Mistral-7B (fine-tuned) demonstrates better overall performance across different metrics, compared to biomedical models such as BioMistral-7B and general models like Zephyr-7B. Closed-source models like GPT-4 (3-shot) also perform well, but fine-tuned Mistral-7B achieves higher scores across different metrics. We name the best performing model, *Echo-Mistral*, and release the corresponding model weights publicly.[2] [3]

[2]Echo-Mistral model weights can be found here: `https://huggingface.co/lamamkh/echomistral`.

[3]Code can be found at: `https://github.com/Mira-MM/echomistral`.

Table 3: Performance metrics for open-source biomedical models, open-source general models, and closed-source general models averaged across 3 runs (higher scores with tighter error margins are better). Fine-tuned open source models are compared to their baseline zero-shot models. Closed-source models are compared to 0-shot and 3-shot in-context learning models. Bolded numbers depict the best model in each of the categories including the open-source models (biomedical and general) and closed-source models.

Evaluation Metric	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	F1	METEOR
Open-source (biomedical)						
BioMistral-7B (0-shot)	0.03957 ± 0.00000	0.24143 ± 0.00000	0.12638 ± 0.00000	0.22213 ± 0.00000	0.25368 ± 0.00000	0.32930 ± 0.00000
BioMistral-7B (fine-tuned)	**0.68290 ± 0.00446**	**0.96694 ± 0.00165**	**0.96100 ± 0.00222**	**0.96659 ± 0.00174**	**0.96699 ± 0.00168**	**0.95483 ± 0.00188**
M42-health (0-shot)	0.18000 ± 0.00110	0.60429 ± 0.00071	0.50154 ± 0.00137	0.57618 ± 0.00044	0.60747 ± 0.00068	0.64191 ± 0.00060
M42-health (fine-tuned)	0.31223 ± 0.03710	0.71158 ± 0.07130	0.64302 ± 0.09202	0.69965 ± 0.07748	0.71220 ± 0.07127	0.70233 ± 0.07844
Meditron7B (0-shot)	0.00023 ± 0.00006	0.08442 ± 0.00244	0.02637 ± 0.00084	0.07909 ± 0.00216	0.08637 ± 0.00269	0.06323 ± 0.00235
Meditron7B (fine-tuned)	0.00104 ± 0.00036	0.11387 ± 0.02270	0.02735 ± 0.00714	0.10594 ± 0.02103	0.11666 ± 0.02126	0.08680 ± 0.01295
PMC-llama-13B (0-shot)	0.00287 ± 0.00033	0.08877 ± 0.00133	0.01519 ± 0.00052	0.08428 ± 0.00147	0.09030 ± 0.00140	0.06801 ± 0.00078
PMC-llama-13B (fine-tuned)	0.00478 ± 0.00153	0.09885 ± 0.01602	0.02617 ± 0.01106	0.09409 ± 0.01377	0.10138 ± 0.01725	0.08549 ± 0.02312
Open-source (general)						
