1 2	Title: EchoGPT: A Large Language Model for Echocardiography Report Summarization
2 3 4 5 6	Authors: Chieh-Ju Chao, MD ^{1,2†} , Imon Banerjee, PhD ³ , Reza Arsanjani, MD ⁴ , Chadi Ayoub, MD, PhD ⁴ , Andrew Tseng, MD ⁵ , Jean-Benoit Delbrouck, PhD ⁶ , Garvan C. Kane, MD, PhD ¹ , Francisco Lopez-Jimenez, MD, MS ¹ , Zachi Attia, PhD ¹ , Jae K Oh, MD ¹ , Li Fei-Fei, PhD ² , Ehsan Adeli, PhD ^{2†} and Curtis Langlotz, MD, PhD ⁶
7	
8 9	[†] Chieh-Ju Chao and Ehsan Adeli are co-corresponding authors for this manuscript.
10	Affiliations:
11	¹ Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota
12	² Stanford Institute for Human-Centered Artificial Intelligence, Palo Alto, California
13	³ Department of Radiology, Mayo Clinic, Scottsdale, Arizona
14	⁴ Department of Cardiovascular Diseases, Mayo Clinic Arizona, Scottsdale, Arizona
15	⁵ Department of Cardiovascular Medicine, Mayo Clinic Florida, Jacksonville, Florida
16	⁶ Center for Artificial Intelligence in Medicine and Imaging (AIMI), Stanford University, Palo Alto,
17	California
18	
19	Address of Correspondence:
20	200 1st Street SW, Room: Gonda 4-478
21	Rochester, MN 55095, USA
22	Chieh-Ju Chao, MD
23	Assistant Professor of Medicine, Department of Cardiovascular Medicine, Mayo Clinic
24	Email: <u>chao.chiehju@mayo.edu</u>
25	Twitter: @chiehjuchao1
26	
27	
28	Keywords: Artificial Intelligence, Large Language Model, Echocardiography, Natural Language
29	Processing, Cardiovascular Imaging
30	
31	
32	
33 24	
34 25	
36	
37	
38	
39	
40	
41	
42	
43	
44	

Abbreviations:

- 2 4C metrics: qualitative review metrics including completeness, correctness, conciseness, and clinical
- 3 utility.
- 4 AI: Artificial Intelligence
- 5 DL: Deep Learning
- 6 Echo: Echocardiography
- 7 FT: Fine-Tuning
- 8 GPT: Generative Pre-trained Transformer
- 9 ICL: In-Context Learning
- 10 NLP: Natural Language Processing
- 11 Seq2seq: Sequence-to-sequence
- 12 TTE: Transthoracic Echocardiography
- 13 TEE: Transesophageal Echocardiography
- 14 LLM: Large Language Model
- 15 NLP: Natural Language Processing
- 16 QLoRA: Quantized Low-Rank Adaption

1 Abstract

2 Background

3 The increasing need for diagnostic echocardiography (echo) tests presents challenges in preserving the

- 4 quality and promptness of reports. While Large Language Models (LLMs) have proven effective in
- 5 summarizing clinical texts, their application in echo remains underexplored. To address this, we proposed
- 6 EchoGPT, a dedicated, domain specific LLM focused on echo report summarization.

7 8 Methods

- 9 Adult echo studies conducted at the Mayo Clinic from January 1, 2017, to December 31, 2017, were
- 10 collected and categorized into two groups: development (all Mayo locations except Arizona) and AZ
- 11 validation (Mayo Arizona) sets. We adapted open-source LLMs (Llama-2, MedAlpaca, Zephyr, and Flan-
- 12 T5) using In-Context Learning (ICL) and Quantized Low-Rank Adaptation (QLoRA) fine-tuning for echo
- 13 report summarization. The models' performance was assessed both quantitatively with automatic metrics
- 14 and qualitatively by cardiologists.
- 15

16 Results

- 17 The development dataset included 97,506 reports from 71,717 unique patients, predominantly male
- (55.4%), with an average age of 64.3±15.8 years. The final split contains 95,506 for training, and 1,000 18
- 19 each for validation and testing. EchoGPT, a QLoRA fine-tuned Llama-2 model, outperformed other
- 20 LLMs with win rates ranging from 87% to 99% in various automatic metrics (BLEU, METEOR,
- 21 ROUGE-L, BERT Score, and RadGraph F1 Score), and produced reports comparable to cardiologists in
- 22 30 randomly selected cases for qualitative human review (significantly preferred in conciseness (p < p
- 23 0.001), with no significant preference in completeness, correctness, and clinical utility).

24

25 Conclusions

- 26 Capable of generating echocardiography reports on par with cardiologists, EchoGPT could be used to 27 generate draft reports for human review and approval, with significant workflow advantages.
- 28

29

32

33

34

35

36

37

39

40

30 **Clinical Perspective** 31

- 1. What is new?
- This study is the first attempt to compare multiple open-source LLMs and different • model adaptation methods in echocardiography report summarization.
 - The resulting system, EchoGPT, can generate echo reports comparable in quality to • cardiologists.
 - Future metrics for echo report quality should emphasize factual correctness, especially on • numerical measurements.
- 38 2. What are the clinical implications?
 - EchoGPT system demonstrated the potential of introducing LLMs into echocardiography practice, to be used as an AI co-pilot to generate echo reports.
- 41 42
- 43
- 44

1 Introduction

2

Echocardiography (echo) is the mainstay imaging modality in the current practice of cardiology¹, providing
vital, non-invasive assessments of heart anatomy and physiology to guide clinical decisions². In the past
decade, the rising demand for diagnostic echo tests³ has posed significant challenges in maintaining the
quality and timeliness of diagnostic reports^{4–7}, underscoring the necessity for automated solutions to
enhance both efficiency and report quality^{8–10}.

8

9 With the recent emergence of artificial intelligence (AI), automated echo reporting has been proposed to 10 use deep learning (DL) models to generate diagnostic predictions and measurements to fill a pre-set report 11 template^{8,10,11}. These frameworks focused on specific image processing tasks^{8,11} rather than the report text, 12 and are technically equivalent to generating individual findings. However, these frameworks were not 13 designed to handle the high-level cognitive activity of synthesizing clinically relevant impressions from 14 detailed findings¹². In practice, physicians usually spend a significant amount of time summarizing detailed

- 15 findings to clinically relevant final impressions^{13,14}. While this task is crucial, it can be time-consuming and
- 16 prone to errors¹⁵.
- 17

18 The advance of large language models (LLM) marked an important milestone for the application of AI in healthcare to automate clinical information summarization^{13,14,16} and expert-level question-answering¹⁷. A 19 major advantage of LLMs is the flexibility of input and output¹⁸, as well as the capability of handling 20 conversations and interaction with human experts¹⁹. While similar functionality can be achieved through 21 commercially available LLMs (e.g., ChatGPT; OpenAI, San Francisco, CA)²⁰, only a few healthcare 22 23 institutions have integrated ChatGPT²¹. Furthermore, fine-tuning ChatGPT for specific tasks still requires 24 uploading data to a central server, which also raises privacy concerns²². In contrast, open-source LLMs are 25 free of charge and can be locally fine-tuned for specific tasks within each healthcare institution's secure 26 confines¹⁸.

27 Previous studies predominantly focused on electronic health records^{13,16} and chest X-rays (CXR)^{13,18} have 28 highlighted the potential of using LLMs to summarize clinical text. In contrast, echo-related studies were 29 mainly on data extraction or classification, rather than report summarization²³⁻²⁵. Tang et al. used rule-based 30 systems and the BART (Bidirectional and Auto-Regressive Transformer) model²⁶ for this purpose and 31 32 demonstrated convincing results. However, BART-generated content was less favored by human experts 33 more than 50% of the time, perhaps due to the smaller number of parameters than current state-of-the-art 34 LLMs²⁶. The application of using billion-parameter LLMs to generate echo reports remains under-35 explored^{27,28}.

36

In this work, we proposed to construct a local, domain-specific LLM (EchoGPT) dedicated to echocardiography report summarization through an instruction fine-tuning approach, which is known to be an effective strategy to adapt LLMs for similar tasks^{13,18}. We anticipate that the fine-tuning procedure improves LLMs' performance on the task of echocardiography report summarization. EchoGPT will fill the knowledge gap for using open-source LLM in the domain of echocardiography reporting and could enhance the efficiency of the current workflow with uncompromised report quality.

