

1 **Title: Evaluating Large Language Models in Echocardiography Reporting: Opportunities and**
2 **Challenges**

3
4 **Authors:** Chieh-Ju Chao, MD^{1,2†}, Imon Banerjee, PhD³, Reza Arsanjani, MD⁴, Chadi Ayoub, MD, PhD⁴,
5 Andrew Tseng, MD⁵, Jean-Benoit Delbrouck, PhD⁶, Garvan C. Kane, MD, PhD¹, Francisco Lopez-
6 Jimenez, MD, MS¹, Zachi Attia, PhD¹, Jae K Oh, MD¹, Bradley Erickson, MD, PhD⁷, Li Fei-Fei, PhD²,
7 Ehsan Adeli, PhD^{2†} and Curtis Langlotz, MD, PhD⁶

8
9 †Chieh-Ju Chao and Ehsan Adeli are co-corresponding authors for this manuscript.

10
11 **Affiliations:**

12 ¹ Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota

13 ² Stanford Institute for Human-Centered Artificial Intelligence, Palo Alto, California

14 ³ Department of Radiology, Mayo Clinic, Scottsdale, Arizona

15 ⁴ Department of Cardiovascular Diseases, Mayo Clinic Arizona, Scottsdale, Arizona

16 ⁵ Department of Cardiovascular Medicine, Mayo Clinic Florida, Jacksonville, Florida

17 ⁶ Center for Artificial Intelligence in Medicine and Imaging (AIMI), Stanford University, Palo Alto,
18 California

19 ⁷ Department of Radiology, Mayo Clinic, Rochester Minnesota

20
21 **Disclosure:** The authors have no conflict of interest to disclose.

22
23 **Address of Correspondence:**

24 200 1st Street SW, Room: Gonda 4-478

25 Rochester, MN 55095, USA

26 Chieh-Ju Chao, MD

27 Assistant Professor of Medicine, Department of Cardiovascular Medicine, Mayo Clinic

28 Email: chao.chiehju@mayo.edu

29 Twitter: @chiehjucho1

30
31
32 **Keywords:** Artificial Intelligence, Large Language Model, Echocardiography, Quality, Cardiovascular
33 Imaging

34

35

36

37

38

39

40

41

42

43

44

1 **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

17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

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.

6

7 **Aims**

8 To evaluate open-source LLMs in echo report summarization.

9

10 **Methods**

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

18

19 **Results**

20 The development dataset included 97,506 reports from 71,717 unique patients, predominantly male
21 (55.4%), with an average age of 64.3 ± 15.8 years. EchoGPT, a QLoRA fine-tuned Llama-2 model,
22 outperformed other LLMs with win rates ranging from 87% to 99% in various automatic metrics, and
23 produced reports comparable to cardiologists in qualitative review (significantly preferred in conciseness
24 ($p < 0.001$), with no significant preference in completeness, correctness, and clinical utility). Correlations
25 between automatic and human metrics were fair to modest, with the best being RadGraph F1 scores
26 versus clinical utility ($r=0.42$) and automatic metrics showed insensitivity (0-5% drop) to changes in
27 measurement numbers.

28

29 **Conclusions**

30 EchoGPT can generate draft reports for human review and approval, helping to streamline the workflow.
31 However, scalable evaluation approaches dedicated to echo reports remains necessary.

32

33 **Clinical Perspectives**

34 1. What is new?

- 35 ● This study evaluated multiple open-source LLMs and different model adaptation methods
36 in echocardiography report summarization.
- 37 ● The resulting system, EchoGPT, can generate echo reports comparable in quality to
38 cardiologists.
- 39 ● Future metrics for echo report quality should emphasize factual correctness, especially on
40 numerical measurements.

41 2. What are the clinical implications?

- 42 ● EchoGPT system demonstrated the potential of introducing LLMs into echocardiography
43 practice to generate draft reports for human review and approval.

44

1 Introduction

2
3 Echocardiography (echo) is the mainstay imaging modality in the current practice of cardiology(1),
4 providing vital, non-invasive assessments of heart anatomy and physiology to guide clinical decisions(2).
5 In the past decade, the rising demand for diagnostic echo tests(3) has posed significant challenges in
6 maintaining the quality and timeliness of diagnostic reports(4–7), underscoring the necessity for automated
7 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
12 report text, and are technically equivalent to generating individual findings. However, these frameworks
13 were not designed to handle the high-level cognitive activity of synthesizing clinically relevant impressions
14 from detailed findings(12). In practice, physicians usually spend a significant amount of time summarizing
15 detailed findings to clinically relevant final impressions(13,14). While this task is crucial, it can be time-
16 consuming and prone to errors(15).

17
18 The advance of large language models (LLM) marked an important milestone for the application of AI in
19 healthcare to automate clinical information summarization(13,14,16) and expert-level question-
20 answering(17). A major advantage of LLMs is the flexibility of input and output(18), as well as the
21 capability of handling conversations and interaction with human experts(19). While similar functionality
22 can be achieved through commercially available LLMs (e.g., ChatGPT; OpenAI, San Francisco, CA)(20),
23 only a few healthcare institutions have integrated ChatGPT(21). Furthermore, fine-tuning ChatGPT for
24 specific tasks still requires uploading data to a central server, which also raises privacy concerns(22). In
25 contrast, open-source LLMs are free of charge and can be locally fine-tuned for specific tasks within each
26 healthcare institution's secure confines(18).

27
28 Previous studies predominantly focused on electronic health records(13,16) and chest X-rays (CXR)(13,18)
29 have highlighted the potential of using LLMs to summarize clinical text. In contrast, echo-related studies
30 were mainly on data extraction or classification, rather than report summarization(23–25). Tang et al. used
31 rule-based systems and a fine-tuned BART (Bidirectional and Auto-Regressive Transformer) model,
32 EchoGen (26) for this purpose and demonstrated convincing results. However, EchoGen reports were less
33 favored by human experts more than 50% of the time, perhaps due to the smaller number of parameters
34 than current state-of-the-art LLMs(26). Meanwhile, despite a recent study exploring the use of an LLM for
35 its question-answering capabilities on echo report texts(27), the potential of using billion-parameter LLMs
36 to generate echo reports remains under-explored(28,29).

37
38 In this work, we proposed to evaluate LLMs in echocardiography reporting and construct a local, domain-
39 specific LLM (EchoGPT) dedicated to echocardiography report summarization through an instruction fine-
40 tuning approach, which is known to be an effective strategy to adapt LLMs for similar tasks(13,18). We
41 anticipate that the fine-tuning procedure improves LLMs' performance on the task of echocardiography
42 report summarization. In addition to provide new insights into the use of open-source LLMs within the
43 domain of echocardiography reporting, we will explore the challenges associated with applying current
44 evaluation standards to echo reports generated by these models.

46 Method

47 *Dataset*

48 Mayo Clinic Reports: All adult (> 18 years old) echocardiography studies performed from 1/1/2017 to
49 12/31/2017 at Mayo Clinic Enterprise were retrieved. The types of studies include transthoracic
50 echocardiography (TTE), transesophageal echocardiography (TEE), and stress echocardiography

1 (including exercise and pharmacological studies). Text in the “Findings” and the “Final Impression”
2 sections of each report was used for the current study (**Figure 1**). The study was approved by the Mayo
3 Clinic IRB (protocol#: 22-010944).
4 MIMIC-III ECHO-NOTE2NUM Dataset (v.1.0.0, referred to as MIMIC-EchoNotes below)(30): This
5 publicly available dataset contains 43,472 valid free-text echocardiography reports from the intensive care
6 unit at the Beth Israel Deaconess Medical Center between 2001 and 2012. A random subset of the dataset
7 was used for external validation.

8 Data Curation and Preprocessing

10 The summarization task was defined as creating the Final Impression section based on the Findings
11 section, mirroring the established workflow of clinical echocardiography reporting at Mayo Clinic. The
12 Final Impression text was used as the ground truth report (**Figure 1**).
13 Mayo Clinic Reports: Echocardiography reports were excluded according to the following criteria: (1)
14 reports without Findings or Final Impression sections, (2) reports whose Findings or Impression section
15 contained less than 15 words, as these are frequently canceled studies in our practice, and (3) labeled in
16 report metadata as limited report. After this filtering process, the Findings section of each report was
17 further processed as follows: (1) Remove capitalized subheadings (e.g., LEFT VENTRICLE, VALVES,
18 OTHER FINDINGS, etc.), (2) Remove template sentences that make comparisons to prior reports, as no
19 information from previous reports has been provided in Findings, and (3) Remove quality control-related
20 sentences such as “study performed per left ventricular function protocol.”
21 MIMIC-EchoNotes Reports: Cases in this dataset were excluded based on criteria (1) and (2) above, as
22 the metadata differed from that of the Mayo Reports. We also removed the subheadings and template
23 sentences as described previously. We observed fundamental differences in report structure, including the
24 "General Comments" section, which typically contains comments related to study quality, and the
25 "Conclusions" section, which usually consists of the physician's interpretation of findings. However, the
26 Impression section often contains only 2-3 sentences summarizing the most pertinent study findings,
27 which was challenging for head-to-head comparison in this study. Given the distinct report structure, the
28 contents under the subheadings "General Comments" and "Conclusions" were integrated into the
29 "Findings" and "Impression" sections, respectively. Common abbreviations in the text were expanded to
30 their full forms.