Llama-8B (0-shot)	0.07216 ± 0.00139	0.27047 ± 0.00141	0.21111 ± 0.00144	0.26227 ± 0.00117	0.29686 ± 0.00145	0.48359 ± 0.00193
Llama-8B (fine-tuned)	0.62419 ± 0.02431	0.95415 ± 0.00374	0.94527 ± 0.00499	0.95408 ± 0.00381	0.95419 ± 0.00373	0.93504 ± 0.00543
Mistral-7B (0-shot)	0.06410 ± 0.00000	0.30277 ± 0.00000	0.19364 ± 0.00000	0.28000 ± 0.00000	0.31844 ± 0.00000	0.46702 ± 0.00000
Mistral-7B (fine-tuned)	0.70062 ± 0.00185	**0.97942 ± 0.00042**	**0.97464 ± 0.00054**	**0.97913 ± 0.00038**	**0.97944 ± 0.00042**	**0.96519 ± 0.00048**
Phi-mini (0-shot)	0.03595 ± 0.00000	0.24040 ± 0.00000	0.10830 ± 0.00000	0.21494 ± 0.00000	0.24707 ± 0.00000	0.28078 ± 0.00000
Phi-mini (fine-tuned)	0.63524 ± 0.00333	0.91028 ± 0.00401	0.89615 ± 0.00386	0.91022 ± 0.00396	0.91033 ± 0.00400	0.89148 ± 0.00450
Zephyr-7B (0-shot)	0.05772 ± 0.00000	0.24476 ± 0.00000	0.16749 ± 0.00000	0.22953 ± 0.00000	0.26266 ± 0.00000	0.41747 ± 0.00000
Zephyr-7B (fine-tuned)	**0.70221 ± 0.00125**	0.97877 ± 0.00212	0.97542 ± 0.00245	0.97875 ± 0.00210	0.97879 ± 0.00212	0.96557 ± 0.00199
Falcon-7B (0-shot)	0.00057 ± 0.00022	0.05637 ± 0.00083	0.01056 ± 0.00057	0.05242 ± 0.00116	0.05917 ± 0.00067	0.06020 ± 0.00054
Falcon-7B (fine-tuned)	0.08515 ± 0.00351	0.27606 ± 0.00347	0.21054 ± 0.00533	0.26710 ± 0.00413	0.27853 ± 0.00349	0.44261 ± 0.00760
Closed-source (general)						
Amazon-titan (0-shot)	0.22389 ± 0.00007	0.63861 ± 0.00033	0.53226 ± 0.00028	0.60887 ± 0.00035	0.64112 ± 0.00031	0.68238 ± 0.00043
Amazon-titan (3-shot)	0.22864 ± 0.00566	0.68582 ± 0.04277	0.60189 ± 0.06315	0.66289 ± 0.04880	0.68741 ± 0.04204	0.70235 ± 0.02002
Claude (0-shot)	0.08953 ± 0.00037	0.32158 ± 0.00074	0.26104 ± 0.00104	0.31722 ± 0.00070	0.34341 ± 0.00067	0.54070 ± 0.00120
Claude (3-shot)	0.06801 ± 0.00041	0.28128 ± 0.00092	0.22510 ± 0.00119	0.27354 ± 0.00111	0.30281 ± 0.00097	0.50598 ± 0.00220
Cohere (0-shot)	0.07602 ± 0.00444	0.39138 ± 0.00819	0.25881 ± 0.00627	0.36379 ± 0.00725	0.39745 ± 0.00802	0.43481 ± 0.01015
Cohere (3-shot)	0.20385 ± 0.00060	0.65451 ± 0.00099	0.58761 ± 0.00127	0.64104 ± 0.00112	0.65501 ± 0.00099	0.66083 ± 0.00055
GPT-4o (0-shot)	0.20759 ± 0.00917	0.57788 ± 0.00885	0.48196 ± 0.00581	0.56247 ± 0.00619	0.58605 ± 0.00879	0.63698 ± 0.00049
GPT-4o (3-shot)	**0.34675 ± 0.00425**	**0.84262 ± 0.00410**	**0.81498 ± 0.00533**	**0.84081 ± 0.00413**	**0.84294 ± 0.00409**	**0.81998 ± 0.00485**