44 Method

45 <u>Echocardiography Report</u>

46 Following the American Society of Echocardiography recommendations²⁹, a standard echocardiography

47 report at the Mayo Clinic contains the following major sections: Final Impressions, Findings, and

- 1 Measurements. The Measurements section contains only measurement values without free text. The key
- 2 measurements such as left ventricular ejection fraction, aortic valve area, and right ventricular systolic
- 3 pressure are included in the Findings section with corresponding statements. Considering the report
- 4 structure above, only the information from the Final Impressions and Findings sections was used in this
- 5 work (Figure 1).
- 6
- 7 <u>Dataset</u>
- 8 Mayo Clinic Reports: All adult (> 18 years old) echocardiography studies performed from 1/1/2017 to
- 9 12/31/2017 at Mayo Clinic Enterprise were retrieved. The types of studies include transthoracic
- 10 echocardiography (TTE), transesophageal echocardiography (TEE), and stress echocardiography
- 11 (including exercise and pharmacological studies). Text in the "Findings" and the "Final Impression"
- 12 sections of each report was extracted for the current study (Figure 2). The study was approved by the
- 13 Mayo Clinic IRB (protocol#: 22-010944).
- 14 MIMIC-III ECHO-NOTE2NUM Dataset (v.1.0.0, referred to as MIMIC-EchoNotes below)³⁰: This
- 15 publicly available dataset contains 43,472 valid free-text echocardiography reports from the intensive care
- 16 unit at the Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset was used for
- 17 external validation.
- 18

19 *Data Curation and Preprocessing*

- 20 Mayo Clinic Reports: Echocardiography reports were excluded according to the following criteria: (1)
- 21 reports without Finding or Final Impression sections, (2) reports whose Finding or Impression section
- contained less than 15 words, as these are frequently canceled studies in our practice, and (3) labeled in
- 23 report metadata as limited report, fetal study, nuclear stress, or vascular study. After this filtering process,
- the text of each report was further processed as follows: (1) Remove capitalized subheadings (e.g., LEFT
- 25 VENTRICLE, VALVES, OTHER FINDINGS, etc.), (2) Remove template sentences that make
- 26 comparisons to prior reports, as no information from previous reports has been provided, and (3) Remove
- 27 quality control-related sentences such as "study performed per left ventricular function protocol" and
- 28 "the goals, risks, and alternatives to moderate sedation were explained to the patient."
- 29 MIMIC-EchoNotes Reports: Cases in this dataset were excluded based on criteria (1) and (2) above, as
- 30 the metadata differed from that of the Mayo Reports. We also removed the subheadings and template
- 31 sentences as described previously. We observed fundamental differences in report structure, including the
- 32 "General Comments" section, which typically contains comments related to study quality, and the
- 33 "Conclusions" section, which usually consists of the physician's interpretation of findings. However, the
- 34 Impression section often contains only 2-3 sentences summarizing the most pertinent study findings,
- 35 which was difficult for head-to-head comparison in this study. Given the distinct report structure, the
- 36 contents under the subheadings "General Comments" and "Conclusions" were integrated into the
- 37 "Findings" and "Impression" sections, respectively. Common abbreviations in the text were expanded to
- their full forms.
- 39
- 40 <u>Data split</u>
- 41 Data from Rochester, Florida, and Mayo Clinic Healthcare sites were used as the model development set.
- 42 Considering variations in practice style among different sites, the data from the Arizona site was
- 43 designated as the external validation set (referred to as the AZ validation set). Within the development set,
- 44 1,000 non-duplicated cases were randomly selected for the test and validation sets, respectively; the rest

1 of the cases were used for fine-tuning (training set). Similarly, from the AZ validation and the MIMIC-

2 EchoNotes datasets, we selected 1,000 non-duplicated random cases from each. For basic dataset

3 statistics, the token length was calculated based on the natural language processing toolkit (NLTK)

4 tokenizer³¹, and the lexical variance was defined as the ratio of the number of unique tokens to the

- 5 number of total tokens in each example¹³.
- 6

7 <u>Model Selection</u>

8 Due to patient privacy policy regulations, proprietary LLMs such as GPT-3.5 and GPT-4 were not

9 considered in this work because versions of those models that were safe for protected health information

10 were not yet available. Among open-source models, we selected representative auto-regressive models

11 including Llama-2-7b-chat²⁷, Zephyr-7b²⁸, and Med-Alpaca³² models considering their performance and

12 max input context length on general natural language processing (NLP) tasks and radiology report

13 summarization ¹³. For sequence-to-sequence (seq2seq) models, we used Flan-T5 (base) as the

14 representative model as it is known for accurate text summarization 13,33 .

15

16 <u>Model Inference Hyperparameter Search</u>

17 LLM inference was conducted by using Hugging Face's (Manhattan, NY) transformer pipeline via the

18 open-source LangChain framework³⁴. After initial tests, text generation and summarization were used as

19 the task type for auto-regressive models and seq2seq models, respectively. A subset (10%, n=100) of

20 examples were randomly selected from the test set for the hyperparameter search. We specifically tested

21 the following configuration parameters that can significantly affect performance: temperature (0.1, 0.5, 0.5)

and 0.9) and repetition penalty (1.1, 1.2, and 1.3). These two parameters were tested separately, when one

23 parameter was being tested, the other was fixed at the lowest value. The generated contents were

evaluated by both automatic metrics and qualitative assessment. We chose the following configuration for

25 model inference: {temperature 0.1, repetition penalty 1.1} after comparing automatic metrics and

26 qualitative assessments; see **Supplemental Table 1**). We did not complete a dedicated search procedure

for the optimal LLM inference configurations, but the configurations used in our study were similar toprior reports, and the generated contents were satisfying on qualitative review. Of note, the configurations

29 were tested in a zero-shot setting, and the best configuration was directly applied to the ICL and QLoRA

- 30 fine-tuned models¹³.
- 31

32 <u>Model Adaptation</u>

33 *Prompt:* A prompt template was created with components of the prefix, instruction, and suffix¹³ (**Table**

1). The final prompt was decided after qualitatively evaluating several different variants of each

35 component on a small subset of the data. We also specified that the summarization should be "concise"

36 and use "a minimal amount of text" to avoid LLMs generating lengthy reports¹³. Likely due to the

37 difference in the reporting style of the MIMIC-EchoNotes dataset, the final prompt above led to

38 suboptimal responses. Therefore, we adopted a new prompt tailored to match the reporting style by

39 incorporating new instructions below: 1) Write a 10-bullet points clinical summary, and 2) Avoid using

40 numbers other than LVEF.

41

42 *In-context Learning (ICL)*: ICL has been proposed to improve LLM's performance without changing the

43 base model weights^{13,35,36}. Also, using relevant in-context examples is shown to have better model

44 performance compared to random examples in ICL. To obtain relevant in-context examples, we adopted

the approach to select m (m=1, 2, 4...) nearest neighbors from the training set for each test set case, after
 embedding both sets by the PubMedBERT model³⁷.

3

Instruction tuning with quantized low-rank adaptation (QLoRA): Due to the size of candidate models, we
opted for quantized low-rank adaptation (QLoRA)³⁸, a type of parameter efficient fine-tuning (PEFT)³⁹ to
optimize our LLMs for echo report summarization tasks. The same prompt template (Table 1) was used,

- 7 and the Final Impression text from the same report was used as the target output 13,14 .
- 8

9 We configured the training process as follows: load model in 4-bit precision, with a LoRA configuration

10 of (alpha = 16, LoRA dropout = 0.1, LoRA r= 64). The batch size and gradient accumulation were

adjusted for each model to achieve an effective batch size of 24 that fits on a single NVIDIA RTX A5000
 24G GPU setting. A paged-AdamW 32-bit optimizer was used, with an initial learning rate of 1e-3, which

decayed to 1e-4 (by a cosine scheduler) after the initial 100 warm-up steps. The above configuration

14 provided the most stable training process after attempting different configurations reported in prior

provided the most stable training process after attempting different configurations reported in prior
 studies^{13,38}.

16

17 <u>Model Performance Evaluation</u>

18 *Automatic NLP evaluation metrics*: To evaluate the models' performance on the information

19 summarization task and compare it to prior works, we utilized four established automatic metrics that

20 have been used in other clinical text summarization studies^{40,41}: BLEU (Bilingual Evaluation

21 Understudy)⁴², METEOR (Metric for Evaluation of Translation with Explicit ORdering)⁴³, ROUGE-L

22 (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence), and the BERT

23 (Bidirectional Encoder Representations from Transformers) score⁴⁴, which represents the similarity

24 between generated contents and the corresponding reference at words/characters (n-gram), single word

25 (unigram), longest sequence of words, and contextual level, respectively For ROUGE-L, we present the

F1 score component^{26,45}. For factual correctness, the RadGraph-F1 metric (level: all) was reported^{46,47}

27 This metric served as the primary evaluation criterion for model performance, considering its significance

28 in ensuring the factual correctness of generated clinical content.

29

Evaluation of Significance of Measurement Numbers in Automatic Metrics: Considering the importance
 of measurements in echo studies, we also attempted to evaluate whether the current automatic metrics can

32 detect changes in measurement numbers. For this purpose, we generated synthetic reports by replacing all

33 the measurement numbers with random numbers ranging from 1 to 99 in the reports. We then compared

34 the automatic metric scores with the corresponding reports with the original measurements.

35

36 *Human expert evaluation metrics*: We designed a human expert evaluation process based on previous

37 clinical text summarization studies^{13,18,26}. The Findings with corresponding ground truth (Final

38 Impression) and LLM-generated summarization of 30 randomly selected cases were presented to four

39 echocardiography-board-certified cardiologists for blinded quality review. We noted that physician-

40 summarized Impressions may contain free-text information beyond the Findings (e.g., documenting

41 events during the study or communication with the ordering provider), and the reviewers were instructed

42 to rate only based on information within the Findings section. Each metric was rated for the preference (5

43 levels) between the two summarizations (Supplemental Figure 1)¹³. The "4C metrics" evaluated by

echocardiography experts are completeness, conciseness, correctness, and clinical utility¹⁴, as described in
 Table 2.