31 Data split

33 Data from Rochester, Florida, and Mayo Clinic Healthcare sites were used as the model development set.
34 Considering variations in practice style among different sites, the data from the Arizona site was
35 designated as the external validation set (referred to as the AZ validation set). Within the development set,
36 1,000 non-duplicated cases were randomly selected for the test and validation sets, respectively; the rest
37 of the cases were used for fine-tuning (training set). Similarly, from the AZ validation and the MIMIC-
38 EchoNotes datasets, we selected 1,000 non-duplicated random cases from each. For basic dataset
39 statistics, the token length was calculated based on the natural language processing toolkit (NLTK)
40 tokenizer(31), and the lexical variance was defined as the ratio of the number of unique tokens to the
41 number of total tokens in each example(13).

42 Model Selection

1 Due to patient privacy policy regulations, proprietary LLMs such as GPT-3.5 and GPT-4 were not
2 considered in this work because versions of those models that were safe for protected health information
3 were not yet available. For our target task, we selected auto-regressive and sequence-to-sequence LLMs
4 with architectures under 7 billion parameters, balancing performance with manageability. Among open-
5 source models, we selected representative auto-regressive models including Llama-2-7b-chat(28),
6 Zephyr-7b(29), and Med-Alpaca(32) models considering their performance and max input context length
7 on general natural language processing (NLP) tasks and radiology report summarization(13). For
8 sequence-to-sequence (seq2seq) models, we used Flan-T5 (base) as the representative model as it is
9 known for accurate text summarization(13,33), given that the EchoGen model(26) was not publicly
10 available.

11

12 Model Inference Hyperparameter Search

13 LLM inference was conducted by using Hugging Face’s (Manhattan, NY) transformer pipeline via the
14 open-source LangChain framework(34). After initial tests, text generation and summarization were used
15 as the task type for auto-regressive models and seq2seq models, respectively. A subset (10%, n=100) of
16 examples were randomly selected from the test set for the hyperparameter search. We specifically tested
17 the following configuration parameters that can significantly affect performance: temperature (0.1, 0.5,
18 and 0.9) and repetition penalty (1.1, 1.2, and 1.3). These two parameters were tested separately, when one
19 parameter was being tested, the other was fixed at the lowest value. The generated contents were
20 evaluated by both automatic metrics and qualitative assessment. We chose the following configuration for
21 model inference: {temperature 0.1, repetition penalty 1.1} after comparing automatic metrics and
22 qualitative assessments; see **Supplemental Table 1**). We did not complete a dedicated search procedure
23 for the optimal LLM inference configurations, but the configurations used in our study were similar to
24 prior reports, and the generated contents were satisfying on qualitative review. Of note, the configurations
25 were tested in a zero-shot setting, and the best configuration was directly applied to the ICL and QLoRA
26 fine-tuned models(13).

27

28 Model Adaptation

29 *Prompt:* A prompt template was created with components of the prefix, instruction, and suffix(13) (**Table**
30 **1**). The final prompt was decided after qualitatively evaluating several different variants of each
31 component on a small subset of the data. We also specified that the summarization should be “concise”
32 and use “a minimal amount of text” to avoid LLMs generating lengthy reports(13). Likely due to the
33 difference in the reporting style of the MIMIC-EchoNotes dataset, the final prompt above led to
34 suboptimal responses. Therefore, we adopted a new prompt tailored to match the reporting style by
35 incorporating new instructions below: 1) Write a 10-bullet points clinical summary, and 2) Avoid using
36 numbers other than LVEF.

37

38 *In-context Learning (ICL):* ICL has been proposed to improve LLM’s performance without changing the
39 base model weights(13,35,36). Also, using relevant in-context examples is shown to have better model
40 performance compared to random examples in ICL. To obtain relevant in-context examples, we adopted
41 the approach to select m (m=1, 2, 4...) nearest neighbors from the training set for each test set case, after
42 embedding both sets by the PubMedBERT model(37).

43

1 *Instruction tuning with quantized low-rank adaptation (QLoRA)*: Due to the size of candidate models, we
2 opted for quantized low-rank adaptation (QLoRA)(38), a type of parameter efficient fine-tuning
3 (PEFT)(39) to optimize our LLMs for echo report summarization tasks. The same prompt template
4 (**Table 1**) was used, and the Final Impression text from the same report was used as the target output
5 (13,14).
6

7 We configured the training process as follows: load model in 4-bit precision, with a LoRA configuration
8 of (alpha = 16, LoRA dropout = 0.1, LoRA r= 64). The batch size and gradient accumulation were
9 adjusted for each model to achieve an effective batch size of 24 that fits on a single NVIDIA RTX A5000
10 24G GPU setting. A paged-AdamW 32-bit optimizer was used, with an initial learning rate of 1e-3, which
11 decayed to 1e-4 (by a cosine scheduler) after the initial 100 warm-up steps. The above configuration
12 provided the most stable training process after attempting different configurations reported in prior
13 studies(13,38).
14

15 *Model Performance Evaluation*

16 *Automatic NLP evaluation metrics*: To evaluate the models' performance on the information
17 summarization task and compare it to prior works, we utilized four established automatic metrics that
18 have been used in other clinical text summarization studies(40,41): BLEU (Bilingual Evaluation
19 Understudy)(42), METEOR (Metric for Evaluation of Translation with Explicit ORdering)(43), ROUGE-
20 L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence), and the BERT
21 (Bidirectional Encoder Representations from Transformers) score(44), which represents the similarity
22 between generated contents and the corresponding reference at words/characters (n-gram), single word
23 (unigram), longest sequence of words, and contextual level, respectively For ROUGE-L, we present the
24 F1 score component(26,45). For factual correctness, the RadGraph-F1 metric (level: all) was
25 reported(46,47) This metric served as the primary evaluation criterion for model performance,
26 considering its significance in ensuring the factual correctness of generated clinical content.
27

28 *Evaluation of Significance of Measurement Numbers in Automatic Metrics*: Considering the importance
29 of measurements in echo studies, we also attempted to evaluate whether the current automatic metrics can
30 detect changes in measurement numbers. For this purpose, we generated synthetic reports by replacing all
31 the measurement numbers with random numbers ranging from 1 to 99 in the reports. We then compared
32 the automatic metric scores with the corresponding reports with the original measurements.
33

34 *Human expert evaluation metrics*: We designed a human expert evaluation process based on previous
35 clinical text summarization studies(13,18,26). The Findings with corresponding ground truth (Final
36 Impression) and LLM-generated summarization of 30 randomly selected cases were presented to four
37 echocardiography-board-certified cardiologists for blinded quality review. We noted that physician-
38 summarized Impressions may contain free-text information beyond the Findings (e.g., documenting
39 events during the study or communication with the ordering provider), and the reviewers were instructed
40 to rate only based on information within the Findings section. Each metric was rated for the preference (5
41 levels) between the two summarizations (**Supplemental Figure 1**)(13). The “4C metrics” evaluated by
42 echocardiography experts are completeness, conciseness, correctness, and clinical utility(14), as described
43 in **Table 2**.
44

1 *Statistical analysis:* Automatic metric performance between each pair of models was compared using a
2 two-tailed paired Student's t-test or Wilcoxon signed-rank test for normally and non-normally distributed
3 data, respectively. Models were also compared based on win rates, which are defined as the percentage of
4 head-to-head victories in performance between two models for each selected metric(13). For the
5 performance bias analysis, data were grouped based on sex (male versus female) and race (white, black,
6 other). In the case of sex, we employed two-tailed Student's t-tests or Mann-Whitney U tests for normally
7 and non-normally distributed data, respectively. For race, we utilized one-way ANOVA. In human expert
8 qualitative analysis, the 4C metrics were compared by a one-sample Wilcoxon signed-rank test(13), and
9 the agreement of ratings between experts was assessed by Fleiss' kappa coefficient(48), and interpreted as
10 recommended by Landis et al(49). We conducted Pearson correlation analyses to explore the relationships
11 between each human expert evaluation metric, assessing their independence. Additionally, we conducted
12 similar analyses to examine the correlation between human and automatic metrics. Statistical analyses
13 were performed using Python 3.8 and SciPy 1.8.0. All the comparisons consider a p-value < 0.05 as
14 significant.

