Title: EchoGPT: A Large Language Model for Echocardiography Report Summarization

Authors: Chieh-Ju Chao, MD[1,2], Imon Banerjee, PhD[3], Reza Arsanjani, MD[4], Chadi Ayoub, MD, PhD[4], Andrew Tseng, MD[5], Garvan C. Kane, MD, PhD[1], Jae K Oh, MD[1], Li Fei-Fei, PhD[2], Ehsan Adeli, PhD[2] and Curtis Langlotz, MD, PhD[6]

Affiliations:
[1] Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota
[2] Stanford Institute for Human-Centered Artificial Intelligence, Palo Alto, California
[3] Department of Radiology, Mayo Clinic, Scottsdale, Arizona
[4] Department of Cardiovascular Diseases, Mayo Clinic Arizona, Scottsdale, Arizona
[5]Department of Cardiovascular Medicine, Mayo Clinic Florida, Jacksonville, Florida
[6] Center for Artificial Intelligence in Medicine and Imaging (AIMI), Stanford University, Palo Alto, California

Corresponding author:
Chieh-Ju Chao, MD
Address: Gonda 4-478, 200 1st Street SW, Rochester, MN 55095
Email: chao.chiehju@mayo.edu

Keywords: Artificial Intelligence, Large Language Model, Echocardiography, Natural Language Processing, Cardiovascular Imaging

Abbreviations:

4C metrics: qualitative review metrics including completeness, correctness, conciseness, and clinical utility
AI: Artificial Intelligence
DL: Deep Learning
Echo: Echocardiography
FT: Fine-Tuning
GPT: Generative Pre-trained Transformer
ICL: In-Context Learning
NLP: Natural Language Processing
Seq2seq: Sequence-to-sequence
TTE: Transthoracic Echocardiography
TEE: Transesophageal Echocardiography
LLM: Large Language Model
NLP: Natural Language Processing
QLoRA: Quantized Low-Rank Adaption

Abstract

Background

The increasing need for diagnostic echocardiography (echo) tests presents challenges in preserving the quality and promptness of reports. While Large Language Models (LLMs) have proven effective in summarizing clinical texts, their application in echo remains underexplored. To address this, we proposed EchoGPT, a dedicated, domain-specific LLM focused on echo report summarization.

Methods

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

Results

The development dataset included 97,506 reports from 71,717 unique patients, predominantly male (54.3%), with an average age of 64.1±16.1 years. The final split contains 95,506 for training, and 1,000 each for validation and testing. EchoGPT, a QLoRA fine-tuned Llama-2 model, outperformed other LLMs with about 90% win rates in various metrics (BLEU, METEOR, ROUGE-L, BERT Score, and RadGraph F1 Score), and produced reports comparable to cardiologists in 30 randomly selected cases for qualitative human review (significantly preferred in conciseness ($p< 0.001$), with no significant preference in completeness, correctness, and clinical utility). In the external validation set (n=1,000), EchoGPT consistently outperformed fine-tuned Zephyr model across the same automatic metrics (all $p < 0.0001$).

Conclusions

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

Clinical Perspective

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.
2. What are the clinical implications?
    - EchoGPT system demonstrated the potential of introducing LLMs into echocardiography practice.
    - EchoGPT could be used as an AI co-pilot to generate echo reports.

Introduction

Echocardiography (echo) is the mainstay imaging modality in the current practice of cardiology[1], providing vital, non-invasive assessments of heart anatomy and physiology to guide clinical decisions[2]. In the past decade, the rising demand for diagnostic echo tests[3] 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].

With the recent emergence of artificial intelligence (AI), automated echo reporting has been proposed to use deep learning (DL) models to generate diagnostic predictions and measurements to fill a pre-set report template[8,10,11]. These frameworks focused on specific image processing tasks[8,11] rather than the report text, and are technically equivalent to generating individual findings. However, these frameworks were not designed to handle the high-level cognitive activity of synthesizing clinically relevant impressions from detailed findings[12]. In practice, physicians usually spend a significant amount of time summarizing detailed findings to clinically relevant final impressions[13,14]. While this task is crucial, it can be time-consuming and prone to errors[15].

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-answering[17]. A major advantage of LLMs is the flexibility of input and output[18], as well as the capability of handling conversations and interaction with human experts[19]. While similar functionality can be achieved through commercially available LLMs (e.g., ChatGPT; OpenAI, San Francisco, CA)[20], only a few healthcare institutions have integrated ChatGPT[21]. Furthermore, fine-tuning ChatGPT for specific tasks still requires uploading data to a central server, which also raises privacy concerns[22]. In contrast, open-source LLMs are free of charge and can be locally fine-tuned for specific tasks within each healthcare institution's secure confines[18].

Previous studies predominantly focused on electronic health records[13,16] and chest X-rays (CXR)[13,18] have highlighted the potential of using LLMs to summarize in clinical text. In contrast, echo-related studies were mainly on data extraction or classification, rather than report summarization[23–25]. Tang et al. used rule-based systems and the BART (Bidirectional and Auto-Regressive Transformer) model[26] for this purpose and demonstrated convincing results. However, BART-generated content was less favored by human experts more than 50% of the time, perhaps due to the smaller number of parameters than current state-of-the-art LLMs[26]. The application of billion-parameter LLMs to generate echo reports remains under-explored[27,28].

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.

Method

Echocardiography Report

Following the American Society of Echocardiography recommendations[29], a standard echocardiography report at the Mayo Clinic contains the following major sections: Final Impressions, Findings, and Measurements. The Measurements section contains only measurement values without free text. The key

measurements such as left ventricular ejection fraction, aortic valve area, and right ventricular systolic pressure are included in the Findings section with corresponding statements. Considering the report structure above, only the information from the Final Impressions and Findings sections was used in this work (**Figure 1**).

Dataset

All adult (> 18 years old) echocardiography studies performed from 1/1/2017 to 12/31/2017 at Mayo Clinic Enterprise were retrieved. The types of studies include transthoracic echocardiography (TTE), transesophageal echocardiography (TEE), stress echocardiography, and intracardiac echocardiography. Text in the "Findings" and the "Final Impression" sections of each report was extracted for the current study (**Figure 2**). The study was approved by Mayo Clinic IRB (protocol#: 22-010944).

Data Curation and Preprocessing

Echocardiography reports were excluded according to the following criteria: (1) 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 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 VENTRICLE, VALVES, OTHER FINDINGS, etc.), (2) Remove template sentences that make comparisons to prior reports, as no information from previous reports has been provided, and (3) Remove quality control-related sentences such as "study performed per left ventricular function protocol" and "the goals, risks, and alternatives to moderate sedation were explained to the patient."

Data split

Data from Rochester, Florida, and Mayo Clinic Healthcare sites were used as the model development set. The Arizona site data was used as the external validation set. Within the development set, 1,000 non-duplicated cases were randomly selected for the test and validation sets, respectively; the rest of the cases were used for fine-tuning (training set). Similarly, 1,000 non-duplicated random examples were selected from the Arizona external validation dataset. For basic dataset statistics, the token length was calculated based on the natural language processing toolkit (NLTK) tokenizer[30], and the lexical variance was defined as the ratio of the number of unique tokens to the number of total tokens in each example[13].