3

4 Statistical analysis: Automatic metric performance between each pair of models was compared using a

- 5 two-tailed paired Student's t-test or Wilcoxon signed-rank test for normally and non-normally distributed
- 6 data, respectively. Models were also compared based on win rates, which are defined as the percentage of
- 7 head-to-head victories in performance between two models for each selected metric¹³. For the
- 8 performance bias analysis, data were grouped based on sex (male versus female) and race (white, black,
- 9 other). In the case of sex, we employed two-tailed Student's t-tests or Mann-Whitney U tests for normally
- 10 and non-normally distributed data, respectively. For race, we utilized one-way ANOVA. In human expert
- 11 qualitative analysis, the 4C metrics were compared by a one-sample Wilcoxon signed-rank test¹³, and the
- agreement of ratings between experts was assessed by Fleiss' kappa coefficient⁴⁸, and interpreted as
- 13 recommended by Landis et al⁴⁹. We conducted Pearson correlation analyses to explore the relationships
- 14 between each human expert evaluation metric, assessing their independence. Additionally, we conducted
- similar analyses to examine the correlation between human and automatic metrics. Statistical analyses
- were performed using Python 3.8 and SciPy 1.8.0. All the comparisons consider a p-value < 0.05 as
 significant.
- 17 s 18

19 Results

- 20 Patient Cohort and Dataset
- 21 Our development set contains 97,506 reports from 71,717 unique patients, with a mean age of 64.3±15.8
- 22 years, 54,005 (55.4%) were male, 89,466 (91.8%) were white. Randomly selected from Mayo Arizona
- studies (19,557 reports/15,853 unique patients), the AZ validation set contains 1,000 reports from 1,000
- 24 unique patients with a mean age of 63.9 ± 16.0 years, 584 (58.4%) were male and 885 (88.5%) were
- 25 white. Detailed demographic information was not available for the MIMIC-EchoNotes dataset. Other
- 26 detailed patient characteristics and statistics of text data are summarized in **Table 3**. Transthoracic
- 27 echocardiography was the predominant study type in development, AZ validation, and MIMIC-Echo
- 28 datasets (81.9%,77.6%, and 86.4% respectively).
- 29

30 <u>Zero-shot, ICL, and QLoRA fine-tuned performance</u>

- 31 Table 4 is a summary of the performance of zero-shot and fine-tuned LLMs, including MedAlpaca,
- 32 Llama-2, and Zephyr. QLoRA fine-tuning significantly improved LLMs' performance from baseline.
- 33 Note that T5 and MedAlpaca were not fine-tuned so only zero-shot results were provided for reference.
- 34 Among the candidate models, Llama-2 generally had the best zero-shot performance, which was
- 35 consistent in ICL. While Flan-T5 had a similar or superior performance to Llama-2 across most metrics, it
- 36 was particularly worse on the RadGraph F1 score (**Table 4; Figure 3**). On qualitative review, we noted
- 37 that T5 provided concise summaries, however important clinical information was missed in this process
- 38 (Supplemental Table 2).
- 39
- 40 In ICL, LLMs that allow longer context length (Llama-2 and Zephyr) had the best performance across all
- 41 metrics when one example was provided (ICL-1). The performance gradually trended down with more
- 42 examples (ICL-2 and ICL-4). In contrast, the performance of LLMs with shorter max context length
- 43 started to trend down with one example (Figure 3).
- 44

- 1 Based on the zero-shot and ICL performance of the candidate models, Llama-2 and Zephyr were selected
- 2 for instruction fine-tuning. Compared to zero-shot, QLoRA significantly improved the performance of
- 3 selected LLMs across all metrics (Table 4). For the head-to-head comparison in model win rates, fine-
- 4 tuned Llama-2 was superior to all other models, including Zephyr (base and fine-tuned), MedAlpaca
- 5 (base), and Flan-T5 (base) across all 5 automatic metrics (Figure 4). Llama-2 maintained similar
- 6 performance in the AZ validation set (n=1,000) and was consistently superior to fine-tuned Zephyr
- 7 (Supplemental Table 4). Regarding potential biases, we did not observe significant biases regarding sex
- 8 and race across the automatic metrics, except for a slightly better RadGraph F1 performance in female
- 9 patients in the AZ validation set (male vs. female: 0.38 ± 0.14 vs. 0.40 ± 0.15 , p=0.04) (Supplemental
- **Figure 2**). Because Llama-2 had the best performance on zero-shot, ICL, and QLoRA fine-tuning
- approaches, fine-tuned Llama-2 was selected as EchoGPT and used for the subsequent expert qualitative
 review.
- 13
- 14 <u>Significance of Measurement Numbers in Automatic Metrics</u>
- 15 After replacing the measurement numbers with random numbers, we observed relatively minor decreases
- 16 in BLEU, METEOR, ROUGE-L, BERT, and RadGraph F1 scores, while statistically significant
- 17 (p<0.0001). One can see that in the provided examples, the report with random numbers doesn't make
- 18 clinical sense when compared to the original content (Table 5).
- 19
- 20 <u>Human Expert Evaluation</u>
- 21 We observed slight agreement for correctness, fair agreement for conciseness and clinical utility, and
- 22 moderate agreement for completeness (Supplemental Table 5). Among the 4C metrics, we observed that
- 23 EchoGPT significantly outperformed human experts in conciseness (p<0.001). There was no significant
- 24 preference among the other three categories (Figure 5A). The 4C metrics were not completely
- 25 independent. There was a high correlation between clinical utility and completeness (Pearson's r= 0.78),
- and modest to moderate correlations between other metrics (Figure 5C). We also observed that across all
- 27 automatic metrics, RadGraph F1 had modest to moderate correlations with all 4 human evaluation metrics
- **28** (Figure 5C).
- 29

30 <u>External Validation on the MIMIC-EchoNotes dataset</u>

- 31 In the MIMIC-EchoNotes dataset (n=1,000), we observed a performance drop in fine-tuned models, while
- 32 EchoGPT was still superior to fine-tuned Zephyr (Supplemental Table 4). Regarding the expert review,
- 33 we observed moderate agreement for completeness and conciseness, but slight agreement for correctness
- 34 and clinical utility (Table 6). The original reports (combined Conclusions and Impression sections) were
- 35 preferred over EchoGPT in completeness, correctness, and clinical utility (p < 0.001); while EchoGPT
- 36 was superior in conciseness (p<0.001) (Supplemental Figure 6A). RadGraph F1 still had the strongest
- 37 correlations with the 4C metrics, although other metrics also had stronger correlations (**Supplemental**
- **38** Figure 3). A representative example demonstrated the difference in reporting structure and style, along
- 39 with reviewers' feedback on the two datasets (**Table 6**).
- 40

41 Discussion

42

- 43 To the best of our knowledge, this study is the first attempt to compare multiple open-source LLMs in
- 44 echocardiography report summarization through different model adaptation methods. Trained on one of

1 the largest echocardiography report datasets in the world, we demonstrated that QLoRA fine-tuning can

- 2 significantly improve LLMs' performance for the desired summarization task, with at least comparable
- 3 qualities to human experts. Our results also indicate that ICL is associated with significant limitations for
- 4 clinical practice, including compromised patient privacy and report quality. Additionally, we
- 5 demonstrated that current automatic metrics are not sensitive to the change in measurement numbers in
- 6 echo reports. The current study provides insights into the construction of a dedicated local LLM for echo
- 7 report summarization and can pave the way for an AI-enabled echocardiography interpretation system
- 8 with a human-AI interaction interface.

9

10 <u>Model Adaption Approaches: ICL vs. Fine-tuning</u>

- 11 Our results suggested that both ICL and QLoRA fine-tuning improved LLMs' performance over zero-
- 12 shot, and fine-tuning performance was consistently above ICL across all automatic metrics (Figure 3).
- 13 Additionally, we observed substantial limitations of ICL for echo reporting, including relatively longer
- 14 context length and compromised patient privacy, as discussed below.
- 15
- 16 Compared to CXR reports, echocardiography reports come with a relatively longer context, which
- 17 directly affects the available choices of LLMs. In contrast to a prior study that used 32 or more
- 18 examples¹³, we were only able to test up to 4 examples for ICL, therefore not able to assess the models'
- 19 behavior with more examples. However, across all metrics, LLMs' had gradually down-trending
- 20 performance when more examples were provided (Figure 3), which is consistent with prior studies^{13,35}. It
- 21 is also important to consider that the computation time and resources required in ICL can increase with
- the number of examples used³⁵. Additionally, we note that LLMs integrate information from the example
- 23 ICL cases, which compromises the report quality and potentially patient privacy (**Supplemental Table**
- **3**). While this behavior was not reported in other studies^{13,35}, we believe it could be a common condition,
- that is easier to identify with numerical values (in echo) compared to narrative statements (in CXR).
- 26 Therefore, even in scenarios where ICL can outperform fine-tuning¹³, fine-tuning may be preferable.
- 27
- 28 The EchoGen study previously demonstrated that Bidirectional Auto-Regressive Transformers (BART)
- 29 was superior to other rule-based approaches for summarizing echocardiogram reports, with BART
- 30 achieving ROUGE-based scores between 0.65 and 0.73, however, human summaries were preferred by
- 31 the majority of the time over those generated by the BART model²⁶. Although EchoGPT didn't match the
- 32 scores in ROUGE-L, it compared favorably to human experts in qualitative assessments. Additionally, in
- 33 our study, we observed that T5 (as the representative seq2seq model) generated summaries that were
- 34 overly brief, so important clinical information was missed (**Supplemental Table 2**). Although the
- 35 EchoGen authors did not provide qualitative examples generated by their BART model, we assume that
- 36 similar behavior occurred with BART which led to the unfavorable rating by physicians.
- 37
- 38 *Evaluation of Echo Report Summarization*
- 39 Automatic evaluation of LLM in clinical text summarization tasks is an emerging area, and there is no
- 40 gold standard metric that can evaluate all aspects of a report^{13,26}. Our study reinforces this conclusion. We
- 41 noted that MedAlpaca and Zephyr can generate medical-professionally-sounding content that frequently
- 42 includes hallucinated information. These differences were mainly reflected by the factual correctness
- 43 metric RadGraph F1 (**Table 4**).
- 44

1 The practice style at each institution could greatly affect the quantitative performance of a model¹⁸. In the

2 AZ validation set, we observed a 5-10% drop in performance of both fine-tuned Llama-2 and Zephyr

3 (Supplemental Table 4). This is likely secondary to the differences in practice style: the AZ validation

4 set, despite having a similar Finding section length, contained an average of 9.5 additional tokens in the

5 Final Impression section (**Table 3**). A more significant drop in performance was observed in reports from

- 6 the MIMIC-EchoNotes datasets (RadGraph F1 from 47.7 to 25.1; **Supplemental Table 4**), which was
- anticipated and within a reasonable range⁵⁰. According to our observations and the input of expert
 reviewers, the key factors leading to the performance drop were:
- 9 10
- 1. The distinct report structure (Findings/Impressions versus Findings/Conclusions/Impressions).
- 11 12

2. The use of reporting languages (templated statements at Mayo versus the free-text style of the MIMIC-EchoNotes dataset).