15 16 **Results**

17 Patient Cohort and Dataset

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

26 27 Zero-shot, ICL, and QLoRA fine-tuned performance

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

36
37 In ICL, LLMs that allow longer context length (Llama-2 and Zephyr) had the best performance across all
38 metrics when one example was provided (ICL-1). The performance gradually trended down with more
39 examples (ICL-2 and ICL-4). In contrast, the performance of LLMs with shorter max context length
40 started to trend down with one example (**Figure 3**). Importantly, we observed that Llama-2 can integrate
41 information from multiple ICL examples, such as "Calculated 2-d linear left ventricular ejection fraction
42 57, 61, 75 (3 reports)", which occurred in 34 (3.4%) and 69 (6.9%) cases for ICL-2 and ICL-4,
43 respectively (**Supplemental Table 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**
10 **Figure 2**). Because Llama-2 had the best performance on zero-shot, ICL, and QLoRA fine-tuning
11 approaches, fine-tuned Llama-2 was selected as EchoGPT and used for the subsequent expert qualitative
12 review.

13 Significance of Measurement Numbers in Automatic Metrics

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

18 Human Expert Evaluation

19 We observed slight agreement for correctness, fair agreement for conciseness and clinical utility, and
20 moderate agreement for completeness (**Supplemental Table 5**). Among the 4C metrics, we observed that
21 EchoGPT significantly outperformed human experts in conciseness ($p<0.001$). There was no significant
22 preference among the other three categories (**Figure 5A**). The 4C metrics were not completely
23 independent. There was a high correlation between clinical utility and completeness (Pearson's $r=0.78$),
24 and modest to moderate correlations between other metrics (**Figure 5C**). We also observed that across all
25 automatic metrics, RadGraph F1 had modest to moderate correlations with all 4 human evaluation metrics
26 (**Figure 5C**).

27 External Validation on the MIMIC-EchoNotes dataset

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

37 **Discussion**

38 This study evaluated the feasibility of multiple open-source LLMs in echocardiography report
39 summarization through different model adaptation methods. Trained on one of the largest
40
41
42
43
44

1 echocardiography report datasets in the world, we demonstrated that QLoRA fine-tuning can significantly
2 improve LLMs' performance for the desired summarization task, with at least comparable qualities to
3 human experts. ICL was overall inferior to QLoRA fine-tuning and faced substantial limitations, such as
4 integrating information from multiple examples. Additionally, we demonstrated that current automatic
5 metrics are not sensitive to the change in measurement numbers in echo reports. The current study offers
6 insights into the development and evaluation of a specialized local LLM tailored for echo report
7 summarization and presenting significant workflow advantages.

8 Model Adaption Approaches: ICL vs. Fine-tuning

9 Our results suggested that both ICL and QLoRA fine-tuning improved LLMs' performance over zero-
10 shot, and fine-tuning performance was consistently above ICL across all automatic metrics (**Figure 3**).
11 Additionally, we observed substantial limitations of ICL for echo reporting, particularly due to the longer
12 context lengths typical of echo reports. The behavior of integrating results from multiple examples also
13 presents challenges, as discussed below.

14
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 examples(13), 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
21 studies(13,35). It is also important to consider that the computation time and resources required in ICL
22 can increase with the number of examples used(35). Additionally, we note that LLMs can integrate
23 information from the example ICL cases, which compromises the report quality (**Supplemental Table 3**).
24 While this behavior was not reported in other studies(13,35), we believe it could be a relatively common
25 condition, that is easier to identify with numerical values (in echo) compared to narrative statements (in
26 CXR). Therefore, even in scenarios where ICL can outperform fine-tuning(13), fine-tuning may be
27 preferable.

28
29 The EchoGen study previously demonstrated that Bidirectional Auto-Regressive Transformers (BART)
30 was superior to other rule-based approaches for summarizing echocardiogram reports, with BART
31 achieving ROUGE-based scores between 0.65 and 0.73, however, human summaries were preferred by
32 the majority of the time over those generated by the BART model(26). Although EchoGPT didn't match
33 the scores in ROUGE-L, it compared favorably to human experts in qualitative assessments. Additionally,
34 in our study, we observed that T5 (as the representative seq2seq model since EchoGen is not publicly
35 available) generated summaries that were overly brief, so important clinical information was missed
36 (**Supplemental Table 2**). We assume that similar behavior could have occurred with BART, leading to
37 its unfavorable ratings by physicians. However, direct comparisons were not feasible as neither the
38 EchoGen model nor qualitative examples generated by it were available for review.

39 Evaluation of Echo Report Summarization

40 Automatic evaluation of LLM in clinical text summarization tasks is an emerging area, and there is no
41 gold standard metric that can evaluate all aspects of a report(13,26). Our study reinforces this conclusion.
42 We noted that MedAlpaca and Zephyr can generate medical-professionally-sounding content that
43

1 frequently includes hallucinated information. These differences were mainly reflected by the factual
2 correctness metric RadGraph F1 (**Table 4**).

3
4 The practice style at each institution could greatly affect the quantitative performance of a model(18). In
5 the AZ validation set, we observed a 5-10% drop in performance of both fine-tuned Llama-2 and Zephyr
6 (**Supplemental Table 4**). This is likely secondary to the differences in practice style: the AZ validation
7 set, despite having a similar Finding section length, contained an average of 9.5 additional tokens in the
8 Final Impression section (**Table 3**). A more significant drop in performance was observed in reports from
9 the MIMIC-EchoNotes datasets (RadGraph F1 from 47.7 to 25.1; **Supplemental Table 4**), which was
10 anticipated and within a reasonable range(50). According to our observations and the input of expert
11 reviewers, the key factors leading to the performance drop were:

- 12
- 13 1. The distinct report structure (Findings/Impressions versus Findings/Conclusions/Impressions).
- 14 2. The use of reporting languages (templated statements at Mayo versus the free-text style of the
15 MIMIC-EchoNotes dataset).
- 16

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

30
31 Regarding the correlations between automatic metrics and human expert preference, our results were
32 similar to the prior studies, showing that most of the metrics were not strongly correlated(13). Notably,
33 the highest correlation was observed between the RadGraph F1 scores and the 4C metrics, particularly in
34 terms of clinical utility ($r=0.42$) (**Figure 5C**). This suggests that the quality of echo reports judged by
35 cardiologists may not be well captured in automated metrics that do not capture notions of factual
36 correctness. While stronger correlations were observed in the MIMIC-EchoNotes examples
37 (**Supplemental Figure 3**), we believe it was reflecting the strong preference secondary to the distinct
38 reporting style.

39
40 As a specific subtype of clinical text, echocardiography reports contain unique terminology, including
41 precise measurements. Clinically, 25% and 55% LV ejection fraction values indicate a significant
42 difference, however, our study demonstrates that this distinction is difficult to capture with current
43 automatic metrics (**Table 5**). While this aspect of reporting can be easily captured in qualitative analyses,
44 such analyses are expensive to conduct at scale because of the limited availability of in-domain experts. A

1 dedicated metric for echocardiography diagnostic quality evaluation, with emphasis on measurement
2 accuracy, is still needed to address this knowledge gap.

3 Application of EchoGPT

4 Our study shows the feasibility of introducing LLMs into echocardiography practice. Through the
5 QLoRA fine-tuning process(38), the EchoGPT model was able to learn clinically relevant knowledge to
6 summarize echo report findings at a quality level comparable to echocardiography-trained cardiologists
7 (Figure 6A).

8
9
10 We envision that EchoGPT could be used as a reporting interface or a co-pilot that could generate echo
11 reports with various inputs(52). EchoGPT still inherits the limitations of LLMs, including
12 hallucination(13,18,53). Although the fine-tuning process can potentially reduce hallucinations, additional
13 efforts such as optimization for factual correctness(46) or paired with a retrieval augmented generation
14 system(54) are still required to minimize hallucinations before clinical implementation.

15 16 **Conclusion**

17 Our study successfully built EchoGPT through QLoRA fine-tuning of open-source LLMs and
18 demonstrated that the model is capable of generating echocardiography reports on par with cardiologists,
19 marking an advancement in integrating LLMs into current echo practice. Our analysis also highlights the
20 challenges of the current LLM evaluation process in echo reporting, particularly dedicated automatic
21 metrics and scalable qualitative evaluation approaches, which necessitate further studies.

22 23 **Limitations**

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

35 36 **Data Availability**

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

40 41 **Code Availability**

42 We released a checkpoint of the fine-tuned Llama-2 model, along with the QLoRA fine-tuning, inference,
43 and statistical analysis code. The code and checkpoint are available on GitHub:
44 <https://github.com/chiehjuchao/EchoGPT.git>

1 **Figure Legends**

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

3
4 **Figure 2.** Overview of the EchoGPT study.

5
6 **Figure 3.** ICL performance of each LLM. Panel A to E correspond to BLEU, METEOR, ROUGE-L,
7 BERT Score, and RadGraph F1 Score, respectively. Zero-shot and fine-tuned Llama-2 (EchoGPT;
8 horizontal purple dashed line) performance was included for reference.