Table 4: Overall bias along six social determinants of health for open-source biomedical models, open-source general models, and closed-source general models averaged across 3 runs (lower scores with tighter error margins are better). Bolded numbers depict the least biased model per dimension. Underlined numbers depict the second least biased model per dimension.

	Social Determinants of Health					
	% of population with a disability	**% of households receiving public assistance**	**% of the population that is unemployed**	**% of heavy drinking adults**	**% of adults reporting 14+ days of poor mental health per month**	**% of current adults smokers**
Open-source finetuned biomedical models						
Biomistral-7B	0.00901 ± 0.00100	0.00688 ± 0.00210	0.00738 ± 0.00030	0.01121 ± 0.00486	0.00759 ± 0.00440	0.00622 ± 0.00042
M42-health	0.00928 ± 0.00292	0.01690 ± 0.00564	0.02084 ± 0.00267	0.02213 ± 0.00579	0.01900 ± 0.00680	0.01938 ± 0.00669
Meditron	0.00437 ± 0.00264	**0.00349 ± 0.00053**	0.00566 ± 0.00232	0.00874 ± 0.00358	0.00448 ± 0.00144	0.00696 ± 0.00101
PMC-llama-13B	0.00515 ± 0.00101	0.00511 ± 0.00160	0.00399 ± 0.00189	0.00808 ± 0.00284	0.00934 ± 0.00310	0.00784 ± 0.00210
Open-source finetuned general models						
Llama-8B	0.01747 ± 0.00176	0.00818 ± 0.00115	0.01154 ± 0.00084	0.00651 ± 0.00162	0.01047 ± 0.00055	0.00572 ± 0.00232
Mistral-7B	**0.00360 ± 0.00125**	0.00358 ± 0.00073	0.00327 ± 0.00139	0.00643 ± 0.00058	0.00591 ± 0.00191	0.00585 ± 0.00177
Phi-mini	0.00996 ± 0.00301	0.01340 ± 0.00113	0.00613 ± 0.00287	0.01538 ± 0.00090	0.01245 ± 0.00078	0.00999 ± 0.00090
Zephyr-7B	0.00405 ± 0.00033	0.00504 ± 0.00148	0.00649 ± 0.00167	**0.00370 ± 0.00071**	**0.00376 ± 0.00065**	**0.00310 ± 0.00107**
Falcon-7B	0.01081 ± 0.00119	0.00561 ± 0.00140	**0.00331 ± 0.00108**	0.00852 ± 0.00247	0.01343 ± 0.00511	0.00974 ± 0.00167
Closed-source general models						
Amazon-titan (3-shot)	0.01579 ± 0.00742	0.02201 ± 0.00950	0.01396 ± 0.00565	0.02274 ± 0.00726	0.01838 ± 0.01001	0.01644 ± 0.00510
Claude (3-shot)	0.30259 ± 0.00815	0.30273 ± 0.00835	0.30145 ± 0.00403	0.30510 ± 0.01168	0.30357 ± 0.00815	0.30056 ± 0.01417
Cohere (3-shot)	0.65874 ± 0.01588	0.66381 ± 0.02542	0.66127 ± 0.01287	0.65936 ± 0.02080	0.65355 ± 0.02268	0.65433 ± 0.02111
GPT-4o (3-shot)	0.84429 ± 0.01620	0.84423 ± 0.01818	0.84304 ± 0.01652	0.84573 ± 0.02880	0.83464 ± 0.01678	0.83799 ± 0.01876

As seen in Table 4, Mistral-7B, the best-performing model, demonstrates the least bias in representing people with disabilities while achieving comparable performance across other socio-demographic dimensions, including public assistance, unemployment, heavy drinking, mental health, and smoking.

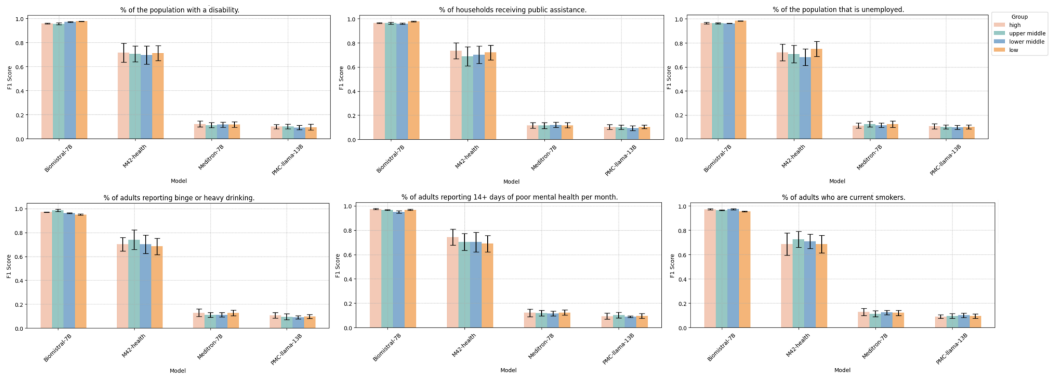

Figure 3: Disparities in performance depicted by F1 and standard error over 3 runs between different groups (high, upper middle, lower middle, low) along the social determinants of health by each examined open-sourced biomedical LLM.