13

14 Specifically, the combination of the Conclusion and Impression sections makes the section almost as long 15 as the Findings (184.3 \pm 50.0 vs. 163.2 \pm 40.4; **Table 3**), and even longer in some cases. Additionally, the 16 physician's interpretation often contains information beyond the Findings, or variations of the original 17 sentences, that the model won't be able to summarize. Moreover, the sentence templates at Mayo include 18 measurement numbers in relevant statements (e.g., ejection fraction, right ventricular systolic pressure, 19 left atrial size index), while MIMIC-EchoNotes did not (Table 6). These factors contributed to the overall 20 less favored completeness, correctness, and clinical utility of EchoGPT summaries on the MIMIC-21 EchoNotes dataset (Supplemental Figure 3). It is important to note the low agreement on correctness and 22 clinical utility metrics, which also implies the challenge on comparing reports with distinct styles 23 (Supplemental Table 5). The difference across institutions, as listed above, could limit the direct 24 generalization of a fine-tuned LLM for report summarization. However, while not comprehensively tested 25 in the current study, we noted that adjusting the prompt to fit the reporting style could lead to better 26 summaries without further fine-tuning^{50,51}.

27

Regarding the correlations between automatic metrics and human expert preference, our results were
 similar to the prior studies, showing that most of the metrics were not strongly correlated¹³. Notably, the

30 highest correlation was observed between the RadGraph F1 scores and the 4C metrics, particularly in

31 terms of clinical utility (r=0.42) (**Figure 5C**). This suggests that the quality of echo reports judged by

32 cardiologists may not be well captured in automated metrics that do not capture notions of factual

33 correctness. While stronger correlations were observed in the MIMIC-EchoNotes examples

34 (Supplemental Figure 3), we believe it was reflecting the strong preference secondary to the distinct35 reporting style.

36

As a specific subtype of clinical text, echocardiography reports contain unique terminology, including
 precise measurements. Clinically, 25% and 55% LV ejection fraction values indicate a significant

39 difference, however, our study demonstrates that this distinction is difficult to capture with current

40 automatic metrics (**Table 5**). While this aspect of reporting can be easily captured in qualitative analyses,

- 41 such analyses are expensive to conduct at scale because of the limited availability of in-domain experts. A
- 42 dedicated metric for echocardiography diagnostic quality evaluation, with emphasis on measurement
- 43 accuracy, is still needed to address this knowledge gap.
- 44

1 <u>Application of EchoGPT</u>

2 Our study shows the feasibility of introducing LLMs into echocardiography practice. Through the

- 3 QLoRA fine-tuning process³⁸, the EchoGPT model was able to learn clinically relevant knowledge to
- 4 summarize echo report findings at a quality level comparable to echocardiography-trained cardiologists
- 5 (Figure 6A).
- 6

7 The current study concentrated solely on the performance of EchoGPT in summarizing echocardiogram

8 reports. However, as an LLM-based system, EchoGPT holds promise for broader applications, including

- 9 providing in-context clinical reasoning, answering questions based on patient data, and interacting with
- 10 human experts or other models 12,16,52 . In contemplating these expanded functionalities, two critical factors
- 11 emerge: the maximum permissible context length and the model's proficiency in managing conversations.
- 12 These considerations might restrict the use of existing seq2seq models in such scenarios 33 .
- 13
- 14 We envision that EchoGPT could be used as a reporting interface or a co-pilot that could generate echo
- 15 reports with various inputs⁵³. EchoGPT inherits the limitations of LLMs, including hallucination^{13,18,54}.
- 16 Although the fine-tuning process can potentially reduce hallucinations, additional efforts such as
- 17 optimization for factual correctness⁴⁶ or paired with a retrieval augmented generation system⁵⁵ are still
- 18 required to minimize hallucinations before clinical implementation.
- 19

25

20 Conclusion

- 21 Our study successfully built EchoGPT through QLoRA fine-tuning of open-source LLMs and
- 22 demonstrated that the model is capable of generating echocardiography reports on par with cardiologists,
- 23 marking an advancement in integrating LLMs into current echo practice. Through further optimizations in
- 24 the future, EchoGPT is envisioned to become a human-AI co-pilot for echo report generation.

26 Limitations

- 27 This study is limited by its retrospective nature and a predominantly white population served by the
- 28 healthcare system. However, we were able to demonstrate that the algorithm is not biased by sex and race.
- 29 Our echocardiography reports are based on standardized statements, with an option to add free text. While
- 30 the lexical variance was high, the corpus could differ from reports composed entirely of free text contents.
- 31 Due to patient privacy regulations, this work did not assess the performance of GPT-3.5 and GPT-4.
- 32 However, we compared the performance of state-of-the-art open-source LLMs, which provided important
- 33 insights for model selection when data privacy is a critical consideration. Instead of full fine-tuning,
- 34 QLoRA was used as the fine-tuning approach, however, it has been demonstrated as an effective
- 35 approach as full fine-tuning is often not feasible for LLMs. Last but not the least, although QLoRA fine-
- 36 tuning demonstrated improvements in echo report summarization tasks, our current approach does not
- 37 include optimization for factual correctness and human expert preference.
- 38

39 Data Availability

- 40 The data that support the findings of this study are not openly available due to reasons of sensitivity and
- 41 patient privacy. Data are located in controlled access data storage at the Mayo Clinic. The MIMIC-
- 42 EchoNotes (ECHO-NOTE2NUM) data is publicly available at <u>https://doi.org/10.13026/xhrz-ht59</u>
- 43
- 44 Code Availability

- 1 We released a checkpoint of the fine-tuned Llama-2 model, along with the QLoRA fine-tuning, inference,
- 2 and statistical analysis code. The code and checkpoint are available on GitHub:
- 3 <u>https://github.com/chiehjuchao/EchoGPT.git</u>
- 4 Figure Legends
- 5 Figure 1. The current workflow of summarizing echo Findings into clinically relevant Final Impressions.
- 7 Figure 2. Overview of the EchoGPT study.
- 8

6

- 9 Figure 3. ICL performance of each LLM. Panel A to E correspond to BLEU, METEOR, ROUGE-L,
- 10 BERT Score, and RadGraph F1 Score, respectively. Zero-shot and fine-tuned Llama-2 (EchoGPT;
- 11 horizontal purple dashed line) performance was included for reference.
- 12
- 13 Figure 4. Model win rates on the test set. Model win rate heatmap illustrates the head-to-head win rate
- 14 comparisons (in percentile) among different models based on the selected metrics. Cool colors indicate
- 15 lower win rates and warmed colors indicate higher win rates. We compared Llama-2 (base and fine-
- 16 tuned), Zephyr (base and fine-tuned), T5 (base), and MedAlpaca (base). Fine-tuned Llama-2 consistently
- 17 outperformed all other models across all 5 automatic metrics. FT: fine-tuned. ZS: zero-shot (base model).
- 18
- 19 Figure 5. Human expert qualitative evaluation results. Panel A. In the 4 categories, EchoGPT
- significantly outperformed human experts in conciseness (p<0.001). We didn't observe significant
- 21 differences among the other three categories (completeness, correctness, and clinical utility). Panel B.
- showed interdependence of the 4C metrics, especially the correlations between clinical utility and
- 23 completeness (Pearson's r= 0.78), and modest to moderate correlations between other metrics. Panel C.
- 24 Correlations between automatic metrics and the 4C metrics. Across all automatic metrics, RadGraph F1
- 25 had modest to moderate correlations with all 4 human evaluation metrics. Preference ratings were
- 26 expressed as mean \pm standard deviation, **indicates p<0.001.
- 27 Supplemental Materials

Supplemental Figure 1. Echocardiography expert review questionnaire. Expert readers were asked to
 rate summaries A and B concerning the 4C metrics without knowing it's a human- or LLM-generated
 summary.