9
10 **Figure 4.** Model win rates on the test set. Model win rate heatmap illustrates the head-to-head win rate
11 comparisons (in percentile) among different models based on the selected metrics. Cool colors indicate
12 lower win rates and warmed colors indicate higher win rates. We compared Llama-2 (base and fine-
13 tuned), Zephyr (base and fine-tuned), T5 (base), and MedAlpaca (base). Fine-tuned Llama-2 consistently
14 outperformed all other models across all 5 automatic metrics. FT: fine-tuned. ZS: zero-shot (base model).

15
16 **Figure 5.** Human expert qualitative evaluation results. **Panel A.** In the 4 categories, EchoGPT
17 significantly outperformed human experts in conciseness ($p < 0.001$). We didn't observe significant
18 differences among the other three categories (completeness, correctness, and clinical utility). **Panel B.**
19 showed interdependence of the 4C metrics, especially the correlations between clinical utility and
20 completeness (Pearson's $r = 0.78$), and modest to moderate correlations between other metrics. **Panel C.**
21 Correlations between automatic metrics and the 4C metrics. Across all automatic metrics, RadGraph F1
22 had modest to moderate correlations with all 4 human evaluation metrics. Preference ratings were
23 expressed as mean \pm standard deviation, **indicates $p < 0.001$.

24
25 **Central Illustration.** Open-source LLMs (Llama-2, MedAlpaca, Zephyr, and Flan-T5) were evaluated
26 using In-Context Learning (ICL) and Quantized Low-Rank Adaptation (QLoRA) fine-tuning for
27 summarizing echocardiography reports from "Findings" to "Impressions." EchoGPT, a QLoRA fine-
28 tuned Llama-2 model, surpassed other LLMs in multiple automatic metrics and produced reports on par
29 with those of cardiologists in qualitative reviews. However, challenges in evaluating generated reports
30 included limited scalability of expert reviews, modest correlations between automatic and human metrics,
31 and the automatic metrics' insensitivity to changes in measurements.

32 33 **Supplemental Materials**

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

37 **Supplemental Figure 2.** This bar chart displays model performance across five automatic metrics,
38 considering sex (male vs. female) and race (white, black, and other) variables. Panels A and B compare
39 metrics by sex and race variables in the test set, while Panels C and D perform the same comparisons in
40 the AZ validation set. There were no significant biases detected for sex or race, except for slightly better
41 RadGraph F1 performance in female patients within the AZ validation set (male vs. female: 0.38 ± 0.14

1 vs. 0.40 ± 0.15 , $p=0.04$). Scores were presented as mean values \pm standard error bars. Demographic
2 information was not available for the same analysis in the MIMIC-EchoNotes dataset.

3 **Supplemental Figure 3.** Human expert qualitative evaluation results on the MIMIC-EchoNotes dataset.
4 **Panel A.** In the 4 categories, the original reports (combined Conclusions and Impression sections) were
5 preferred over EchoGPT in completeness, correctness, and clinical utility ($p < 0.001$); while EchoGPT
6 was superior in conciseness ($p < 0.001$). **Panel B.** Interdependence of the 4C metrics, especially the
7 correlations between clinical utility, completeness, and correctness (Pearson's $r = 0.48$ and 0.66 ,
8 respectively). **Panel C.** Correlations between automatic metrics and the 4C metrics. Across all automatic
9 metrics, RadGraph F1 still had the strongest correlations with the 4C metrics, although other metrics also
10 had stronger correlations. Preference ratings were expressed as mean \pm standard deviation, **indicates
11 $p < 0.001$.

12

13 Tables

14

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.” |

15

16

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. |

17

1

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

| | Development Set | | | | 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 | | | | | | |
| <i>White</i> | 89466 (91.8%) | 87632 (91.8%) | 914 (91.4%) | 920 (92.0%) | 885 (88.5%) | -- |
| <i>Black</i> | 3326 (3.4%) | 3257 (3.4%) | 35 (3.5%) | 34 (3.4%) | 42 (4.2%) | -- |
| <i>Other</i> | 2850 (2.9%) | 2790 (2.9%) | 28 (2.8%) | 32 (3.2%) | 35 (3.5%) | -- |
| <i>Asian</i> | 1415 (1.5%) | 1390 (1.5%) | 16 (1.6%) | 9 (0.9%) | 22 (2.2%) | -- |
| <i>Native American</i> | 362 (0.4%) | 353 (0.4%) | 5 (0.5%) | 4 (0.4%) | 13 (1.3%) | -- |
| <i>Pacific Islander</i> | 87 (0.1%) | 84 (0.1%) | 2 (0.2%) | 1 (0.1%) | 3 (0.3%) | -- |
| Sex | | | | | | |
| <i>Male</i> | 54010 (55.4%) | 52920 (55.4%) | 563 (56.3%) | 527 (52.7%) | 584 (58.4%) | -- |
| <i>Female</i> | 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 | | | | | | |
| <i>Adult TTE</i> | 79902 (81.9%) | 78243 (81.9%) | 833 (83.3%) | 826 (82.6%) | 776 (77.6%) | 881 (88.1%) |
| <i>Adult TEE</i> | 7829 (8.0%) | 7685 (8.0%) | 65 (6.5%) | 79 (7.9%) | 63 (6.3%) | 108 (10.8%) |
| <i>Exercise Stress</i> | 6781 (7.0%) | 6652 (7.0%) | 64 (6.4%) | 65 (6.5%) | 131 (13.1%) | 11 (1.1%) |
| <i>Pharmacological Stress</i> | 2994 (3.1%) | 2926 (3.1%) | 38 (3.8%) | 30 (3.0%) | 30 (3.0%) | -- |
| Data Characteristics | | | | | | |
| <i>Average Number of Tokens (Findings)</i> | 215.7 ± 55.8 | 215.7 ± 55.9 | 213.1 ± 55.2 | 214.8 ± 53.3 | 217.1 ± 77.3 | 184.3 ± 50.0 |
| <i>Average Number of Tokens (Final Impression)</i> | 88.2 ± 34.3 | 88.2 ± 34.3 | 87.1 ± 33.7 | 86.8 ± 33.6 | 97.7 ± 37.9 | 163.2 ± 40.4 |
| <i>Average Lexical Variance (Findings)</i> | 0.51 | 0.51 | 0.52 | 0.51 | 0.52 | 0.54 |
| <i>Average Lexical Variance (Final Impression)</i> | 0.65 | 0.65 | 0.65 | 0.65 | 0.64 | 0.52 |
| CAD: coronary artery disease, HTN: hypertension, DM: diabetes, CKD: chronic kidney disease, CHF: congestive heart failure. TTE: transthoracic echocardiography, TEE: transesophageal echocardiography. | | | | | | |

2

3

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. 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
3

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 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 |

4
5

Table 6. Representative Examples in Reader Study

| Mayo Echo Report | Findings | Impression | EchoGPT Summary |
|---------------------|----------|------------|-----------------|
|---------------------|----------|------------|-----------------|

| | | | |
|--------------------------------------|---|---|--|
| <p>Representative Example</p> | <p>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. Trivial tricuspid 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 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 velocities. No abdominal aortic aneurysm. Normal left ventricular diastolic function. Calculated 2-d monoplane volumetric left ventricular ejection fraction 59.</p> | <p>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.</p> | <p>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%)</p> |
| <p>Comments</p> | <p>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.</p> | <p>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.</p> | <p>EchoGPT captured the relevant findings, although not in the same order. LV ejection fraction measurement was included in the summary (61%).</p> |
| <p>MIMIC-EchoNote Report</p> | <p>Findings</p> | <p>Conclusions/Impression</p> | <p>EchoGPT Summary</p> |

| | | | |
|--------------------------------------|--|---|--|
| <p>Representative Example</p> | <p>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.</p> | <p>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.</p> | <p>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.</p> |
| <p>Comments</p> | <p>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.</p> | <p>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..."</p> | <p>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.</p> |

1
2
3
4

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 \pm 5.4 | 8.8 \pm 6.7 | 8.0 \pm 6.1 | 5.5 \pm 4.6 | 1.1 \pm 1.8 |
| METEOR | 23.3 \pm 8.2 | 24.1 \pm 8.2 | 23.0 \pm 8.1 | 20.1 \pm 6.9 | 12.9 \pm 5.7 |
| ROUGE-L | 22.3 \pm 7.4 | 23.2 \pm 7.8 | 23.3 \pm 9.3 | 18.7 \pm 7.0 | 11.6 \pm 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 |
|-------------------------------|--|--|---|---|--|
| Representative Example | Global hypokinesia 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 hypokinesia with superimposed regional variation. No intracardiac mass or thrombus. | LVEF: 25% with superimposed regional variation (hypokinesia) - 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. |

| | | | | | |
|-----------------|----|---|---|--|--|
| 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. |
|-----------------|----|---|---|--|--|