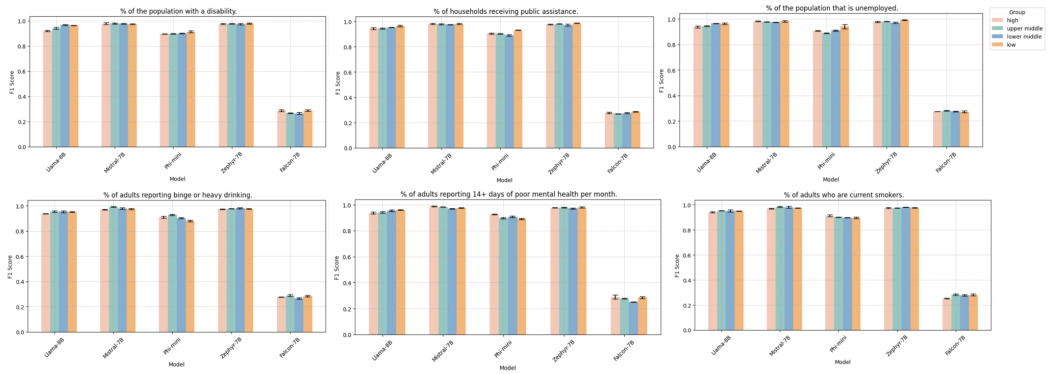

Figure 4: Disparities in performance depicted by F1 and standard error over 3 runs between different groups (high, upper middle, lower middle, low) along the social determinants of health by each examined open-sourced general LLM.

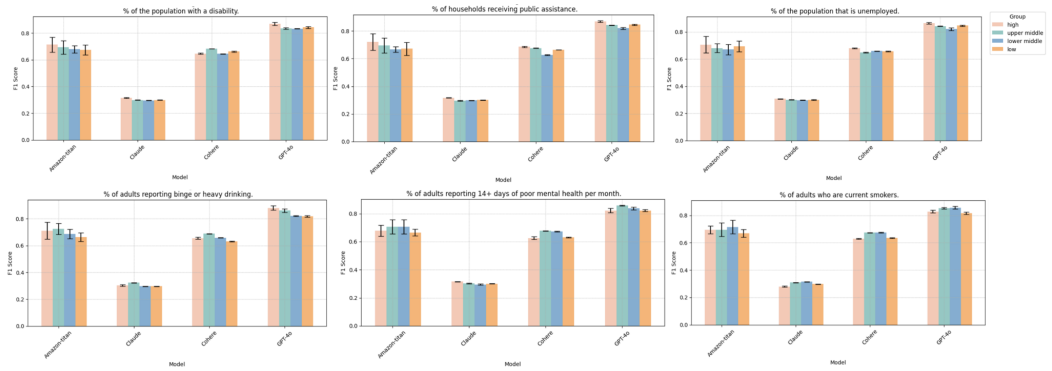

Figure 5: Disparities in performance depicted by F1 and standard error over 3 runs between different groups (high, upper middle, lower middle, low) along the social determinants of health by each examined closed-sourced general LLM.

This highlights its ability to navigate the trade-off between minimizing bias and maintaining performance. However, there are biases across certain groups, where models like Falcon-7B, Meditron, and Zephyr-7B slightly outperform Mistral-7B in specific dimensions, such as unemployment , public

assistance, and heavy drinking and mental health. These variations highlight the need for mitigation strategies before deployment to address any remaining biases and ensure fair and equitable use of the model in healthcare applications.

On a fine-grained level, as seen in Figure 3, the disparity in F1 scores among the four groups (high, upper-middle, lower-middle, and low) within each open-source biomedical model is comparable when we compare groups for a given model. This pattern is consistent across all social determinants of health for the open-source biomedical models. For open-source general models, as seen in Figure 4, Phi-mini attains a higher F1 for the households with low level of public assistance compared to the other groups while the disparity in F1 among the four groups with different levels of public assistance is more moderate for the other open-source models. Moreover, Falcon-7B shows slightly lower F1 for adults who are heavy smokers compared to the other groups. Additionally, Llama-8B depicts lower performance for populations with a high percentage of individuals with disabilities, and unemployed individuals. The best-performing model in terms of overall F1 score, Mistral-7B, demonstrates moderate disparities among the four groups across all social determinants of health. Finally, for closed source general models, Figure 5 shows that GPT-4o achieves a higher F1 score for the high group compared to the low group across various social determinant attributes, including the percentage of the population with disabilities, percentage of households receiving public assistance, percentage of the population unemployed, and percentage of adults reporting binge or heavy drinking.