- 31 Supplemental Figure 2. This bar chart displays model performance across five automatic metrics,
- 32 considering sex (male vs. female) and race (white, black, and other) variables. Panels A and B compare
- 33 metrics by sex and race variables in the test set, while Panels C and D perform the same comparisons in
- 34 the AZ validation set. There were no significant biases detected for sex or race, except for slightly better
- **35** RadGraph F1 performance in female patients within the AZ validation set (male vs. female: 0.38 ± 0.14
- 36 vs. 0.40 ± 0.15 , p=0.04). Scores were presented as mean values \pm standard error bars. Demographic
- 37 information was not available for the same analysis in the MIMIC-EchoNotes dataset.
- 38 Supplemental Figure 3. Human expert qualitative evaluation results on the MIMIC-EchoNotes dataset.
- 39 Panel A. In the 4 categories, the original reports (combined Conclusions and Impression sections) were
- 40 preferred over EchoGPT in completeness, correctness, and clinical utility (p < 0.001); while EchoGPT

1 2 3 4 5 6	was superior in conciseness (p<0.001). Panel B. Interdependence of the 4C metrics, especially the correlations between clinical utility, completeness, and correctness (Pearson's $r= 0.48$ and 0.66, respectively). Panel C. Correlations between automatic metrics and the 4C metrics. Across all automatic metrics, RadGraph F1 still had the strongest correlations with the 4C metrics, although other metrics also had stronger correlations. Preference ratings were expressed as mean \pm standard deviation, **indicates p<0.001.
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	
26	
27	
28	

1 Tables

2

Table 1. Prompt Template

	Prompt Component
Prefix	"You are a knowledgeable cardiologist."
Instruction	"For the following echocardiography report findings, please write a concise summary with a minimal amount of text."
Suffix (ICL only)	"Use the following examples to guide word choice."

Table 2. Definition of the 4C Human Expert Evaluation Metrics

Metric	Definition
Completeness	This metric evaluates whether the generated contents include all relevant details, elements, or aspects, leaving no important information gaps.
Conciseness	This metric measures the quality of being clear and succinct, presenting information or content in a brief and straightforward manner without unnecessary elaboration or redundancy.
Correctness	This metric assesses whether the generated information is true, free from mistakes, and aligned with established facts or standards.
Clinical Utility	This metric evaluates whether the generated information is useful in clinical practice.

Table 3. Data distribution of the development and AZ validation sets.

		Developm	AZ validation set	MIMIC-EchoNotes		
	All	Train	Validation	Test		
	n=97,506	n= 95,506	n=1,000	n=1,000	n=1,000	n=1,000
Age	64.3 ± 15.8	64.3 ± 15.8	64.9 ± 16.0	64.7 ± 15.7	63.9 ± 16.0	
Race						
White	89466 (91.8%)	87632 (91.8%)	914 (91.4%)	920 (92.0%)	885 (88.5%)	
Black	3326 (3.4%)	3257 (3.4%)	35 (3.5%)	34 (3.4%)	42 (4.2%)	
Other	2850 (2.9%)	2790 (2.9%)	28 (2.8%)	32 (3.2%)	35 (3.5%)	
Asian	1415 (1.5%)	1390 (1.5%)	16 (1.6%)	9 (0.9%)	22 (2.2%)	
Native American	362 (0.4%)	353 (0.4%)	5 (0.5%)	4 (0.4%)	13 (1.3%)	
Pacific Islander	87 (0.1%)	84 (0.1%)	2 (0.2%)	1 (0.1%)	3 (0.3%)	
Sex						
Male	54010 (55.4%)	52920 (55.4%)	563 (56.3%)	527 (52.7%)	584 (58.4%)	
Female	43496 (44.6%)	42586 (44.6%)	437 (43.7%)	473 (47.3%)	416 (41.6%)	
HTN	11578 (11.9%)	11320 (11.9%)	115 (11.5%)	143 (14.3%)	125 (12.5%)	
DM	14406 (14.8%)	14105 (14.8%)	142 (14.2%)	159 (15.9%)	115 (11.5%)	

CAD	4983 (5.1%)	4863 (5.1%)	51 (5.1%)	69 (6.9%)	57 (5.7%)	
CHF	38908 (39.9%)	38105 (39.9%)	384 (38.4%)	419 (41.9%)	293 (29.3%)	
CKD	15071 (15.5%)	14759 (15.5%)	140 (14.0%)	172 (17.2%)	170 (17.0%)	
Stroke	2879 (3.0%)	2812 (2.9%)	30 (3.0%)	37 (3.7%)	21 (2.1%)	
Echo Study Type						
Adult TTE	79902 (81.9%)	78243 (81.9%)	833 (83.3%)	826 (82.6%)	776 (77.6%)	881 (88.1%)
Adult TEE	7829 (8.0%)	7685 (8.0%)	65 (6.5%)	79 (7.9%)	63 (6.3%)	108 (10.8%)
Exercise Stress	6781 (7.0%)	6652 (7.0%)	64 (6.4%)	65 (6.5%)	131 (13.1%)	11 (1.1%)
Pharmacological Stress	2994 (3.1%)	2926 (3.1%)	38 (3.8%)	30 (3.0%)	30 (3.0%)	
Data Characteristics						
Average Number of Tokens	2157+558	215.7 ± 55.9	213 1 + 55 2	214 8 + 53 3	217 1 + 77 3	184.3 ± 50.0
(Findings)	215.7 ± 55.0	215.7 ± 55.9	215.1 ± 55.2	214.0 ± 55.5	217.1 ± 77.5	104.5 ± 50.0
Average Number of Tokens	88 2 + 34 3	88 2 + 34 3	871 + 337	868 + 336	07.7 ± 37.0	163.2 ± 40.4
(Final Impression)	00.2 ± 57.5	00.2 ± 54.5	07.1 ± 33.7	80.8 ± 55.0)1.1 ± 31.)	105.2 ± 40.4
Average Lexical Variance	0.51	0.51	0.52	0.51	0.52	0.54
(Findings)	0.31	0.51	0.52	0.51	0.52	0.54
Average Lexical Variance	0.65	0.65	0.65	0.65	0.64	0.52
(Final Impression)	0.05	0.05	0.05	0.05	0.04	0.52

CAD: coronary artery disease, HTN: hypertension, DM: diabetes, CKD: chronic kidney disease, CHF: congestive heart failure. TTE: transthoracic echocardiography, TEE: transesophageal echocardiography.

1 2

Table 4. Quantitative performance of zero-shot and QLoRA fine-tuned LLMs

Model	Flan-T5	MedAlpaca		Llama-2 Zephyr					
Metric	zero-shot	zero-shot	zero-shot	QLoRA	P-value*	zero-shot	QLoRA	P-value*	P-value**
BLEU	9.8 ± 9.8	2.0 ± 4.4	6.8 ± 5.5	45.9 ± 18.9	< 0.0001	3.5 ± 4.2	20.6 ± 15.5	< 0.0001	< 0.0001
METEOR	35.6 ± 16.0	18.8 ± 9.6	21.6 ± 7.3	62.4 ± 18.0	< 0.0001	21.8 ± 7.9	35.0 ± 16.1	< 0.0001	< 0.0001
ROUGE-L	22.4 ± 12.0	17.5 ± 9.6	21.3 ± 8.5	55.7 ± 17.8	< 0.0001	19.1 ± 7.4	32.8 ± 15.3	< 0.0001	< 0.0001
BERT									
Score	85.9 ± 2.5	81.6 ± 3.4	85.4 ± 2.0	91.6 ± 3.0	< 0.0001	83.5 ± 2.7	87.4 ± 3.9	< 0.0001	< 0.0001
RadGraph									
F1	17.6 ± 10.8	11.2 ± 8.8	24.2 ± 11.6	47.7 ± 14.9	< 0.0001	14.7 ± 9.4	29.3 ± 10.4	< 0.0001	< 0.0001
***********		OI aDA manf		1	1. **C		and afor a		a un al

*Compared zero-shot to QLoRA performance of the same model; **Compared performance of QLoRA Llama-2 and Zephyr. BLEU: Bilingual Evaluation Understudy, METEOR: Metric for Evaluation of Translation with Explicit ORdering, ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence, BERT: Bidirectional Encoder Representations from Transformers.

3 4

Table 5. Comparison of Automatic Metrics between the Original Response and the Synthetic

Response

Model: EchoGPT	Original Response	Synthetic Response: Replaced Measurements with Random Numbers	QLoRA	Random Measurements	P-value
BLEU	Grade 1/4 left ventricular	Grade 9/53 left ventricular	45.9 ± 18.9	40.4 ± 16.4	< 0.0001
METEOR	consistent with low to	consistent with low to normal left ventricular filling pressure. estimated	62.4 ± 18.0	58.7 ± 16.8	< 0.0001
ROUGE-L	filling pressure. estimated		55.7 ± 17.8	52.2 ± 16.5	< 0.0001
BERT Score	pressure 24 (systolic blood	pressure 68 (systolic blood	91.6±3.0	91.2 ± 2.8	<0.0001

RadGraph F1	pressure 108). calculated ejection fraction; 64%. no regional wall motion abnormalities. mildly enlarged mid ascending aorta diameter of 40.	pressure 52). normal left ventricular chamber size. calculated ejection fraction; 98%. no regional wall motion abnormalities. mildly enlarged mid ascending aorta diameter of 12.	47.7 ± 14.9	45.3 ± 13.7	<0.0001
----------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------	-------------	---------