Model selection

Due to patient privacy policy regulations, proprietary LLMs such as GPT-3.5 and GPT-4 were not considered in this work because versions of those models that were safe for protected health information were not yet available. Among open-source models, we selected representative auto-regressive models including Llama-2-7b-chat[27], Zephyr-7b[28], and Med-Alpaca[31] models considering their performance and max input context length on general natural language processing (NLP) tasks and radiology report summarization [13]. For sequence-to-sequence (seq2seq) models, we used Flan-T5 (base) as the representative model as it is known for accurate text summarization[13,32].

Model inference hyperparameter search

LLM inference was conducted by using HuggingFace's (Manhattan, NY) transformer pipeline via the open-source LangChain framework[33]. After initial tests, text generation and summarization were used as

the task type for auto-regressive models and seq2seq models, respectively. A subset (10%, n=100) of examples were randomly selected from the test set for the hyperparameter search. We specifically tested the following configuration parameters that can significantly affect performance: temperature (0.1, 0.5, and 0.9) and repetition penalty (1.1, 1.2, and 1.3). These two parameters were tested separately, when one 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 model inference: {temperature 0.1, repetition penalty 1.1} after comparing automatic metrics and 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 to prior reports, and the generated contents were satisfying on qualitative review. Of note, the configurations were tested in a zero-shot setting, and the best configuration was directly applied to the ICL and QLoRA fine-tuned models[13].

Model Adaptation

In-context Learning (ICL): ICL has been proposed to improve LLM's performance without changing the base model weights[13,34,35]. Also, using relevant in-context examples is shown to have better model 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 model[36]. A prompt template was created with components of the prefix, instruction, and suffix[13] (**Table 1**). The final prompt was decided after qualitatively evaluating several different variants of each component on a small subset of the data. We also specified that the summarization should be "concise" and use "a minimal amount of text" to avoid LLMs generating lengthy reports[13].

Instruction tuning with quantized low-rank adaptation (QLoRA):

Due to the size of candidate models, we opted for quantized low-rank adaptation (QLoRA)[37], a type of parameter efficient fine-tuning (PEFT)[38] to optimize our LLMs for echo report summarization tasks. The same prompt template was used, and the Final Impression text from the same report was used as the target output [13,14].

We configured the training process as follows: load model in 4-bit precision, with a LoRA configuration 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 provided the most stable training process after attempting different configurations reported in prior studies[13,37].

Model Performance Evaluation

Automatic NLP evaluation metrics: To evaluate the models' performance on the information summarization task and compare it to prior works, we utilized four established automatic metrics: BLEU (Bilingual Evaluation Understudy)[39], METEOR (Metric for Evaluation of Translation with Explicit ORdering)[40], ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence), and the BERT (Bidirectional Encoder Representations from Transformers) score[41]. For

ROUGE-L, we present the F1 score component[26,42]. For factual correctness, the RadGraph-F1 metric (level: all) was reported[43,44]. Considering the importance of measurements in echo studies, we attempted to evaluate whether the current automatic metrics can detect changes in measurement numbers. For this purpose, we generated synthetic reports by replacing all the measurement numbers with random numbers ranging from 1 to 99 in the reports. We then compared the automatic metric scores with the corresponding reports with the original measurements.

Human expert evaluation metrics: We designed a human expert evaluation process based on previous clinical text summarization studies[13,18,26]. The Findings with corresponding ground truth (Final Impression) and LLM-generated summarization of 30 randomly selected cases were presented to four echocardiography-board-certified cardiologists for blinded quality review. We noted that physician-summarized Impressions may contain free-text information beyond the Findings (e.g., documenting events during the study or communication with the ordering provider), and the reviewers were instructed to ignore these. Each metric was rated for the preference (5 levels) between the two summarizations (**Supplemental Figure 1**)[13]. The "4C metrics" evaluated by echocardiography experts are completeness, conciseness, correctness, and clinical utility[14], as described in **Table 2**.

Automatic metric performance between each pair of models was compared using a two-tailed paired Student's t-test or Wilcoxon signed-rank test for normally and non-normally distributed data, respectively. Models were also compared based on win rates, which are defined as the percentage of head-to-head victories in performance between two models for each selected metric[13]. In human expert qualitative analysis, the 4C metrics were compared by a one-sample Wilcoxon signed-rank test[13], and the agreement of ratings between experts was assessed by Fleiss' kappa coefficient[45], and interpreted as recommended by Landis et al[46]. We conducted Pearson correlation analyses to explore the relationships between each human expert evaluation metric, assessing their independence. Additionally, we conducted similar analyses to examine the correlation between human and automatic metrics. All the comparisons consider a p-value < 0.05 as significant.

Results

Patient cohort and dataset

Our development set contains 97,506 reports from 71,717 unique patients with a mean age of 64.1±16.1 years, and 38,912 (54.3%) were male. The external validation dataset contains 19,557 reports from 15,853 unique patients with a mean age of 61.2 ±15.7 years, and 8,970 (56.6%) were male. Statistics of text data are summarized in **Table 3**. Transthoracic echocardiography was the predominant study type in both development and external validation datasets (81.9% and 76.5%, respectively); detailed distribution was summarized in **Supplemental Table 2**.

Zero-shot, ICL, and QLoRA fine-tuned performance

Table 4 is a summary of the performance of zero-shot and fine-tuned LLMs, including MedAlpaca, Llama-2, and Zephyr. QLoRA fine-tuning significantly improved LLMs' performance from baseline. Note that T5 and MedAlpaca were not fine-tuned so only zero-shot results were provided for reference. Among the candidate models, Llama-2 generally had the best zero-shot performance, which was consistent in ICL. While Flan-T5 had a similar or superior performance to Llama-2 across most metrics, it was particularly worse on the RadGraph F1 score (**Table 4; Figure 3**). On qualitative review, we noted

that T5 provided concise summaries, however important clinical information was missed in this process (**Supplemental Table 3**).

In ICL, LLMs that allow longer context length (Llama-2 and Zephyr) had the best performance across all metrics when one example was provided (ICL-1). The performance then gradually trended down with more examples (ICL-2 and ICL-4). In contrast, the performance of LLMs with shorter max context length started to trend down with one example (**Figure 3**).

Based on the zero-shot and ICL performance of the candidate models, Llama-2 and Zephyr were selected for instruction fine-tuning. Compared to zero-shot, QLoRA significantly improved the performance of selected LLMs across all metrics (**Table 4**). For the head-to-head comparison in model win rates, fine-tuned Llama-2 was superior to all other models, including Zephyr (base and fine-tuned), MedAlpaca (base), and Flan-T5 (base) across all 5 automatic metrics (**Figure 4**). Llama-2 maintained similar performance in the external validation data set (n=1,000) and was consistently superior to fine-tuned Zephyr (**Supplemental Table 5**). Because Llama-2 had the best performance on both zero-shot and ICL approaches, it was selected as EchoGPT and used for the subsequent expert qualitative review.