5 Conclusion

In this paper, we introduce a novel and comprehensive question-answeing dataset using the MIMIC echocardiogram reports. This dataset is designed to enhance QA systems within the cardiology domain. To demonstrate the dataset's utility, we validated it using 13 LLMs, showing that the instruction fine-tuned Mistral-7B open-source model performs better than biomedical-specific models and general state-of-the-art closed-source model. Given Mistral-7B performed the best, we named our best fine-tuned model *Echo-Mistral*. Our fairness audit reveals variability in model performance across different social and marginalized communities, highlighting the trade-off between performance and fairness. We hope our comprehensive benchmark, featuring multiple LLMs and various evaluation metrics, will serve as a baseline, facilitating progress in medical real-world question-answering tasks in the cardiology space. Figure depicts the workflow of the paper.

Ethics

The dataset originates from the MIMIC-IV database, which is a de-identified dataset accessed through the PhysioNet Credentialed Health Data Use Agreement (v1.5.0) that we have been granted permission to use. The ethics approval of the dataset follows from that of the parent MIMIC dataset.

Acknowledgements

This work was supported in part by the Falcon 40B Challenge, an initiative by Abu Dhabi's Technology Innovation Institute (TII). H.C. is supported by the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant number: HI21C107409).

References

[1] Sherif F Nagueh, Otto A Smiseth, Christopher P Appleton, Benjamin F Byrd, Hisham Dokainish, Thor Edvardsen, Frank A Flachskampf, Thierry C Gillebert, Allan L Klein, Patrizio Lancellotti, et al. Recommendations for the evaluation of left ventricular diastolic function by echocardiography: an update from the american society of echocardiography and the european association of cardiovascular imaging. *European Journal of Echocardiography*, 17(12):1321–1360, 2016.

[2] Roberto M Lang, Luigi P Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A Flachskampf, Elyse Foster, Steven A Goldstein, Tatiana Kuznetsova, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. *European Heart Journal-Cardiovascular Imaging*, 16(3):233–271, 2015.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[6] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[7] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[8] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[9] Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[10] Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. Publicly shareable clinical large language model built on synthetic clinical notes. *arXiv preprint arXiv:2309.00237*, 2023.

[11] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[12] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[13] R Wilkinson. Social determinants of health: the solid facts. *World Health Organization Regional Office for Europe*, 2003.

[14] Mira Moukheiber, Lama Moukheiber, Dana Moukheiber, and Hyung-Chul Lee. Unmasking societal biases in respiratory support for icu patients through social determinants of health.

[15] Michael P Cary Jr, Anna Zink, Sijia Wei, Andrew Olson, Mengying Yan, Rashaud Senior, Sophia Bessias, Kais Gadhoumi, Genevieve Jean-Pierre, Demy Wang, et al. Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: A scoping review: Scoping review examines racial and ethnic bias in clinical algorithms. *Health Affairs*, 42(10):1359–1368, 2023.

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[17] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*, 2024.

[18] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[19] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[20] Qianying Liu, Sicong Jiang, Yizhong Wang, and Sujian Li. Liveqa: A question answering dataset over sports live. In *Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19*, pages 316–328. Springer, 2020.

[21] Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. Bridging the gap between consumers' medication questions and trusted answers. In *MedInfo*, pages 25–29, 2019.

[22] Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*, 2018.

[23] Sarvesh Soni, Meghana Gudala, Atieh Pajouhi, and Kirk Roberts. Radqa: A question answering dataset to improve comprehension of radiology reports. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6250–6259, 2022.

[24] MD Chieh-Ju Chao, Imon Banerjee, MD Andrew Tseng, Garvan C Kane, and Chieh-Ju Chao. Echogpt: A large language model for echocardiography report summarization.

[25] Achille Sowa and Robert Avram. Fine tuned large language models can generate expert-level echocardiography reports, 2024.

[26] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

[27] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on health, inference, and learning*, pages 204–233. PMLR, 2022.

[28] Paula Braveman and Laura Gottlieb. The social determinants of health: it's time to consider the causes of the causes. *Public health reports*, 129(1_suppl2):19–31, 2014.

[29] Min Chen, Xuan Tan, and Rema Padman. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. *Journal of the American Medical Informatics Association*, 27(11):1764–1773, 2020.

[30] A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). `https://doi.org/10.13026/1n74-ne17`, 2023.