1 2

Table 6. Representative Examples in Reader Study

Mayo Echo Report	Findings	Impression	EchoGPT Summary
Representativ e Example	Calculated 2-d linear left ventricular ejection fraction 61. No regional wall motion abnormalities. aortic valve strands (on the ventricular side of the valve). Estimated right ventricular systolic pressure 23 (systolic blood pressure 108). Normal abdominal aorta doppler flow pattern. Normal left ventricular wall thickness. No atrial level shunt by color flow imaging. left atrial volume index 30. Trivial aortic valve regurgitation. Trivial mitral valve regurgitation. Trivial pulmonary valve regurgitation. No pericardial effusion. Normal aortic valve. thickness. Normal pulmonary valve. Normal mitral valve. Normal tricuspid valve. Trileaflet aortic valve. Normal left ventricular chamber size. Normal right atrial size. Normal left atrial size with normal inspiratory collapse (>50%) inferior vena cava size normal. Normal ascending aorta diameter. Normal right ventricular systolic function. Normal pulmonary valve systolic function. Normal pulmonary valve systolic function. Normal right ventricular systolic function. Normal right ventricular systolic function. Normal pulmonary valve systolic function. Normal right ventricular systolic function. Normal pulmonary valve systolic function. Calculated 2- d monoplane volumetric left ventricular ejection fraction 59.	Normal left ventricular chamber size. Calculated ejection fraction 61%. No regional wall motion abnormalities. Normal right ventricular chamber size and function. Estimated right ventricular systolic pressure 23 mmHg. Inferior vena cava size: normal, with normal inspiratory collapse (>50%). Normal left ventricular diastolic function.	No regional wall motion abnormalities. Normal cardiac valves. Estimated right ventricular systolic pressure 23 (systolic blood pressure 108). No pericardial effusion. Normal left ventricular chamber size and wall thickness. Calculated left ventricular ejection fraction 61%. Normal right ventricular chamber size and systolic function. Inferior vena cava size: normal, with normal inspiratory collapse (>50%)
Comments	Numbers were utilized in pertinent statements, such as right ventricular systolic pressure, systolic blood pressure, ejection fraction, and left atrial volume index. Typical reports detail the method of LVEF measurement alongside the value.	This section typically contains a 7–10-point summary that covers the overall condition of the heart. In this case, no information beyond the Findings was included.	EchoGPT captured the relevant findings, although not in the same order. LV ejection fraction measurement was included in the summary (61%).
MIMIC- EchoNote Report	Findings	Conclusions/Impression	EchoGPT Summary

Representativ e Example	Highly mobile echodensity (0.8x0.5 cm) in the right atrium, best seen on right ventricle inflow views, which appears to be attached to a catheter (catheter not fully visualized, but patient with history of port-A-cath). Doppler across tricuspid valve suggestive of tricuspid valve prolapse. Normal left atrium size. Normal right atrium size. right atrium mass. Normal inferior vena cava diameter (<=2.1cm) with >50% decrease with sniff (estimated right atrium pressure (0-5 mmHg). Normal left ventricle wall thickness, cavity size and regional/global systolic function (left ventricular ejection fraction >55%). Normal right ventricle chamber size and free wall motion. Normal diameter of aorta at the sinus, ascending and arch levels. No 2D or Doppler evidence of distal arch coarctation. Normal aortic valve leaflets (3). No aortic stenosis. No aortic regurgitation. Normal mitral valve leaflets with trivial mitral regurgitation. TVP. Mild tricuspid regurgitation. Normal pulmonic valve leaflet. No PS. Physiologic pulmonary regurgitation. No pericardial effusion. If clinically indicated, a transesophageal echocardiographic examination is recommended. Echocardiographic results were reviewed by telephone with the MD caring for the patient.	The left atrium is normal in size. A highly mobile mass, measuring 0.8x0.5 cm, is seen in the right atrium (possibly attached to a catheter?) which may represent thrombus or vegetation. The estimated right atrial pressure is 0-5 mmHg. Left ventricular wall thickness, cavity size and regional/global systolic function are normal (left ventricular ejection fraction >55%). Right ventricular chamber size and free wall motion are normal. The diameters of aorta at the sinus, ascending and arch levels are normal. The aortic valve leaflets (3) appear structurally normal with good leaflet excursion and no aortic stenosis or aortic regurgitation. Tricuspid valve prolapse is present based on doppler findings with mild tricuspid regurgitation. The mitral valve appears structurally normal with trivial mitral regurgitation. There is no pericardial effusion. Highly mobile mass is seen in the right atrium (possibly attached to a catheter?) which may represent thrombus or vegetation. Clinical correlation advised. Normal biventricular systolic function. Tricuspid valve prolapse with mild tricuspid regurgitation. If clinically indicated, a transesophageal echocardiographic examination is recommended. Drs. and were notified of these findings at 4:45 p.m. today.	Mass noted in right atrium, measuring 0.8 x 0.5cm, attached to a catheter (catheter not fully visualized but has history of port-A-cath). Doppler across tricuspid valve suggestive of tricuspid valve prolapse. Normal left ventricular chamber size, normal regional/global systolic function, estimated ejection fraction 60%. Normal right ventricular chamber size and function. Normal diameter of aorta at the sinus, ascending and arch levels. No 2D or Doppler evidence of distal arch coarctation. No pericardial effusion. If clinically indicated, a transesophageal echocardiographic examination is recommended.
Comments	In the findings section, normal LVEF was described as 'LVEF > 55%' without further specification. Measurements such as right ventricular systolic pressure, and LA volume index were not included in the report.	The combined section was longer than the Findings section. The blue and red text belongs to the Conclusions and Impression sections, respectively. Information beyond the Findings was in the text: "Drs. and were notified"	Relatively brief summary generated by EchoGPT, capturing the important findings (right atrial mass attached to catheter) and relevant recommendations. However, EchoGPT hallucinated estimated ejection fraction- which was a template sentence used at Mayo.

1

Supplemental Tables

2 3 4

Supplemental Table 1. Hyperparameter Search Results

				Repetition Pena	lty (Temp fixed			
	Tempe	erature (RP fixed	at	0.1)				
	0.1	1.2	1.3					
	LLaMA-2							
BLEU	BLEU 8.0 ± 5.4 8.8 ± 6.7 8.0 ± 6.1 5.5 ± 4.6 1.1 ± 1.8							
METEOR 23.3 ± 8.2 24.1 ± 8.2 23.0 ± 8.1 20.1 ± 6.9 12.9 ± 5								
ROUGE-L	22.3 ± 7.4	23.2 ± 7.8	23.3 ± 9.3	18.7 ± 7.0	11.6 ± 5.3			

BERT Score	85.7 ± 1.5	86.0 ± 1.4	85.8 ± 1.6	84.7 ± 2.0	82.4 ± 1.6			
RadGraph F1	26.2 ± 9.9	25.9 ± 10.7	26.2 ± 12.4	20.4 ± 9.9	10.3 ± 7.0			
MedAlpaca								
BLEU	0.9 ± 2.8	0.5 ± 1.6	0.7 ± 2.0	0.1 ± 0.8	0.0 ± 0.0			
METEOR	12.3 ± 7.4	13.9 ± 8.9	13.1 ± 9.4	9.2 ± 6.1	5.6 ± 3.2			
ROUGE-L	9.9 ± 7.3	10.4 ± 7.9	9.4 ± 7.5	6.7 ± 4.9	3.9 ± 2.7			
BERT Score	81.3 ± 2.7	81.6 ± 2.8	81.4 ± 3.1	80.4 ± 2.2	79.4 ± 1.6			
RadGraph F1	7.1 ± 8.2	7.4 ± 7.1	7.9 ± 8.3	3.2 ± 3.9	1.0 ± 2.1			
	Zephyr							
BLEU	4.4 ± 6.3	4.0 ± 4.5	4.4 ± 5.0	1.3 ± 2.7	0.1 ± 0.6			
METEOR	23.5 ± 8.3	22.8 ± 7.6	22.6 ± 9.1	18.1 ± 6.6	12.8 ± 5.8			
ROUGE-L	20.2 ± 7.8	19.6 ± 7.4	19.0 ± 7.6	14.6 ± 5.5	10.6 ± 5.1			
BERT Score	83.6 ± 2.8	83.9 ± 2.5	83.5 ± 2.9	82.8 ± 2.0	81.5 ± 1.7			
RadGraph F1	16.5 ± 9.6	16.4 ± 9.8	14.8 ± 10.8	7.9 ± 7.0	3.3 ± 4.7			
Flan-T5								
BLEU	9.6 ± 9.3	10.3 ± 10.2	9.8 ± 8.7	9.3 ± 8.2	9.9 ± 9.3			
METEOR	37.0 ± 14.8	37.9 ± 15.3	37.5 ± 15.0	37.8 ± 15.0	37.7 ± 14.0			
ROUGE-L	22.8 ± 10.2	23.2 ± 10.4	22.7 ± 9.6	22.6 ± 9.8	22.8 ± 10.0			
BERT Score	85.8 ± 2.2	86.1 ± 2.2	85.9 ± 2.3	86.0 ± 2.3	85.9 ± 2.1			
RadGraph F1	18.2 ± 10.3	19.5 ± 10.9	18.3 ± 10.0	17.9 ± 10.4	18.5 ± 9.7			
Temp: temperature, Repetition Penalty: RP, BLEU: Bilingual Evaluation Understudy, METEOR: Metric								

for Evaluation of Translation with Explicit ORdering, ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence, BERT: Bidirectional Encoder Representations from Transformers.