Significance of Measurement Numbers in Automatic Metrics

After replacing the measurement numbers with random numbers, we observed statistically significant ($p<0.0001$) but minor decreases in BLEU, METEOR, ROUGE-L, BERT, and RadGraph F1 scores. One can see that in the provided examples, the report with random numbers doesn't make clinical sense when compared to the original content (**Table 5**).

Human expert evaluation

We observed fair agreement for conciseness and correctness, moderate agreement for completeness, and good agreement for clinical utility (**Table 6**). Among the 4C metrics, we observed that EchoGPT significantly outperformed human experts in conciseness ($p<0.001$). There was no significant difference among the other three categories (**Figure 5A**). The 4C metrics were not completely 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 5B**). We also observed that across all automatic metrics, RadGraph F1 had modest to moderate correlations with all 4 human evaluation metrics (**Figure 5C**).

Discussion

To the best of our knowledge, this study is the first attempt to compare multiple open-source LLMs in echocardiography report summarization through different model adaptation methods. Trained on one of the largest echocardiography report datasets in the world, we demonstrated that QLoRA fine-tuning can significantly improve LLMs' performance for the desired summarization task, with at least comparable qualities to human experts. Our results also indicate that ICL is associated with significant limitations for clinical practice, including compromised patient privacy and report quality. Additionally, we demonstrated that current automatic metrics are not sensitive to the change in measurement numbers in echo reports. The current study provides insights into the construction of a dedicated local LLM for echo

report summarization and can pave the way for an AI-enabled echocardiography interpretation system with a human-AI interaction interface.

Model adaption approaches: ICL vs. fine-tuning

Our results suggested that both ICL and QLoRA fine-tuning improved LLMs' performance over zero-shot, and fine-tuning performance was consistently above ICL across all automatic metrics (**Figure 3**). Additionally, we observed substantial limitations of ICL for echo reporting, including relatively longer context length and compromised patient privacy, as discussed below.

Compared to CXR reports, echocardiography reports come with a relatively longer context, which directly affects the available choices of LLMs. In contrast to a prior study that used 32 or more examples[13], we were only able to test up to 4 examples for ICL, therefore not able to assess the models' behavior with more examples. However, across all metrics, LLMs' had gradually down-trending performance when more examples were provided (**Figure 3**), which is consistent with prior studies[13,34]. It is also important to consider that the computation time and resources required in ICL can increase with the number of examples used[34]. Additionally, we note that LLMs integrate information from the example ICL cases, which compromises the report quality as well as patient privacy (**Supplemental Table 4**). While this behavior was not reported in other studies[13,34], we believe it could be a common condition, that is easier to identify with numerical values (in echo) compared to narrative statements (in CXR). Therefore, even in scenarios where ICL can outperform fine-tuning[13], fine-tuning may be preferable.

The EchoGen study previously demonstrated that BART and other rule-based approaches are effective for summarizing echocardiogram reports, with BART achieving ROUGE-based scores between 0.65 and 0.73[26]. Although EchoGPT didn't match these scores in ROUGE-L, it compared favorably to human experts in qualitative assessments. In contrast, human summaries were preferred a majority of the time over those generated by the BART model[26]. Additionally, in our study, we observed that T5 (as the representative seq2seq model) generated summaries that were overly brief so important clinical information was missed (**Supplemental Table 3**). Although the EchoGen authors did not provide qualitative examples generated by their BART model, we assume that similar behavior occurred with BART which led to the unfavorable rating by physicians.

Evaluation of Echo Report Summarization

Automatic evaluation of LLM in clinical text summarization tasks is an emerging area, and there is no single gold standard metric that can evaluate all aspects of a report[13,26]. Our study reinforces this conclusion. We noted that MedAlpaca and Zephyr can generate medical-professionally-sounding content that frequently includes hallucinated information. These differences were mainly reflected by the factual correctness metric RadGraph F1 (**Table 4**).

Concerning 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 correlated[13]. Notably, the highest correlation was observed between the RadGraph F1 scores and the 4C metrics, particularly in terms of clinical utility (r=0.42) (**Figure 5C**). This suggests that the quality of echo reports judged by cardiologists may not be well captured in automated metrics that do not capture notions of factual correctness.

As a specific subtype of clinical text, echocardiography reports contain unique terminology, including many precise measurements. Clinically, 25% and 55% LV ejection fraction values indicate a significant difference, however, our study demonstrates that this distinction is difficult to capture with current automatic metrics (**Table 5**). While this aspect of reporting can be easily captured in qualitative analyses, such analyses are expensive to conduct at scale because of the limited availability of in-domain experts. A dedicated metric for echocardiography diagnostic quality evaluation, with emphasis on measurement accuracy, is still needed to address this knowledge gap.

Application of EchoGPT

Constructing EchoGPT shows the feasibility of introducing LLMs into echocardiography practice. Through the QLoRA fine-tuning process[37], the model was able to learn clinically relevant knowledge to summarize echo report findings at a quality level comparable to echocardiography-trained cardiologists (**Figure 5A**).

The current study concentrated solely on the performance of EchoGPT in summarizing echocardiogram reports. However, as an LLM-based system, EchoGPT holds promise for broader applications, including providing in-context clinical reasoning, answering questions based on patient data, and interacting with human experts or other models[12,16,47]. In contemplating these expanded functionalities, two critical factors emerge: the maximum permissible context length and the model's proficiency in managing conversations. These considerations might restrict the use of existing seq2seq models in such scenarios[32].

We envision that EchoGPT could be used as a reporting interface, or a co-pilot that could generate echo reports with various inputs[48]. EchoGPT inherits the limitations of LLMs, including hallucination[13,18,49]. Although the number of hallucinations improved after the fine-tuning process, additional efforts such as optimization for factual correctness[43] or paired with a retrieval augmented generation system[50] are still required to minimize hallucinations before clinical implementation.

Conclusion

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

Limitations

Our echocardiography reports are based on standardized statements, with an option to add free text. While the lexical variance was high, the corpus could differ from reports composed entirely of free text contents. Due to patient privacy regulations, this work did not assess the performance of GPT-3.5 and GPT-4. However, we compared the performance of SOTA open-source LLMs, which provided important insights for model selection when data privacy is a critical consideration. Instead of full fine-tuning, QLoRA was used as the fine-tuning approach, however, it has been demonstrated as an effective approach as full fine-tuning is often not feasible for LLMs. Although QLoRA fine-tuning demonstrated improvements in echo report summarization tasks, our current approach does not include optimization for factual correctness and human expert preference.