1
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, 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 | |

4
5 **Supplemental Table 4. Performance of Fine-tuned LLMs on the AZ Validation and MIMIC-**
6 **EchoNote Datasets**

| Dataset | Mayo AZ validation | | | MIMIC-EchoNote | | |
|-------------|--------------------|-------------|----------|----------------|-------------|----------|
| | LLaMA-2 | Zephyr | | LLaMA-2 | Zephyr | |
| Model | 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

7
8
9
10
11
12
13
14
15
16
17 **Supplemental Table 5. Agreement of the Ratings Between Echo-Expert Readers.**

| Metric | Fleiss' Kappa between 4 raters | |
|------------------|--------------------------------|------------|
| | Mayo | MIMIC-Echo |
| Completeness | 0.49 | 0.49 |
| Conciseness | 0.22 | 0.60 |
| Correctness | 0.17 | 0.12 |
| Clinical Utility | 0.34 | 0.14 |

1 **References**

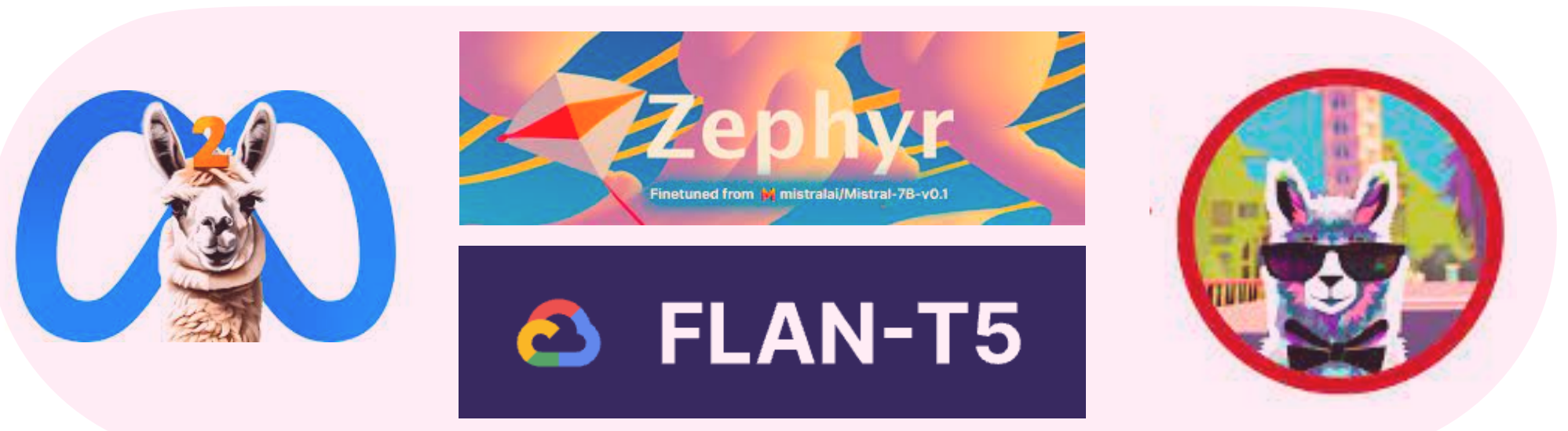
- 2 1. Daubert MA., Tailor T., James O., Shaw LJ., Douglas PS., Koweik L. Multimodality cardiac imaging
3 in the 21st century: evolution, advances and future opportunities for innovation. *Br J Radiol*
4 2021;94(1117):20200780. Doi: 10.1259/bjr.20200780.
- 5 2. Carli MFD., Geva T., Davidoff R. The Future of Cardiovascular Imaging. *Circulation*
6 2016;133(25):2640–61. Doi: 10.1161/circulationaha.116.023511.
- 7 3. Reeves RA., Halpern EJ., Rao VM. Cardiac Imaging Trends from 2010 to 2019 in the Medicare
8 Population. *Radiol Cardiothorac Imaging* 2021;3(5):e210156. Doi: 10.1148/ryct.2021210156.
- 9 4. Tiver KD., Horsfall M., Swan A., et al. Accuracy of Highly Limited Echocardiographic Screening
10 Images for Determining a Structurally Normal Heart: The Quick-Six Study. *Hear Lung Circ*
11 2022;31(4):462–8. Doi: 10.1016/j.hlc.2021.08.021.
- 12 5. Habash-Bseiso DE., Rokey R., Berger CJ., Weier AW., Chyou P-H. Accuracy of Noninvasive Ejection
13 Fraction Measurement in a Large Community-Based Clinic. *Clin Medicine Res* 2005;3(2):75–82. Doi:
14 10.3121/cm.3.2.75.
- 15 6. Berlin L. Defending the “Missed” Radiographic Diagnosis. *Am J Roentgenol* 2001;176(2):317–22.
16 Doi: 10.2214/ajr.176.2.1760317.
- 17 7. Berlin L., Hendrix RW. Perceptual errors and negligence. *Am J Roentgenol* 1998;170(4):863–7. Doi:
18 10.2214/ajr.170.4.9530024.
- 19 8. Tromp J., Seekings PJ., Hung C-L., et al. Automated interpretation of systolic and diastolic function on
20 the echocardiogram: a multicohort study. *Lancet Digital Heal* 2022;4(1):e46–54. Doi: 10.1016/s2589-
21 7500(21)00235-1.
- 22 9. Nolan MT., Thavendiranathan P. Automated Quantification in Echocardiography. *JACC Cardiovasc*
23 Imaging 2019;12(6):1073–92. Doi: 10.1016/j.jcmg.2018.11.038.
- 24 10. Ghorbani A., Ouyang D., Abid A., et al. Deep learning interpretation of echocardiograms. *Npj Digital*
25 Medicine 2020;3(1):10. Doi: 10.1038/s41746-019-0216-8.
- 26 11. Zhang J., Gajjala S., Agrawal P., et al. Fully Automated Echocardiogram Interpretation in Clinical
27 Practice. *Circulation* 2018;138(16):1623–35. Doi: 10.1161/circulationaha.118.034338.
- 28 12. Rajpurkar P., Lungren MP. The Current and Future State of AI Interpretation of Medical Images. *New*
29 Engl J Med 2023;388(21):1981–90. Doi: 10.1056/nejmra2301725.
- 30 13. Veen DV., Uden CV., Blankemeier L., et al. Clinical Text Summarization: Adapting Large Language
31 Models Can Outperform Human Experts. *ArXiv* 2023.
- 32 14. Liu Z., Zhong A., Li Y., et al. Radiology-GPT: A Large Language Model for Radiology. *ArXiv* 2023.
- 33 15. Gershanik EF., Lacson R., Khorasani R. Critical finding capture in the impression section of
34 radiology reports. *AMIA Annu Symp Proc AMIA Symp* 2011;2011:465–9.
- 35 16. Fleming SL., Lozano A., Haberkorn WJ., et al. MedAlign: A Clinician-Generated Dataset for
36 Instruction Following with Electronic Medical Records. *ArXiv* 2023. Doi: 10.48550/arxiv.2308.14089.
- 37 17. Singhal K., Tu T., Gottweis J., et al. Towards Expert-Level Medical Question Answering with Large
38 Language Models. *ArXiv* 2023. Doi: 10.48550/arxiv.2305.09617.
- 39 18. Liu Z., Zhong A., Li Y., et al. Radiology-GPT: A Large Language Model for Radiology. *ArXiv* 2023.
40 Doi: 10.48550/arxiv.2306.08666.
- 41 19. Dave T., Athaluri SA., Singh S. ChatGPT in medicine: an overview of its applications, advantages,
42 limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023;6:1169595. Doi:
43 10.3389/frai.2023.1169595.