1 2

Supplemental Table 2. Representative Example of Zero-shot Results

Model	Ground Truth	LLaMA-2	Flan-T5	Zephyr	MedAlpaca
Representati ve Example	Global hypokinesis with superimposed regional variation (please see wall motion graphics). estimated right ventricular systolic pressure 32 (systolic blood pressure 120). no pericardial effusion. no significant valvular heart disease. moderate-severely enlarged left ventricular chamber size; calculated ejection fraction 25%. normal right ventricular chamber size. with borderline decrease in systolic function. with normal inspiratory collapse (>50%) inferior vena cava size normal. indeterminate left ventricular diastolic function."	The patient presents with global hypokinesia and regional variation in left ventricular wall motion. There is no evidence of intracardiac mass or thrombus, but the left atrial appendage cannot be adequately visualized for thrombus exclusion. The estimated right ventricular systolic pressure is 32 mmHg, and there is an anterior epicardial fat layer. The pulmonary valve is partially visualized, and the abdominal aorta has an incomplete Doppler flow pattern. The left ventricular wall thickness is normal, and there is no atrial level shunt. However, the left atrium volume index is slightly elevated, and there is mild mitral valve regurgitation and trivial pulmonary valve regurgitation. Additionally, the patient has moderate-severe left ventricular chamber enlargement, normal right ventricular chamber size, and normal aortic valve function. Overall, the patient's cardiac function appears compromised, likely due to chronic heart failure. Further	Calculated 2-d linear left ventricular ejection fraction 25. global hypokinesis with superimposed regional variation. No intracardiac mass or thrombus.	LVEF: 25% with superimposed regional variation (hypokinesis) - No intracardiac mass, but LAA cannot be ruled out for thrombus (transthoracic only) - Mild MR; trivial TR and PVR - Anterior epicardial fat; pulmonary valve not well visualized - Normal RV systolic pressures, AV, and RA sizes - Moderate- severe chamber enlargement; borderline RV syst.	Summary: This is a good study from which I can conclude that the patient has mild MR and TR. His LVESD is slightly above normal and his RVSV is slightly decreased. His PVR is also low. He does have a thrombus in the left atrial appendage, although it may not be noticeable on this particular study. His pulmonary artery velocity is normal but the abdominal aorta could not be measured adequately. So more information is needed for a full diagnosis. There is some fat in the epicardium, which is also nonspecific and might be seen in many patients. There is evidence of LV strain but again, more information is needed. The LA size is normal as is the RV size, and the IVC was normal. The aorta was normal in size. In general, there are no significant abnormalities and no cause for concern at this time.

	evaluation and management are warranted.			
Comments	 Summary of the findings with transitioning words that did not alter the original meaning.	Concise summary but omitted clinical important information.	Concise bullet point summary with false information.	Medical Professional- sounding, but hallucinated information.

2 Supplemental Table 3. Representative Example: In-context Learning Integrated Information from

3 Multiple Examples

	Sentence	Model Summary
Case	Calculated 2-d linear left ventricular ejection fraction 57.	"Calculated 2-d linear left ventricular ejection fraction 57,
ICL Examples	Calculated 2-d linear left ventricular ejection fraction 56.	61, 75 (3 reports)"
	Calculated 2-d linear left ventricular ejection fraction 75	
	Calculated 2-d linear left ventricular ejection fraction 69	
	Calculated 2-d linear left ventricular ejection fraction 61	

1 Supplemental Table 4. Performance of Fine-tuned LLMs on the AZ Validation and MIMIC-

2 EchoNote Datasets

1		
5	ι.	
	,	
	5	3

	Dataset	Dataset Mayo AZ vali			MI	te	
	Model	LLaMA-2	Zephyr		LLaMA-2	Zephyr	
		QLoRA	QLoRA	P-value*	QLoRA	QLoRA	P-value*
	BLEU	37.8 ± 17.5	17.6 ± 13.8	< 0.0001	9.6 ± 10.6	3.6 ± 4.5	< 0.0001
	METEOR	54.4 ± 17.3	32.9 ± 14.9	< 0.0001	40.2 ± 13.3	23.3 ± 10.1	< 0.0001
	ROUGE-L	47.4 ± 16.9	29.1 ± 13.7	< 0.0001	25.1 ± 12.8	14.6 ± 7.2	< 0.0001
	BERT Score	90.1 ± 2.9	86.6 ± 3.8	< 0.0001	85.4 ± 2.0	80.2 ± 15.2	< 0.0001
	RadGraph F1	39.0 ± 14.5	25.4 ± 12.2	< 0.0001	25.1 ± 13.9	15.4 ± 9.8	< 0.0001
	*Compared performance of QLoRA Llama-2 and Zephyr						
Supplemental Table 5. Agreement of the Ratings Between Echo-Expert Readers.							
	Metric Fleiss' Kappa between 4 rate					4 raters	
				Mayo	MI	MIC-Echo	
		Completene	ess	0.49		0.49	

21

20

- 22
- 23
- 24
- 25

26 References

27 1. Daubert MA, Tailor T, James O, Shaw LJ, Douglas PS, Koweek L. Multimodality cardiac imaging in

0.22

0.17

0.34

0.60

0.12

0.14

the 21st century: evolution, advances and future opportunities for innovation. Br J Radiol.

Conciseness

Correctness

Clinical Utility

29 2021;94:20200780.

30 2. Carli MFD, Geva T, Davidoff R. The Future of Cardiovascular Imaging. Circulation. 2016;133:2640-

- 1 2661.
- 2 3. Reeves RA, Halpern EJ, Rao VM. Cardiac Imaging Trends from 2010 to 2019 in the Medicare
- 3 Population. *Radiol Cardiothorac Imaging*. 2021;3:e210156.
- 4 4. Tiver KD, Horsfall M, Swan A, Pasquale CD, Horsfall E, Chew DP, Pasquale CGD. Accuracy of
- 5 Highly Limited Echocardiographic Screening Images for Determining a Structurally Normal Heart: The
- 6 Quick-Six Study. *Hear Lung Circ*. 2022;31:462–468.
- 7 5. Habash-Bseiso DE, Rokey R, Berger CJ, Weier AW, Chyou P-H. Accuracy of Noninvasive Ejection
- 8 Fraction Measurement in a Large Community-Based Clinic. *Clin Medicine Res.* 2005;3:75–82.
- 9 6. Berlin L. Defending the "Missed" Radiographic Diagnosis. *Am J Roentgenol*. 2001;176:317–322.
- 10 7. Berlin L, Hendrix RW. Perceptual errors and negligence. Am J Roentgenol. 1998;170:863–867.
- 11 8. Tromp J, Seekings PJ, Hung C-L, Iversen MB, Frost MJ, Ouwerkerk W, Jiang Z, Eisenhaber F, Goh
- 12 RSM, Zhao H, Huang W, Ling L-H, Sim D, Cozzone P, Richards AM, Lee HK, Solomon SD, Lam CSP,
- 13 Ezekowitz JA. Automated interpretation of systolic and diastolic function on the echocardiogram: a
- 14 multicohort study. *Lancet Digital Heal*. 2022;4:e46–e54.
- 15 9. Nolan MT, Thavendiranathan P. Automated Quantification in Echocardiography. *JACC Cardiovasc*
- 16 *Imaging*. 2019;12:1073–1092.
- 17 10. Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, Liang DH, Ashley EA, Zou JY.
- 18 Deep learning interpretation of echocardiograms. *Npj Digital Medicine*. 2020;3:10.
- 19 11. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras
- 20 MA, Jordan C, Fleischmann KE, Melisko M, Qasim A, Shah SJ, Bajcsy R, Deo RC. Fully Automated
- 21 Echocardiogram Interpretation in Clinical Practice. Circulation. 2018;138:1623–1635.
- 22 12. Rajpurkar P, Lungren MP. The Current and Future State of AI Interpretation of Medical Images. New
- 23 Engl J Med. 2023;388:1981–1990.
- 24 13. Veen DV, Uden CV, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M,
- 25 Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Clinical Text
- 26 Summarization: Adapting Large Language Models Can Outperform Human Experts. *arXiv*. 2023;
- 27 14. Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z, Ma C, Shu P, Chen C, Kim S, Dai H, Zhao L, Zhu D,
- Liu J, Liu W, Shen D, Li X, Li Q, Liu T. Radiology-GPT: A Large Language Model for Radiology.
- **29** *arXiv*. 2023;
- 30 15. Gershanik EF, Lacson R, Khorasani R. Critical finding capture in the impression section of radiology
- 31 reports. AMIA Annu Symp Proc AMIA Symp. 2011;2011:465–9.
- 32 16. Fleming SL, Lozano A, Haberkorn WJ, Jindal JA, Reis EP, Thapa R, Blankemeier L, Genkins JZ,
- 33 Steinberg E, Nayak A, Patel BS, Chiang C-C, Callahan A, Huo Z, Gatidis S, Adams SJ, Fayanju O, Shah
- 34 SJ, Savage T, Goh E, Chaudhari AS, Aghaeepour N, Sharp C, Pfeffer MA, Liang P, Chen JH, Morse KE,
- 35 Brunskill EP, Fries JA, Shah NH. MedAlign: A Clinician-Generated Dataset for Instruction Following
- 36 with Electronic Medical Records. *arXiv*. 2023;
- 37 17. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D,
- 38 Schaekermann M, Wang A, Amin M, Lachgar S, Mansfield P, Prakash S, Green B, Dominowska E,
- 39 Arcas BA y, Tomasev N, Liu Y, Wong R, Semturs C, Mahdavi SS, Barral J, Webster D, Corrado GS,
- 40 Matias Y, Azizi S, Karthikesalingam A, Natarajan V. Towards Expert-Level Medical Question
- 41 Answering with Large Language Models. *arXiv*. 2023;
- 42 18. Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z, Ma C, Shu P, Chen C, Kim S, Dai H, Zhao L, Zhu D,
- 43 Liu J, Liu W, Shen D, Li X, Li Q, Liu T. Radiology-GPT: A Large Language Model for Radiology.
- 44 *arXiv*. 2023;