Reference

1. Daubert MA, Tailor T, James O, Shaw LJ, Douglas PS, Koweek L. Multimodality cardiac imaging in the 21st century: evolution, advances and future opportunities for innovation. *Br J Radiol*. 2021;94:20200780.
2. Carli MFD, Geva T, Davidoff R. The Future of Cardiovascular Imaging. *Circulation*. 2016;133:2640–2661.
3. Reeves RA, Halpern EJ, Rao VM. Cardiac Imaging Trends from 2010 to 2019 in the Medicare Population. *Radiol Cardiothorac Imaging*. 2021;3:e210156.
4. Tiver KD, Horsfall M, Swan A, Pasquale CD, Horsfall E, Chew DP, Pasquale CGD. Accuracy of Highly Limited Echocardiographic Screening Images for Determining a Structurally Normal Heart: The Quick-Six Study. *Hear Lung Circ*. 2022;31:462–468.
5. Habash-Bseiso DE, Rokey R, Berger CJ, Weier AW, Chyou P-H. Accuracy of Noninvasive Ejection Fraction Measurement in a Large Community-Based Clinic. *Clin Medicine Res*. 2005;3:75–82.
6. Berlin L. Defending the "Missed" Radiographic Diagnosis. *Am J Roentgenol*. 2001;176:317–322.
7. Berlin L, Hendrix RW. Perceptual errors and negligence. Am J Roentgenol. 1998;170:863–867.
8. Tromp J, Seekings PJ, Hung C-L, Iversen MB, Frost MJ, Ouwerkerk W, Jiang Z, Eisenhaber F, Goh RSM, Zhao H, Huang W, Ling L-H, Sim D, Cozzone P, Richards AM, Lee HK, Solomon SD, Lam CSP, Ezekowitz JA. Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. *Lancet Digital Heal*. 2022;4:e46–e54.
9. Nolan MT, Thavendiranathan P. Automated Quantification in Echocardiography. *JACC Cardiovasc Imaging*. 2019;12:1073–1092.
10. Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, Liang DH, Ashley EA, Zou JY. Deep learning interpretation of echocardiograms. *Npj Digital Medicine*. 2020;3:10.
11. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, Fleischmann KE, Melisko M, Qasim A, Shah SJ, Bajcsy R, Deo RC. Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation. 2018;138:1623–1635.
12. Rajpurkar P, Lungren MP. The Current and Future State of AI Interpretation of Medical Images. *New Engl J Med*. 2023;388:1981–1990.
13. Veen DV, Uden CV, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. *arXiv*. 2023;
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. *arXiv*. 2023;
15. Gershanik EF, Lacson R, Khorasani R. Critical finding capture in the impression section of radiology reports. *AMIA Annu Symp Proc AMIA Symp*. 2011;2011:465–9.
16. Fleming SL, Lozano A, Haberkorn WJ, Jindal JA, Reis EP, Thapa R, Blankemeier L, Genkins JZ, Steinberg E, Nayak A, Patel BS, Chiang C-C, Callahan A, Huo Z, Gatidis S, Adams SJ, Fayanju O, Shah SJ, Savage T, Goh E, Chaudhari AS, Aghaeepour N, Sharp C, Pfeffer MA, Liang P, Chen JH, Morse KE, Brunskill EP, Fries JA, Shah NH. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. *arXiv*. 2023;
17. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D, Schaekermann M, Wang A, Amin M, Lachgar S, Mansfield P, Prakash S, Green B, Dominowska E,

Arcas BA y, Tomasev N, Liu Y, Wong R, Semturs C, Mahdavi SS, Barral J, Webster D, Corrado GS, Matias Y, Azizi S, Karthikesalingam A, Natarajan V. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv. 2023;
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, Liu J, Liu W, Shen D, Li X, Li Q, Liu T. Radiology-GPT: A Large Language Model for Radiology. *arXiv*. 2023;
19. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595.
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 X, Qiang N, Shen D, Liu T, Ge B. Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiol*. 2023;1:100017.
21. Diaz N. 6 hospitals, health systems testing out ChatGPT [Internet]. 2023 [cited 2023 Jun 2];Available from: https://www.beckershospitalreview.com/innovation/4-hospitals-health-systems-testing-out-chatgpt.html
22. Latif E, Zhai X. Fine-tuning ChatGPT for Automatic Scoring. *arXiv*. 2023;
23. Nath C, Albaghdadi MS, Jonnalagadda SR. A Natural Language Processing Tool for Large-Scale Data Extraction from Echocardiography Reports. *PLoS ONE*. 2016;11:e0153749.
24. Dong T, Sunderland N, Nightingale A, Fudulu DP, Chan J, Zhai B, Freitas A, Caputo M, Dimagli A, Mires S, Wyatt M, Benedetto U, Angelini GD. Development and Evaluation of a Natural Language Processing System for Curating a Trans-Thoracic Echocardiogram (TTE) Database. *Bioengineering*. 2023;10:1307.
25. Zheng C, Sun BC, Wu Y-L, Ferencik M, Lee M-S, Redberg RF, Kawatkar AA, Musigdilok VV, Sharp AL. Automated interpretation of stress echocardiography reports using natural language processing. *Eur Hear J - Digit Heal*. 2022;3:626–637.
26. Tang L, Kooragayalu S, Wang Y, Ding Y, Durrett G, Rousseau JF, Peng Y. EchoGen: Generating Conclusions from Echocardiogram Notes. *Proc 21st Work Biomed Lang Process*. 2022;2022:359–368.
27. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux M-A, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T. 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 C, Habib N, Sarrazin N, Sanseviero O, Rush AM, Wolf T. Zephyr: Direct Distillation of LM Alignment. *arXiv*. 2023;
29. Gardin JM, Adams DB, Douglas PS, Feigenbaum H, Forst DH, Fraser AG, Grayburn PA, Katz AS, Keller AM, Kerber RE, Khandheria BK, Klein AL, Lang RM, Pierard LA, Quinones MA, Schnittger I, Echocardiography AS of. Recommendations for a standardized report for adult transthoracic echocardiography: A report from the American Society of Echocardiography's Nomenclature and Standards Committee and Task Force for a Standardized Echocardiography Report. *J Am Soc Echocardiogr*. 2002;15:275–290.
30. Loper E, Bird S. NLTK: the Natural Language Toolkit. *Proc ACL-02 Work Eff tools Methodol Teach*

Nat Lang Process Comput linguistics -. 2002;63–70.
31. Han T, Adams LC, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, Truhn D, Bressem KK. MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data. *arXiv*. 2023;
32. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv*. 2019;
33. Cowan BR, Clark L, Følstad A, Skjuve M. Chatbots for customer service: user experience and motivation. *Proc 1st Int Conf Conversational User Interfaces*. 2019;1–9.
34. Li M, Gong S, Feng J, Xu Y, Zhang J, Wu Z, Kong L. In-Context Learning with Many Demonstration Examples. *arXiv*. 2023;
35. Choi E, Jo Y, Jang J, Seo M. Prompt Injection: Parameterization of Fixed Inputs. *arXiv*. 2022;
36. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv*. 2020;
37. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv*. 2023;
38. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv*. 2021;
39. 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.
40. Lavie A, Agarwal A. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. 2007;228–231.
41. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. *arXiv*. 2019;
42. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries [Internet]. Association for Computational Linguistics; p. 74–81.Available from: https://aclanthology.org/W04-1013
43. Delbrouck J-B, Chambon P, Bluethgen C, Tsai E, Almusa O, Langlotz CP. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. *arXiv*. 2022;
44. 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;
45. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22:276–82.
46. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–74.
47. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616:259–265.
48. Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models. *arXiv*. 2023;
49. Mallio CA, Sertorio AC, Bernetti C, Zobel BB. Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. *Radiol Med*. 2023;1–5.
50. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv*. 2020;

Figure Legends

Figure 1. The current workflow of summarizing echo Findings into clinically relevant Final Impressions.

Figure 2. Overview of the EchoGPT study.