- 1 20. Liu Y., Han T., Ma S., et al. Summary of ChatGPT-Related research and perspective towards the
2 future of large language models. *Meta-Radiol* 2023;1(2):100017. Doi: 10.1016/j.metrad.2023.100017.
- 3 21. Diaz N. 6 hospitals, health systems testing out ChatGPT. Available at:
4 <https://www.beckershospitalreview.com/innovation/4-hospitals-health-systems-testing-out-chatgpt.html>.
5 Accessed June 2, 2023.
- 6 22. Latif E., Zhai X. Fine-tuning ChatGPT for Automatic Scoring. *ArXiv* 2023. Doi:
7 10.48550/arxiv.2310.10072.
- 8 23. Nath C., Albaghdadi MS., Jonnalagadda SR. A Natural Language Processing Tool for Large-Scale
9 Data Extraction from Echocardiography Reports. *PLoS ONE* 2016;11(4):e0153749. Doi:
10 10.1371/journal.pone.0153749.
- 11 24. Dong T., Sunderland N., Nightingale A., et al. Development and Evaluation of a Natural Language
12 Processing System for Curating a Trans-Thoracic Echocardiogram (TTE) Database. *Bioengineering*
13 2023;10(11):1307. Doi: 10.3390/bioengineering10111307.
- 14 25. Zheng C., Sun BC., Wu Y-L., et al. Automated interpretation of stress echocardiography reports using
15 natural language processing. *Eur Hear J - Digit Heal* 2022;3(4):626–37. Doi: 10.1093/ehjdh/ztac047.
- 16 26. Tang L., Kooragayalu S., Wang Y., et al. EchoGen: Generating Conclusions from Echocardiogram
17 Notes. *Proc 21st Work Biomed Lang Process* 2022;2022:359–68. Doi: 10.18653/v1/2022.bionlp-1.35.
- 18 27. Vaid A., Duong SQ., Lampert J., et al. Local large language models for privacy-preserving
19 accelerated review of historic echocardiogram reports. *J Am Méd Inform Assoc* 2024:ocae085. Doi:
20 10.1093/jamia/ocae085.
- 21 28. Touvron H., Martin L., Stone K., et al. Llama 2: Open Foundation and Fine-Tuned Chat Models.
22 *ArXiv* 2023. Doi: 10.48550/arxiv.2307.09288.
- 23 29. Tunstall L., Beeching E., Lambert N., et al. Zephyr: Direct Distillation of LM Alignment. *ArXiv*
24 2023. Doi: 10.48550/arxiv.2310.16944.
- 25 30. Kwak G., Moukheiber D., Moukheiber M., et al. EchoNotes Structured Database derived from
26 MIMIC-III (ECHO-NOTE2NUM). *PhysioNet* 2024. Doi: 10.13026/xhrz-ht59.
- 27 31. Loper E., Bird S. NLTK: the Natural Language Toolkit. *Proc ACL-02 Work Eff Tools Methodol*
28 *Teach Nat Lang Process Comput Linguistics - 2002*:63–70. Doi: 10.3115/1118108.1118117.
- 29 32. Han T., Adams LC., Papaioannou J-M., et al. MedAlpaca -- An Open-Source Collection of Medical
30 Conversational AI Models and Training Data. *ArXiv* 2023. Doi: 10.48550/arxiv.2304.08247.
- 31 33. Raffel C., Shazeer N., Roberts A., et al. Exploring the Limits of Transfer Learning with a Unified
32 Text-to-Text Transformer. *ArXiv* 2019. Doi: 10.48550/arxiv.1910.10683.
- 33 34. Cowan BR., Clark L., Følstad A., Skjuve M. Chatbots for customer service: user experience and
34 motivation. *Proc 1st Int Conf Conversational User Interfaces* 2019:1–9. Doi: 10.1145/3342775.3342784.
- 35 35. Li M., Gong S., Feng J., et al. In-Context Learning with Many Demonstration Examples. *ArXiv* 2023.
36 Doi: 10.48550/arxiv.2302.04931.
- 37 36. Choi E., Jo Y., Jang J., Seo M. Prompt Injection: Parameterization of Fixed Inputs. *ArXiv* 2022. Doi:
38 10.48550/arxiv.2206.11349.
- 39 37. Gu Y., Tinn R., Cheng H., et al. Domain-Specific Language Model Pretraining for Biomedical
40 Natural Language Processing. *ArXiv* 2020. Doi: 10.48550/arxiv.2007.15779.
- 41 38. Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized
42 LLMs. *ArXiv* 2023. Doi: 10.48550/arxiv.2305.14314.
- 43 39. Hu EJ., Shen Y., Wallis P., et al. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv*
44 2021. Doi: 10.48550/arxiv.2106.09685.

- 1 40. Veen DV., Uden CV., Blankemeier L., et al. Adapted large language models can outperform medical
2 experts in clinical text summarization. *Nat Med* 2024;1–9. Doi: 10.1038/s41591-024-02855-5.
- 3 41. Tang L., Sun Z., Iday B., et al. Evaluating large language models on medical evidence
4 summarization. *Npj Digit Med* 2023;6(1):158. Doi: 10.1038/s41746-023-00896-7.
- 5 42. Papineni K., Roukos S., Ward T., Zhu W-J. BLEU: a method for automatic evaluation of machine
6 translation. *Proc 40th Annu Meet Assoc Comput Linguistics - ACL '02* 2002:311–8. Doi:
7 10.3115/1073083.1073135.
- 8 43. Lavie A., Agarwal A. Meteor: an automatic metric for MT evaluation with high levels of correlation
9 with human judgments 2007:228–31. Doi: 10.1145/1626355.1626389.
- 10 44. Zhang T., Kishore V., Wu F., Weinberger KQ., Artzi Y. BERTScore: Evaluating Text Generation
11 with BERT. *ArXiv* 2019. Doi: 10.48550/arxiv.1904.09675.
- 12 45. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. vol. *Text Summarization*
13 *Branches Out*. Association for Computational Linguistics; n.d. p. 74–81.
- 14 46. Delbrouck J-B., Chambon P., Bluethgen C., Tsai E., Almusa O., Langlotz CP. Improving the Factual
15 Correctness of Radiology Report Generation with Semantic Rewards. *ArXiv* 2022. Doi:
16 10.48550/arxiv.2210.12186.
- 17 47. Jain S., Agrawal A., Saporta A., et al. RadGraph: Extracting Clinical Entities and Relations from
18 Radiology Reports. *ArXiv* 2021. Doi: 10.48550/arxiv.2106.14463.
- 19 48. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22(3):276–82.
- 20 49. Landis JR., Koch GG. The measurement of observer agreement for categorical data. *Biometrics*
21 1977;33(1):159–74.
- 22 50. Zhao TZ., Wallace E., Feng S., Klein D., Singh S. Calibrate Before Use: Improving Few-Shot
23 Performance of Language Models. *ArXiv* 2021. Doi: 10.48550/arxiv.2102.09690.
- 24 51. Wang L., Chen X., Deng X., et al. Prompt engineering in consistency and reliability with the
25 evidence-based guideline for LLMs. *Npj Digit Med* 2024;7(1):41. Doi: 10.1038/s41746-024-01029-4.
- 26 52. Wang S., Zhao Z., Ouyang X., Wang Q., Shen D. ChatCAD: Interactive Computer-Aided Diagnosis
27 on Medical Image using Large Language Models. *Arxiv* 2023. Doi: 10.48550/arxiv.2302.07257.
- 28 53. Mallio CA., Sertorio AC., Bernetti C., Zobel BB. Large language models for structured reporting in
29 radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. *Radiol Med* 2023:1–5. Doi:
30 10.1007/s11547-023-01651-4.
- 31 54. Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP
32 Tasks. *ArXiv* 2020. Doi: 10.48550/arxiv.2005.11401.

medRxiv preprint doi: <https://doi.org/10.1101/2024.01.18.24301503>; this version posted June 28, 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](#) .

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



medRxiv preprint doi: <https://doi.org/10.1101/2024.01.18.24301503>; this version posted June 28, 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.