- 1 19. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages,
- 2 limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595.
- 3 20. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, He H, Li A, He M, Liu Z, Wu Z, Zhao L, Zhu D, Li
- 4 X, Qiang N, Shen D, Liu T, Ge B. Summary of ChatGPT-Related research and perspective towards the
- 5 future of large language models. *Meta-Radiol.* 2023;1:100017.
- 6 21. Diaz N. 6 hospitals, health systems testing out ChatGPT [Internet]. 2023 [cited 2023 Jun 2]; Available
- 7 from: https://www.beckershospitalreview.com/innovation/4-hospitals-health-systems-testing-out-
- 8 chatgpt.html
- 9 22. Latif E, Zhai X. Fine-tuning ChatGPT for Automatic Scoring. *arXiv*. 2023;
- 10 23. Nath C, Albaghdadi MS, Jonnalagadda SR. A Natural Language Processing Tool for Large-Scale
- 11 Data Extraction from Echocardiography Reports. *PLoS ONE*. 2016;11:e0153749.
- 12 24. Dong T, Sunderland N, Nightingale A, Fudulu DP, Chan J, Zhai B, Freitas A, Caputo M, Dimagli A,
- 13 Mires S, Wyatt M, Benedetto U, Angelini GD. Development and Evaluation of a Natural Language
- 14 Processing System for Curating a Trans-Thoracic Echocardiogram (TTE) Database. *Bioengineering*.
- 15 2023;10:1307.
- 16 25. Zheng C, Sun BC, Wu Y-L, Ferencik M, Lee M-S, Redberg RF, Kawatkar AA, Musigdilok VV,
- 17 Sharp AL. Automated interpretation of stress echocardiography reports using natural language
- 18 processing. *Eur Hear J Digit Heal*. 2022;3:626–637.
- 19 26. Tang L, Kooragayalu S, Wang Y, Ding Y, Durrett G, Rousseau JF, Peng Y. EchoGen: Generating
- 20 Conclusions from Echocardiogram Notes. Proc 21st Work Biomed Lang Process. 2022;2022:359–368.
- 21 27. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P,
- 22 Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W,
- 23 Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V,
- 24 Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux M-A, Lavril T, Lee J, Liskovich D, Lu Y, Mao
- 25 Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K,
- 26 Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P,
- 27 Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T.
- 28 Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv. 2023;
- 28. Tunstall L, Beeching E, Lambert N, Rajani N, Rasul K, Belkada Y, Huang S, Werra L von, Fourrier
- 30 C, Habib N, Sarrazin N, Sanseviero O, Rush AM, Wolf T. Zephyr: Direct Distillation of LM Alignment.
- 31 *arXiv*. 2023;
- 32 29. Gardin JM, Adams DB, Douglas PS, Feigenbaum H, Forst DH, Fraser AG, Grayburn PA, Katz AS,
- 33 Keller AM, Kerber RE, Khandheria BK, Klein AL, Lang RM, Pierard LA, Quinones MA, Schnittger I,
- 34 Echocardiography AS of. Recommendations for a standardized report for adult transthoracic
- 35 echocardiography: A report from the American Society of Echocardiography's Nomenclature and
- 36 Standards Committee and Task Force for a Standardized Echocardiography Report. J Am Soc
- 37 *Echocardiogr.* 2002;15:275–290.
- 30. Kwak G, Moukheiber D, Moukheiber M, Moukheiber L, Moukheiber S, Butala N, Celi L, Chen C.
- 39 EchoNotes Structured Database derived from MIMIC-III (ECHO-NOTE2NUM). *PhysioNet*. 2024;
- 40 31. Loper E, Bird S. NLTK: the Natural Language Toolkit. *Proc ACL-02 Work Eff tools Methodol Teach*
- 41 Nat Lang Process Comput linguistics -. 2002;63–70.
- 42 32. Han T, Adams LC, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, Truhn D, Bressem KK.
- 43 MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data.
- 44 *arXiv*. 2023;

- 1 33. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the
- 2 Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv*. 2019;
- 3 34. Cowan BR, Clark L, Følstad A, Skjuve M. Chatbots for customer service: user experience and
- 4 motivation. Proc 1st Int Conf Conversational User Interfaces. 2019;1–9.
- 5 35. Li M, Gong S, Feng J, Xu Y, Zhang J, Wu Z, Kong L. In-Context Learning with Many
- 6 Demonstration Examples. *arXiv*. 2023;
- 7 36. Choi E, Jo Y, Jang J, Seo M. Prompt Injection: Parameterization of Fixed Inputs. *arXiv*. 2022;
- 8 37. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-Specific
- 9 Language Model Pretraining for Biomedical Natural Language Processing. *arXiv*. 2020;
- 10 38. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized
- 11 LLMs. arXiv. 2023;
- 12 39. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-Rank
- 13 Adaptation of Large Language Models. *arXiv*. 2021;
- 14 40. Veen DV, Uden CV, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M, Reis
- 15 EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly
- 16 J, Chaudhari AS. Adapted large language models can outperform medical experts in clinical text
- 17 summarization. *Nat Med*. 2024;1–9.
- 18 41. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, Xu Z, Ding Y, Durrett G, Rousseau JF,
- Weng C, Peng Y. Evaluating large language models on medical evidence summarization. *npj Digit Med.*2023;6:158.
- 21 42. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine
- translation. Proc 40th Annu Meet Assoc Comput Linguistics ACL '02. 2002;311–318.
- 23 43. Lavie A, Agarwal A. Meteor: an automatic metric for MT evaluation with high levels of correlation
- 24 with human judgments. 2007;228–231.
- 25 44. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with
- 26 BERT. *arXiv*. 2019;
- 27 45. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries [Internet]. Association for
- 28 Computational Linguistics; p. 74–81. Available from: https://aclanthology.org/W04-1013
- 29 46. Delbrouck J-B, Chambon P, Bluethgen C, Tsai E, Almusa O, Langlotz CP. Improving the Factual
- 30 Correctness of Radiology Report Generation with Semantic Rewards. arXiv. 2022;
- 31 47. Jain S, Agrawal A, Saporta A, Truong SQ, Duong DN, Bui T, Chambon P, Zhang Y, Lungren MP,
- Ng AY, Langlotz CP, Rajpurkar P. RadGraph: Extracting Clinical Entities and Relations from Radiology
 Reports. *arXiv*. 2021;
- 48. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22:276–82.
- 35 49. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*.
- 36 1977;33:159–74.
- 37 50. Zhao TZ, Wallace E, Feng S, Klein D, Singh S. Calibrate Before Use: Improving Few-Shot
- **38** Performance of Language Models. *arXiv*. 2021;
- 39 51. Wang L, Chen X, Deng X, Wen H, You M, Liu W, Li Q, Li J. Prompt engineering in consistency and
- 40 reliability with the evidence-based guideline for LLMs. npj Digit Med. 2024;7:41.
- 41 52. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation
- 42 models for generalist medical artificial intelligence. *Nature*. 2023;616:259–265.
- 43 53. Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. ChatCAD: Interactive Computer-Aided Diagnosis on
- 44 Medical Image using Large Language Models. *arXiv*. 2023;

- 1 54. Mallio CA, Sertorio AC, Bernetti C, Zobel BB. Large language models for structured reporting in
- 2 radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. *Radiol Med.* 2023;1–5.
- 3 55. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W,
- 4 Rocktäschel T, Riedel S, Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

5 *arXiv*. 2020;

Open Source LLMs (LLaMA-2, Zephyr, MedAlpaca, Flan T5)



Completeness Qualitative evaluation Conciseness 30 Cases Correctness **EchoGPT Clinical Utility**

Experts: the 4C metrics

Figure 2

















RadGraph F1 Win Rate Comparisons

Figure 4 - win rate























Figure 5- Human eval



Findings: no intracardiac mass or thrombus, but the left atrial appendage cannot be visualized adequately with transthoracic echo to exclude thrombus in this location. normal left ventricular wall thickness. colorflow and spectral doppler were performed to assess valvular heart disease. echocardiographic images interpreted at mcf - clinic campus. no pericardial effusion. normal left ventricular chamber size. normal right ventricular chamber size. normal right ventricular chamber size. normal right ventricular systolic function. normal left ventricular diastolic function. calculated 2-d biplane volumetric left ventricular ejection fraction of 62. global averaged left ventricular longitudinal peak systolic strain is at -20 (normal = more negative than -18%).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.18.24301503; this version posted April 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Summary A: no pericardial effusion. normal left ventricular chamber size. , wall thickness and regional wall motion. ef 62% normal right ventricular chamber size. and function normal left ventricular diastolic function. global averaged left ventricular longitudinal peak systolic strain is at -20 (normal = more negative than -18%). strain is slightly worse at the base compared to the prior study and overall global average strain has declined, but still remains in the normal range

Summary B: no regional wall motion abnormalities. normal cardiac valves. normal sized atria. no pericardial effusion. normal left ventricular chamber size. calculated left ventricular ejection fraction; 62%. normal right ventricular chamber size. and systolic function. unable to estimate right ventricular systolic pressure. global averaged left ventricular longitudinal peak systolic strain is at -20 (normal = more negative than -18%). *



Supplemental Figure 1





Supplemental Figure 2- bias





Α







Supplemental Figure 3 - Human eval on MIMIC-Echo