Figure 3. ICL performance of each LLM. Panel A to E correspond to BLEU, METEOR, ROUGE-L, BERT Score, and RadGraph F1 Score, respectively. Zero-shot and fine-tuned Llama-2 (EchoGPT; horizontal purple dashed line) performance was included for reference.

Figure 4. Model win rate heatmap illustrating the head-to-head win rate comparisons (in percentile) among different models based on the selected metrics. Cool colors indicate lower win rates and warmed colors indicate higher win rates. We compared Llama-2 (base and fine-tuned), Zephyr (base and fine-tuned), T5 (base), and MedAlpaca (base). Fine-tuned Llama-2 consistently outperformed all other models across all 5 automatic metrics.

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 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 completeness (Pearson's $r= 0.78$), and modest to moderate correlations between other metrics. **Panel C.** Correlations between automatic metrics and the 4C metrics. Across all automatic metrics, RadGraph F1 had modest to moderate correlations with all 4 human evaluation metrics.

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.

Tables

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 external validation sets.

	Number of Samples	**Average Number of Token (Findings)**	**Average Number of Token (Final Impression)**	**Average Lexical Variance (Findings)**	**Average Lexical Variance (Final Impression)**
Development set	97.5k	215.7 ± 55.8	88.2 ± 34.3	0.51	0.65
External validation set	19.6k	211.8 ± 80.5	99.8 ± 39.0	0.52	0.64

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

*Compared zero-shot to QLoRA performance of the same model; **Compared performance of QLoRA Llama-2 and Zephyr.

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

Model: Llama-2	**Original Response**	**Synthetic Response: Replaced Measurements with Random Numbers**	**QLoRA**	**Random Measurements**	**P-value**
BLEU	Grade 1/4 left ventricular diastolic dysfunction, consistent with low to normal left ventricular filling pressure. estimated right ventricular systolic pressure 24 (systolic blood pressure 108). calculated ejection fraction; 64%. no regional wall motion abnormalities. mildly enlarged mid ascending aorta diameter of 40.	Grade 9/53 left ventricular diastolic dysfunction, consistent with low to normal left ventricular filling pressure. estimated right ventricular systolic pressure 68 (systolic blood pressure 52). normal left ventricular chamber size. calculated ejection fraction; 98%. no regional wall motion abnormalities. mildly enlarged mid ascending aorta diameter of 12.	45.9 ± 18.9	40.4 ± 16.4	<0.0001
METEOR			62.4 ± 18.0	58.7 ± 16.8	<0.0001
ROUGE-L			55.7 ± 17.8	52.2 ± 16.5	<0.0001
BERT Score			91.6 ± 3.0	91.2 ± 2.8	<0.0001
RadGraph F1			47.7 ± 14.9	45.3 ± 13.7	<0.0001

Table 6. Agreement of the rating between echo-expert readers.

Metric	**Fleiss' Kappa between 4 raters**
Completeness	0.43
Conciseness	0.32
Correctness	0.34
Clinical Utility	0.61

Supplemental Tables

Supplemental Table 1. Hyperparameter Search Results

	Temperature (RP fixed at 1.1)			Repetition Penalty (Temp fixed at 0.1)	
	0.1	0.5	0.9	1.2	1.3
LLaMA-2					
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 ± 5.7
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.

Supplemental Table 2. Data Distribution

	Development Set		External Validation Set	
Category	**Count**	**Percentage**	**Count**	**Percentage**
Adult TTE	79902	81.9	14967	76.5
Adult TEE	7829	8.0	1247	6.4
Exercise Stress	6781	7.0	2509	12.8
Pharmacological Stress	2994	3.1	569	2.9
Carotid Ultrasound	0	0.0	262	1.3
Intracardiac echocardiography	0	0.0	3	0.0
Total	97506	100	19557	100
TTE: transthoracic echocardiography, TEE: transesophageal echocardiography				

Supplemental Table 3. Representative Example of Zero-shot Results

Model	Ground Truth	LLaMA-2	Flan-T5	Zephyr	MedAlpaca
Representative 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 evaluation and management are warranted.	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.
Comment	--	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.

Supplemental Table 4. Representative Example: In-context Learning Integrated Information from Multiple Examples

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

Supplemental Table 5. Performance of Fine-tuned LLMs on the External Validation Set

Model	LLaMA-2	Zephyr	
	QLoRA	QLoRA	P-value*
BLEU	31.0 ± 18.6	17.5 ± 13.6	<0.0001
METEOR	45.8 ± 19.1	32.9 ± 14.6	<0.0001
ROUGE-L	42.0 ± 16.8	29.0 ± 13.8	<0.0001
BERT Score	88.9 ± 3.2	86.6 ± 3.9	<0.0001
RadGraph F1	36.9 ± 13.8	25.4 ± 12.3	<0.0001

*Compared performance of QLoRA Llama-2 and Zephyr.

Findings

- Calculated 2-d linear left ventricular ejection fraction 25. global hypokinesis with superimposed regional variation (please see wall motion graphics).
- No intracardiac mass or thrombus, but the left atrial appendage cannot be visualized adequately with transthoracic echo to exclude thrombus in this location.
- Estimated right ventricular systolic pressure 32 (systolic blood pressure 120).
- Prominent anterior epicardial fat layer. pulmonary valve not well visualized.
- Normal abdominal aorta doppler flow pattern.
- Normal left ventricular wall thickness.
- No atrial level shunt by color flow imaging.

⋮

- Abdominal aorta incompletely visualized.
- Left ventricular strain assessment not performed because of image quality.
- Left atrial volume index 28.
- No aortic valve regurgitation. Mild mitral valve regurgitation. Trivial pulmonary valve regurgitation. Trivial tricuspid valve regurgitation.

Summarization

Final Impressions

- Moderate-severely enlarged left ventricular chamber size ; calculated ejection fraction 25%.
- Global hypokinesis with superimposed regional variation (please see wall motion graphics).
- Normal right ventricular chamber size with borderline decrease in systolic function.
- Estimated right ventricular systolic pressure 32 (systolic blood pressure 120).
- No pericardial effusion.
- No significant valvular heart disease.
- Inferior vena cava size normal with normal inspiratory collapse (>50%).