[31] Gloria Hyunjung Kwak, Dana Moukheiber, Mira Moukheiber, Lama Moukheiber, Sulaiman Moukheiber, Neel Butala, Leo Anthony Celi, and Christina Chen. Echonotes structured database derived from mimic-iii (echo-note2num), February 2024. URL `https://www.physionet.org/content/echo-note-to-num/1.0.0/`.

[32] Thomas R Porter, Sasha K Shillcutt, Mark S Adams, Georges Desjardins, Kathryn E Glas, Joan J Olson, and Richard W Troughton. Guidelines for the use of echocardiography as a monitor for therapeutic intervention in adults: a report from the american society of echocardiography. *Journal of the American Society of Echocardiography*, 28(1):40–56, 2015.

[33] Lawrence G Rudski, Wyman W Lai, Jonathan Afilalo, Lanqi Hua, Mark D Handschumacher, Krishnaswamy Chandrasekaran, Scott D Solomon, Eric K Louie, and Nelson B Schiller. Guidelines for the echocardiographic assessment of the right heart in adults: a report from the american society of echocardiography: endorsed by the european association of echocardiography, a registered branch of the european society of cardiology, and the canadian society of echocardiography. *Journal of the American society of echocardiography*, 23(7):685–713, 2010.

[34] Renato De Vecchis, Cesare Baldi, Giuseppe Giandomenico, Marco Di Maio, Anna Giasi, and Carmela Cioppa. Estimating right atrial pressure using ultrasounds: an old issue revisited with new methods. *Journal of clinical medicine research*, 8(8):569, 2016.

[35] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[36] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[37] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

[38] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

[39] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

[40] Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*, 2024.

[41] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024.

[42] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[43] Amazon Web Services. Amazon titan, 2024. URL `https://aws.amazon.com/bedrock/titan/`. Accessed: April 12, 2024.

[44] Anthropic. Introducing the next generation of claude. `https://www.anthropic.com/news/claude-3-family`, 2024. Accessed: April 12, 2024.

[45] Cohere. Cohere models documentation. `https://docs.cohere.com/v2/docs/models`, 2024. Accessed: April 12, 2024.

[46] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[47] What is the personally identifying information (pii) detection feature in azure cognitive service for language? - azure cognitive services. `https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/personally-identifiable-information/overview`, n.d. Accessed: 2022-11-24.

[48] Amazon bedrock. `https://aws.amazon.com/bedrock`, n.d. Accessed: 2024-11-30.

[49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[50] Dell Zhang, Jun Wang, and Xiaoxue Zhao. Estimating the uncertainty of average f1 scores. In *Proceedings of the 2015 International conference on the theory of information retrieval*, pages 317–320, 2015.

[51] Chin-Yew Lin and FJ Och. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir workshop*, 2004.

[52] Marcello Barbella and Genoveffa Tortora. Rouge metric evaluation for text summarization techniques. *Available at SSRN 4120317*, 2022.

[53] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[54] Ming Ying Yang, Gloria Hyunjung Kwak, Tom Pollard, Leo Anthony Celi, and Marzyeh Ghassemi. Evaluating the impact of social determinants on health prediction in the intensive care unit. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 333–350, 2023.

[55] Courtney Mansfield, Amandalynne Paullada, and Kristen Howell. Behind the mask: Demographic bias in name detection for pii masking. *arXiv preprint arXiv:2205.04505*, 2022.

A Prompt Templates

```
Below is an echocardiography report followed by a
question. Write an answer by extracting the sentence
from the report to answer the question.

"[REPORT]"
"[QUESTION]"
```

Figure 5: Zero-shot prompt provided to LLM models for question-answering.

```
Below is an echocardiography report followed by a
question. Write an answer by extracting the sentence
from the report to answer the question.


Example 1:
Report: "[REPORT EXAMPLE 1]"
Question: "[QUESTION EXAMPLE 1]"
Answer: "[ANSWER EXAMPLE 1]"

Example 2:
Report: "[REPORT EXAMPLE 2]"
Question: "[QUESTION EXAMPLE 2]"
Answer: "[ANSWER EXAMPLE 2]"

Example 3:
Report: "[REPORT EXAMPLE 3]"
Question: "[QUESTION EXAMPLE 3]"
Answer: "[ANSWER EXAMPLE 3]"

Now, answer the following question based on the report:
Report: "[REPORT]"
Question: "[QUESTION]"
```

Figure 6: Three-shot prompt provided to LLM models for question-answering.