Figure 1

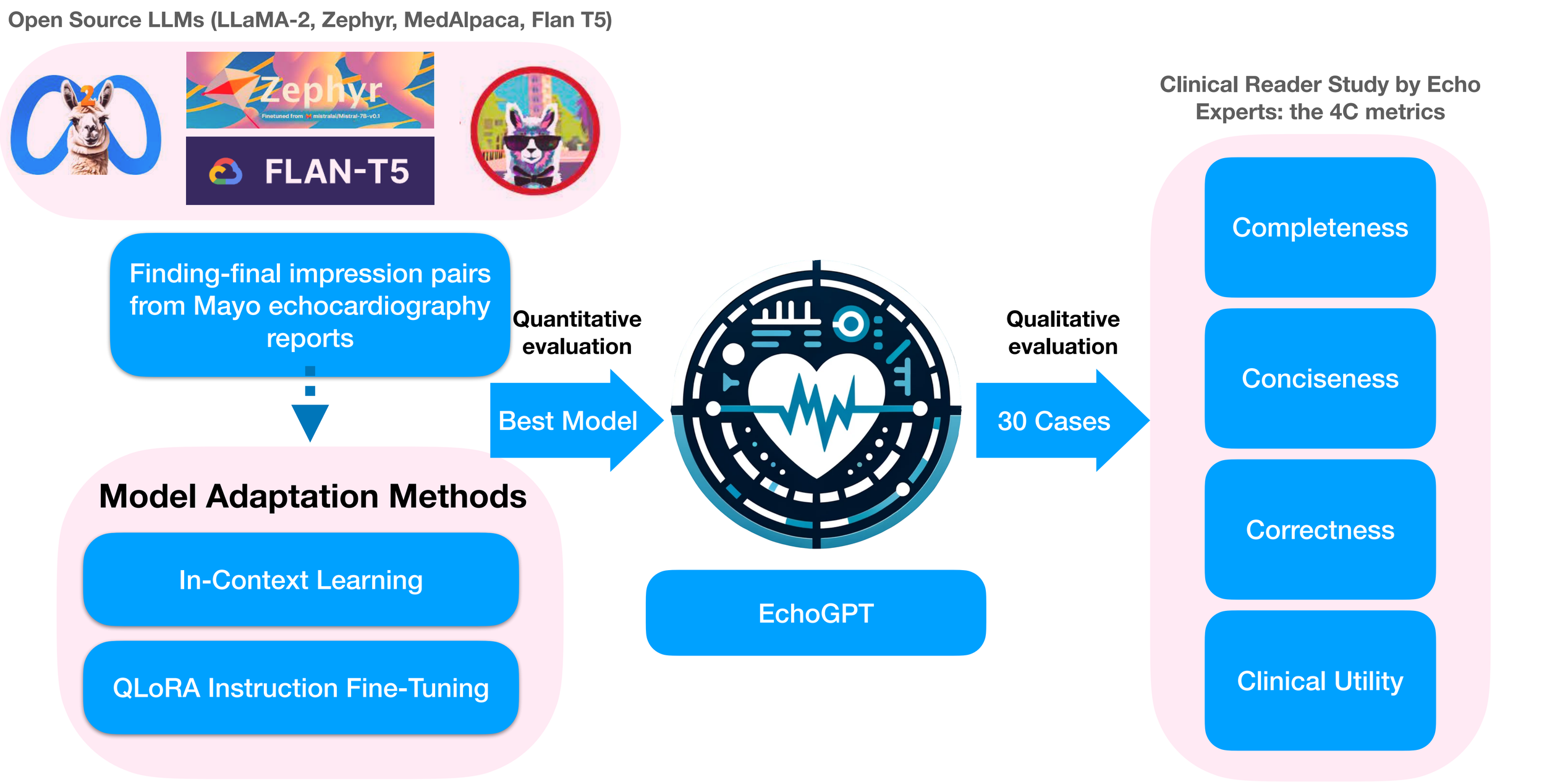

Open Source LLMs (LLaMA-2, Zephyr, MedAlpaca, Flan T5)
Zephyr
FLAN-T5
Finding-final impression pairs from Mayo echocardiography reports
Model Adaptation Methods
In-Context Learning
QLoRA Instruction Fine-Tuning
Quantitative evaluation
Best Model
EchoGPT
Qualitative evaluation
30 Cases
Clinical Reader Study by Echo Experts: the 4C metrics
Completeness
Conciseness
Correctness
Clinical Utility

Figure 2

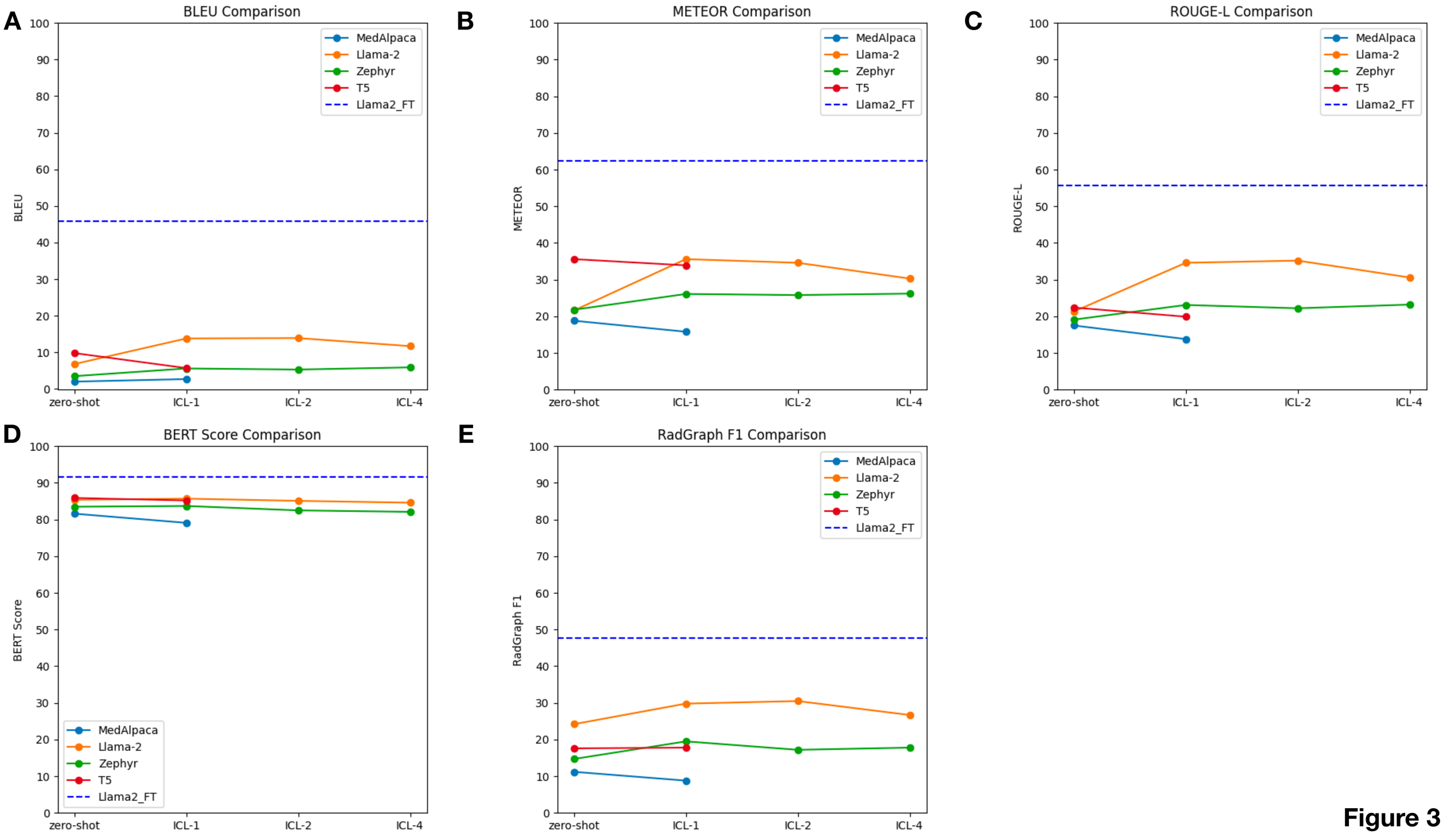

A
BLEU Comparison
BLEU
B
METEOR Comparison
METEOR
C
ROUGE-L Comparison
ROUGE-L
D
BERT Score Comparison
BERT Score
E
RadGraph F1 Comparison
RadGraph F1
MedAlpaca
Llama-2
Zephyr
T5
Llama2_FT
zero-shot
ICL-1
ICL-2
ICL-4

Figure 3

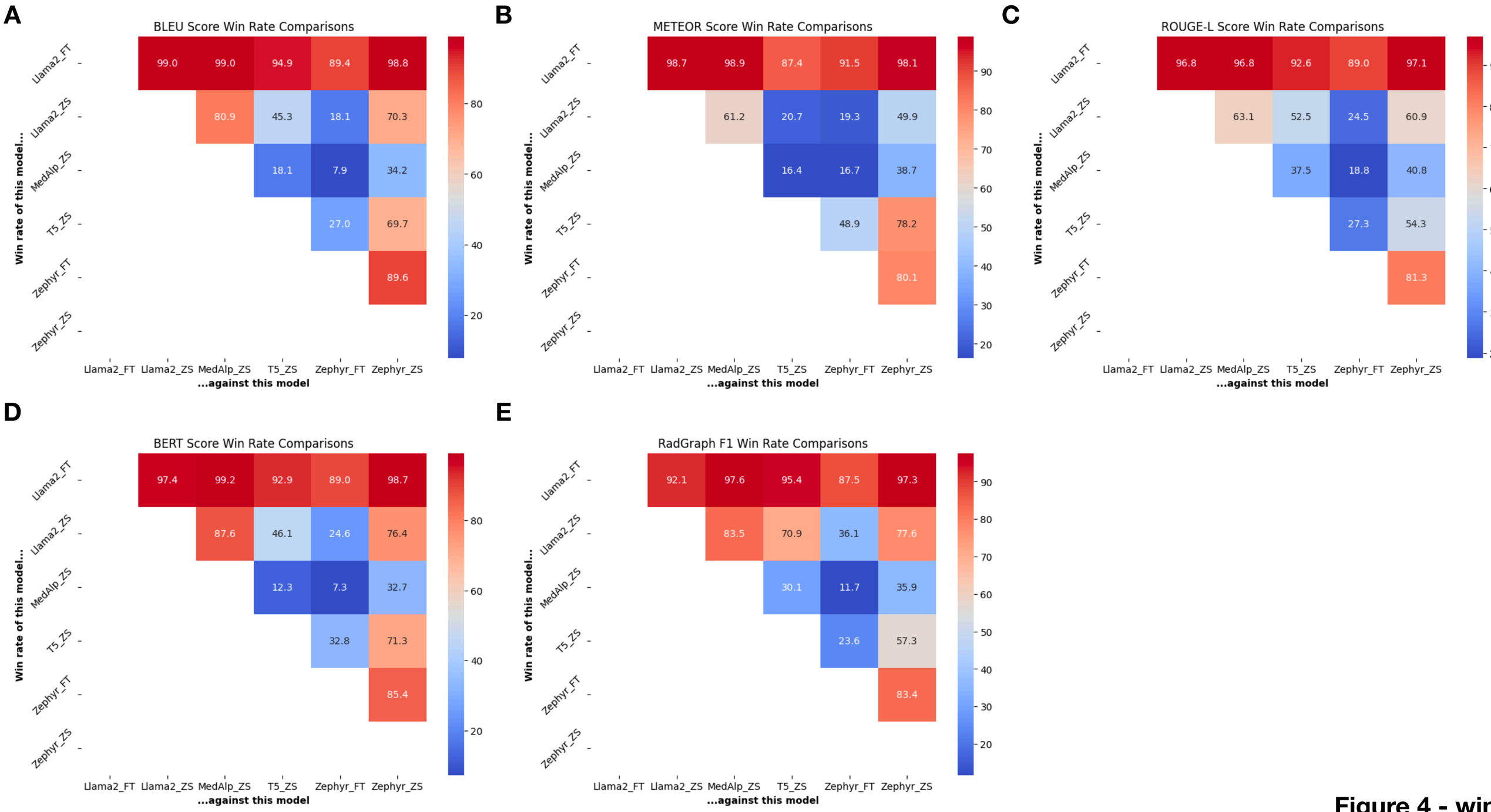
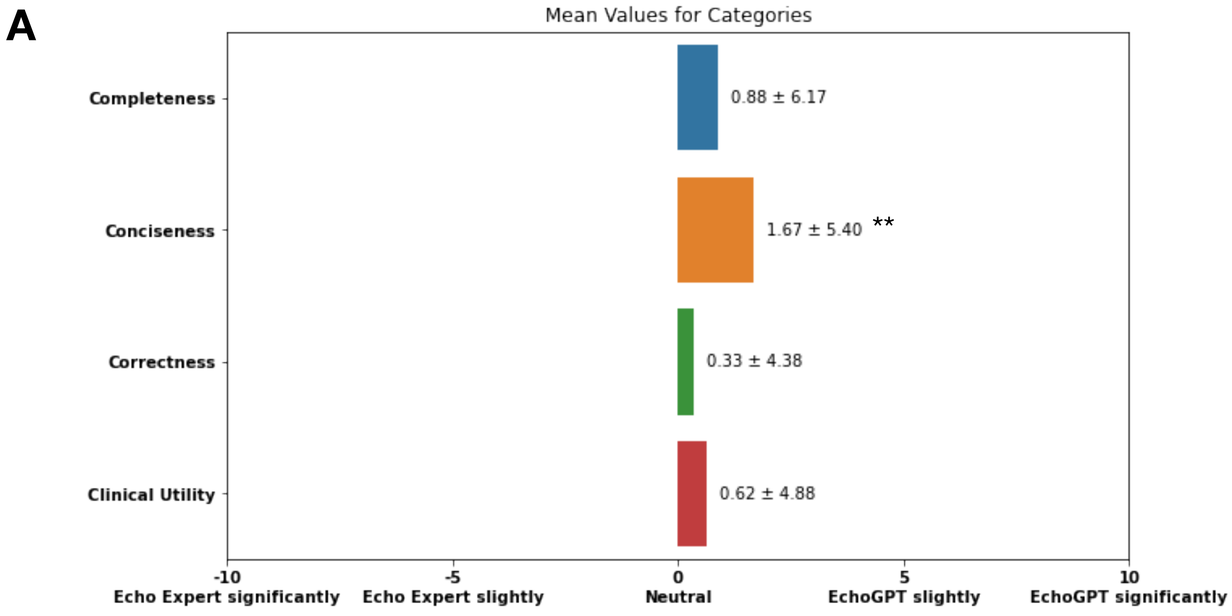
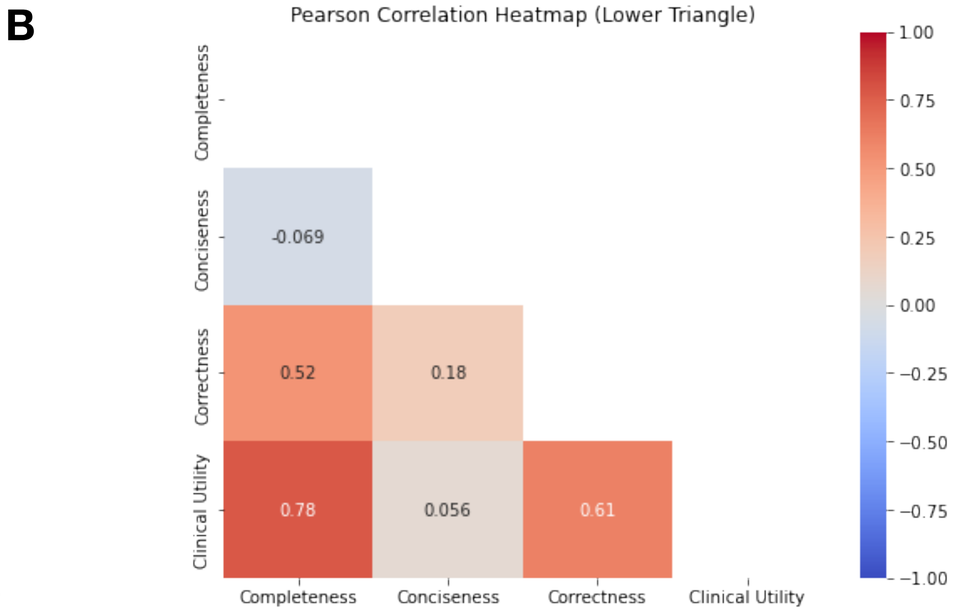

A
BLEU Score Win Rate Comparisons
Llama2_FT 99.0 99.0 94.9 89.4 98.8
Llama2_ZS 80.9 45.3 18.1 70.3
MedAlp_ZS 18.1 7.9 34.2
T5_ZS 27.0 69.7
Zephyr_FT 89.6
Zephyr_ZS
Llama2_FT Llama2_ZS MedAlp_ZS T5_ZS Zephyr_FT Zephyr_ZS
...against this model
Win rate of this model...
B
METEOR Score Win Rate Comparisons
Llama2_FT 98.7 98.9 87.4 91.5 98.1
Llama2_ZS 61.2 20.7 19.3 49.9
MedAlp_ZS 16.4 16.7 38.7
T5_ZS 48.9 78.2
Zephyr_FT 80.1
Zephyr_ZS
Llama2_FT Llama2_ZS MedAlp_ZS T5_ZS Zephyr_FT Zephyr_ZS
...against this model
Win rate of this model...
C
ROUGE-L Score Win Rate Comparisons
Llama2_FT 96.8 96.8 92.6 89.0 97.1
Llama2_ZS 63.1 52.5 24.5 60.9
MedAlp_ZS 37.5 18.8 40.8
T5_ZS 27.3 54.3
Zephyr_FT 81.3
Zephyr_ZS
Llama2_FT Llama2_ZS MedAlp_ZS T5_ZS Zephyr_FT Zephyr_ZS
...against this model
Win rate of this model...
D
BERT Score Win Rate Comparisons
Llama2_FT 97.4 99.2 92.9 89.0 98.7
Llama2_ZS 87.6 46.1 24.6 76.4
MedAlp_ZS 12.3 7.3 32.7
T5_ZS 32.8 71.3
Zephyr_FT 85.4
Zephyr_ZS
Llama2_FT Llama2_ZS MedAlp_ZS T5_ZS Zephyr_FT Zephyr_ZS
...against this model
Win rate of this model...
E
RadGraph F1 Win Rate Comparisons
Llama2_FT 92.1 97.6 95.4 87.5 97.3
Llama2_ZS 83.5 70.9 36.1 77.6
MedAlp_ZS 30.1 11.7 35.9
T5_ZS 23.6 57.3
Zephyr_FT 83.4
Zephyr_ZS
Llama2_FT Llama2_ZS MedAlp_ZS T5_ZS Zephyr_FT Zephyr_ZS
...against this model
Win rate of this model...

Figure 4 - win





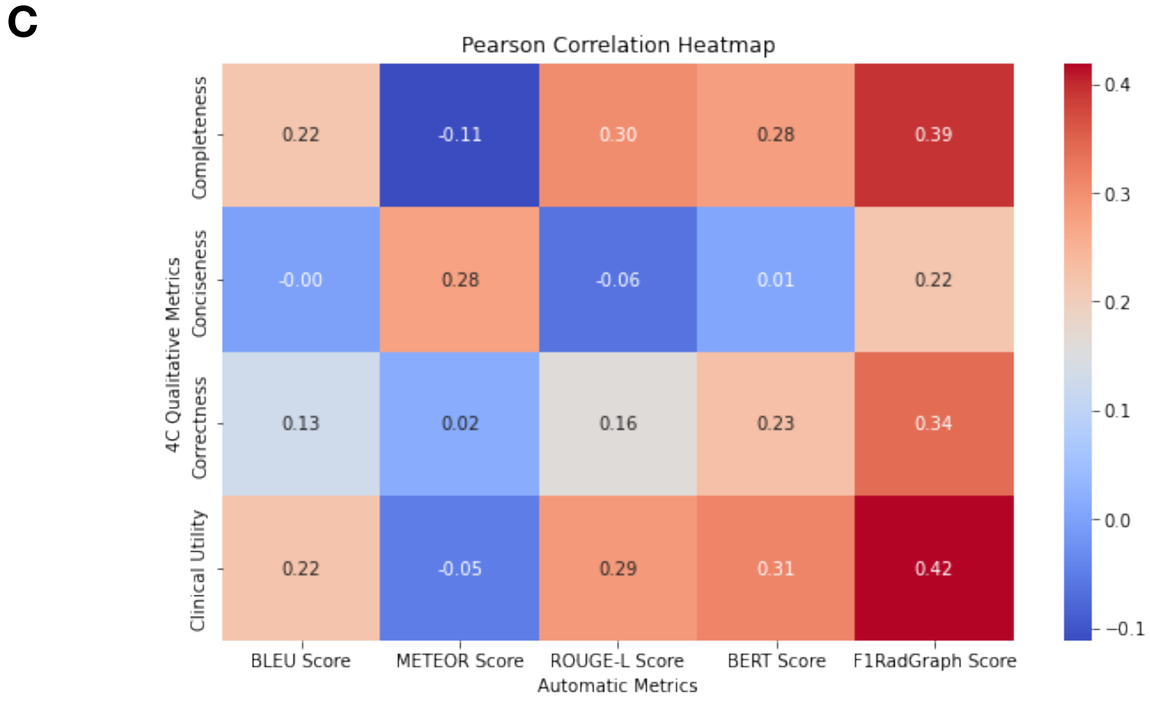


Figure 5 - Human eval

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 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%).

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%). *

	Summary A	Summary A slightly	Neutral	Summary B slightly	Summary B
Completeness	○	○	○	◉	○
Conciseness	○	○	◉	○	○
Correctness	○	◉	○	○	○
Clinical Utility	○	○	◉	○	○

Supplemental Figure 1