Finding-final impression pairs from Mayo echocardiography reports

Quantitative evaluation

Best Model



EchoGPT

Qualitative evaluation

30 Cases

Clinical Reader Study by Echo Experts: the 4C metrics

Completeness

Conciseness

Correctness

Clinical Utility

Model Adaptation Methods

- In-Context Learning
- QLoRA Instruction Fine-Tuning

Figure 2

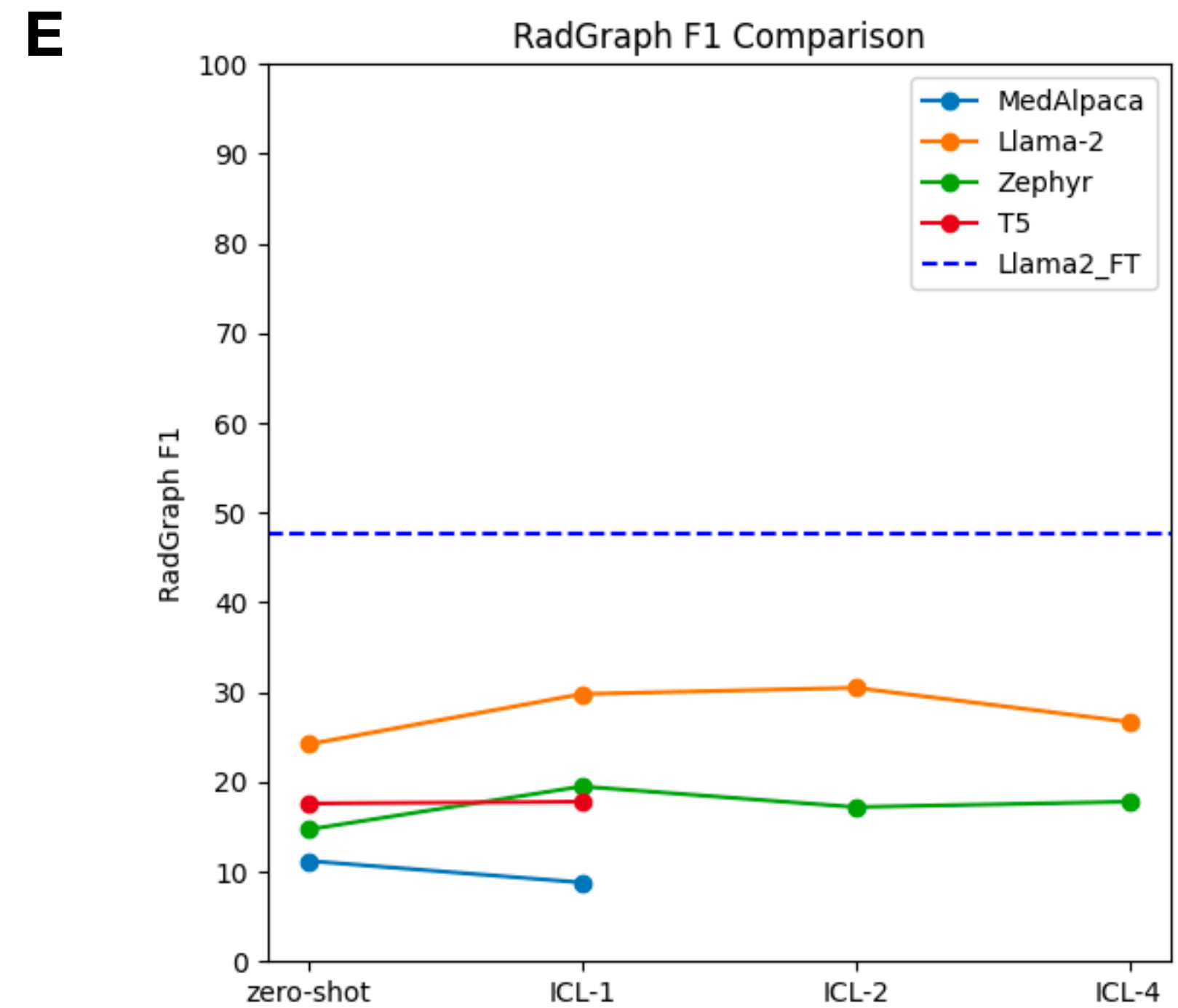
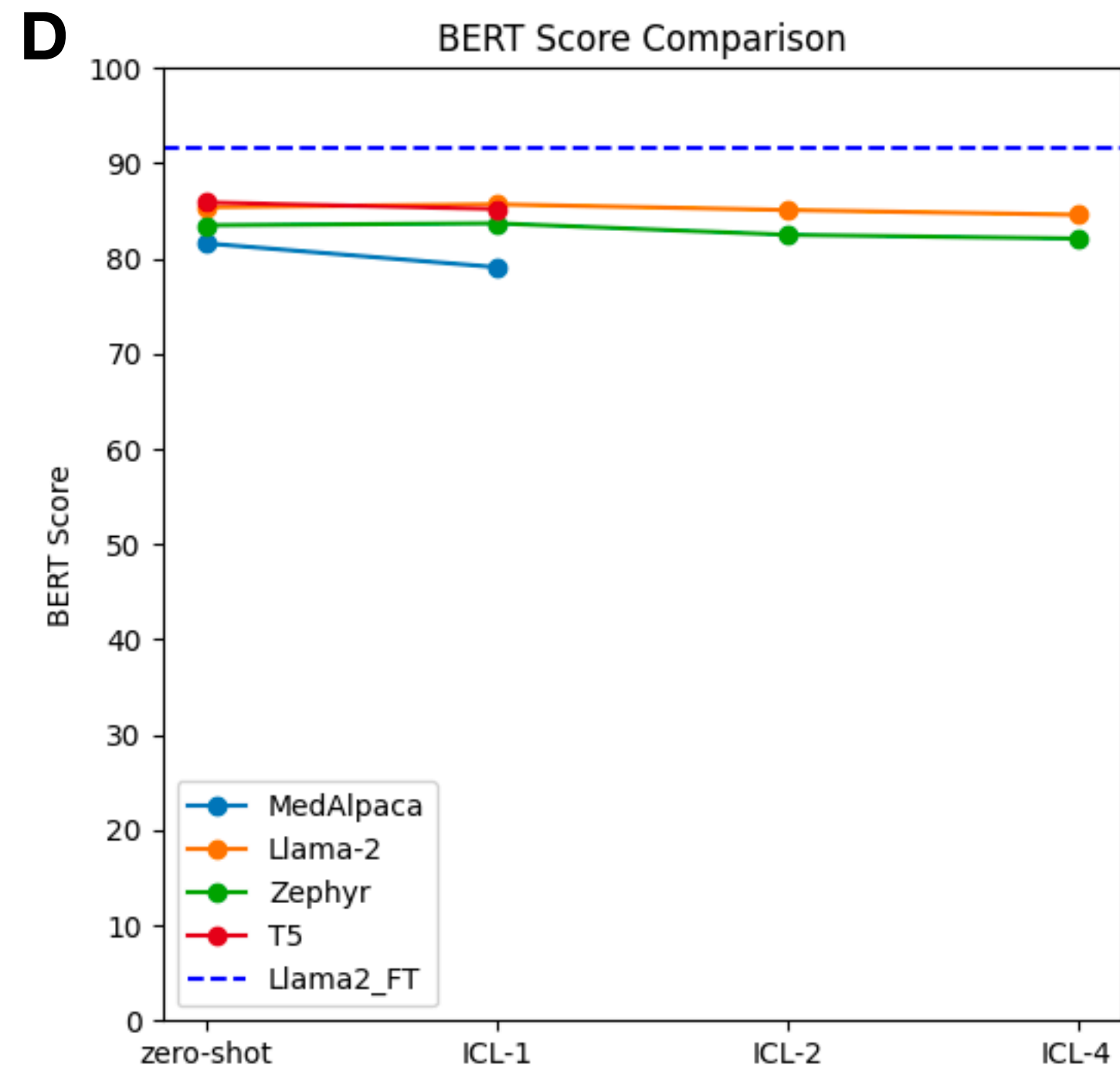
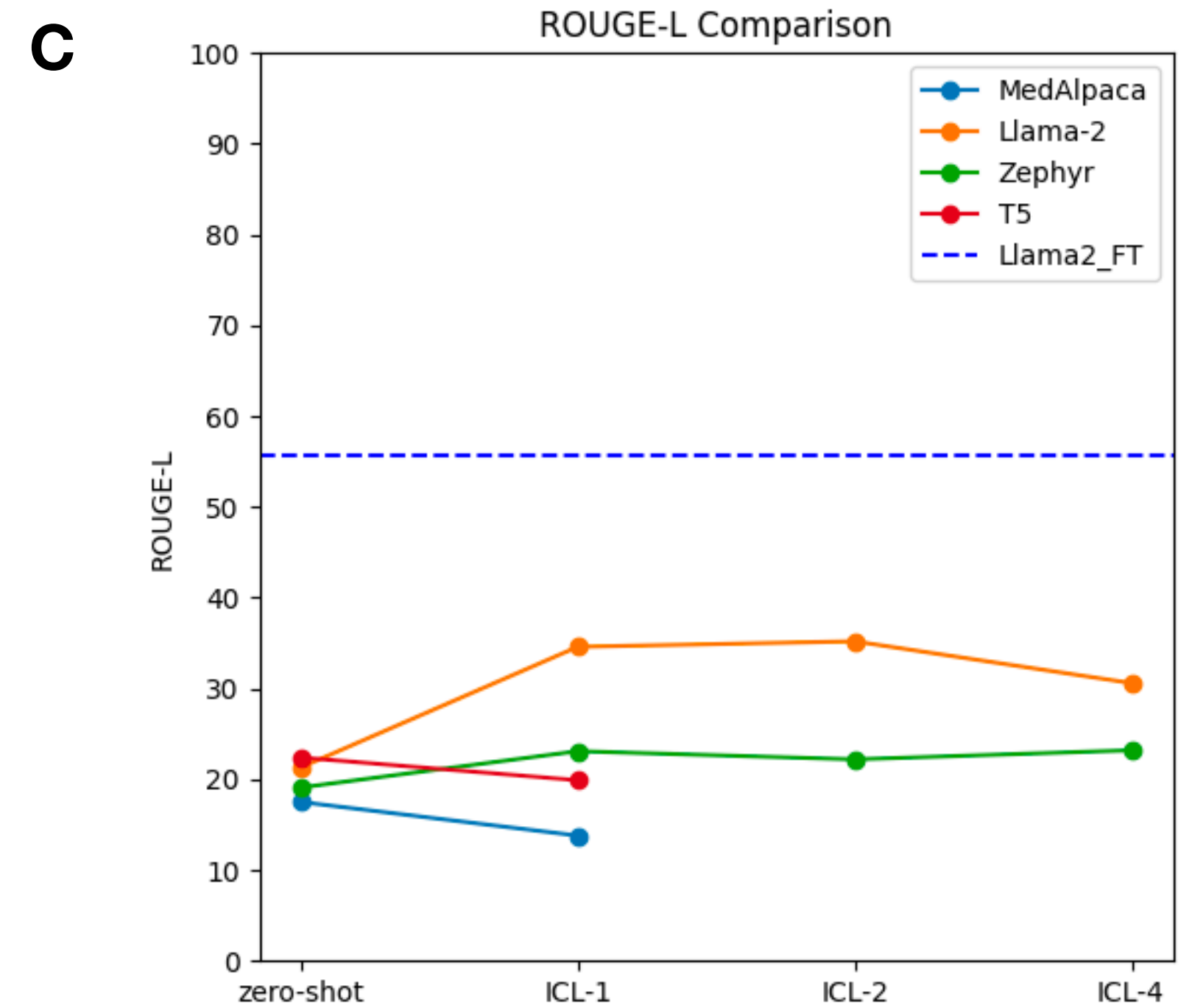
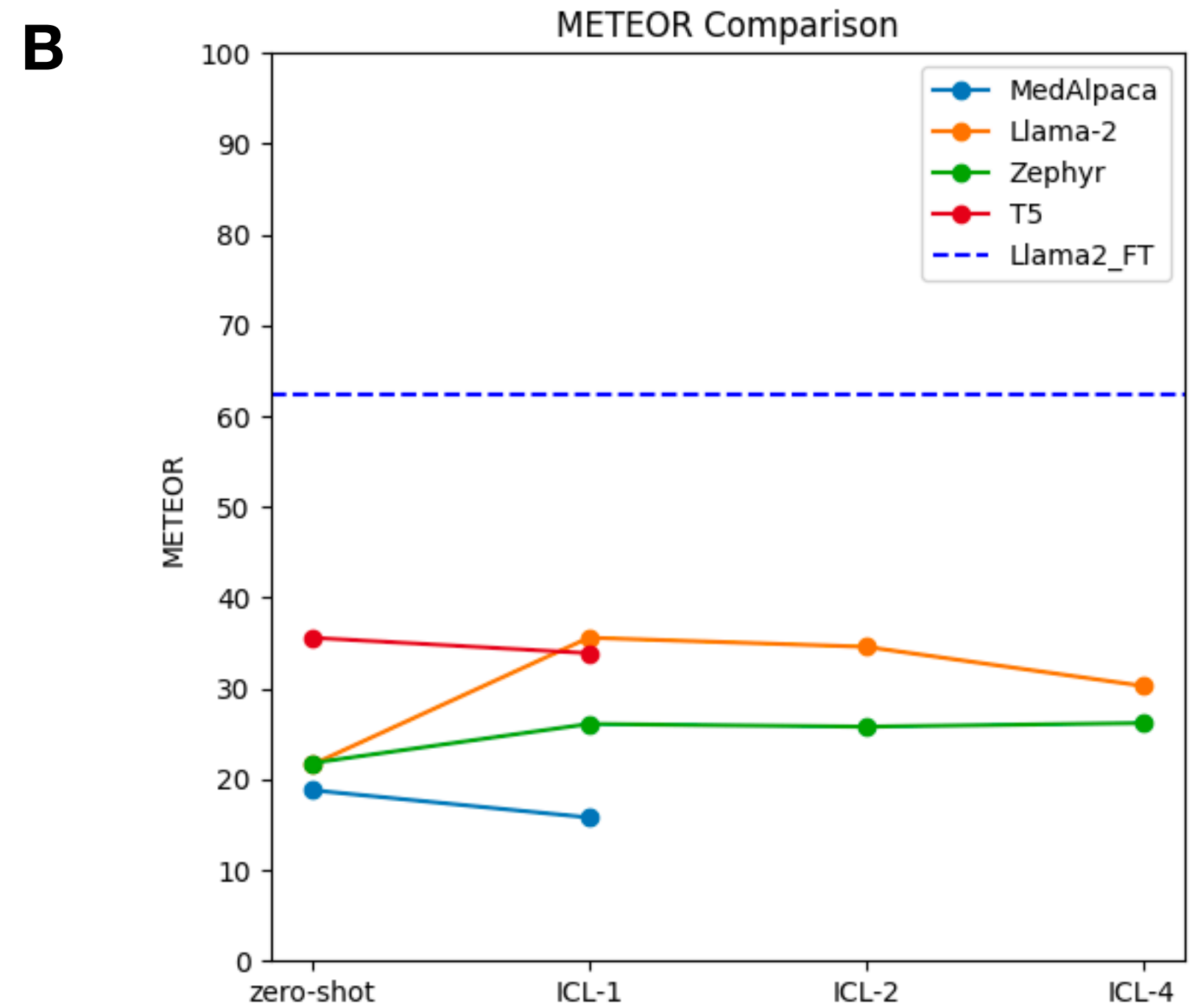
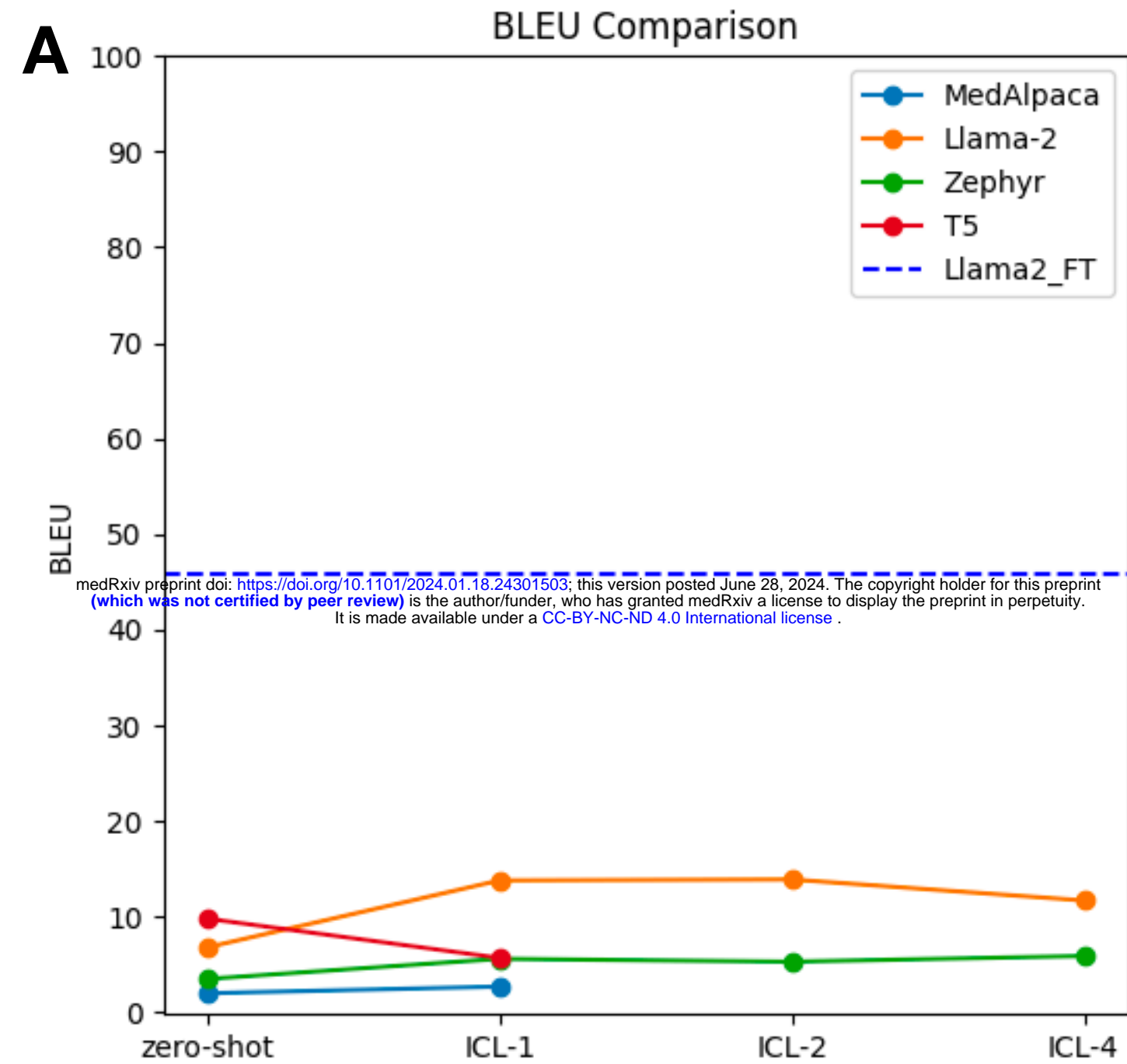
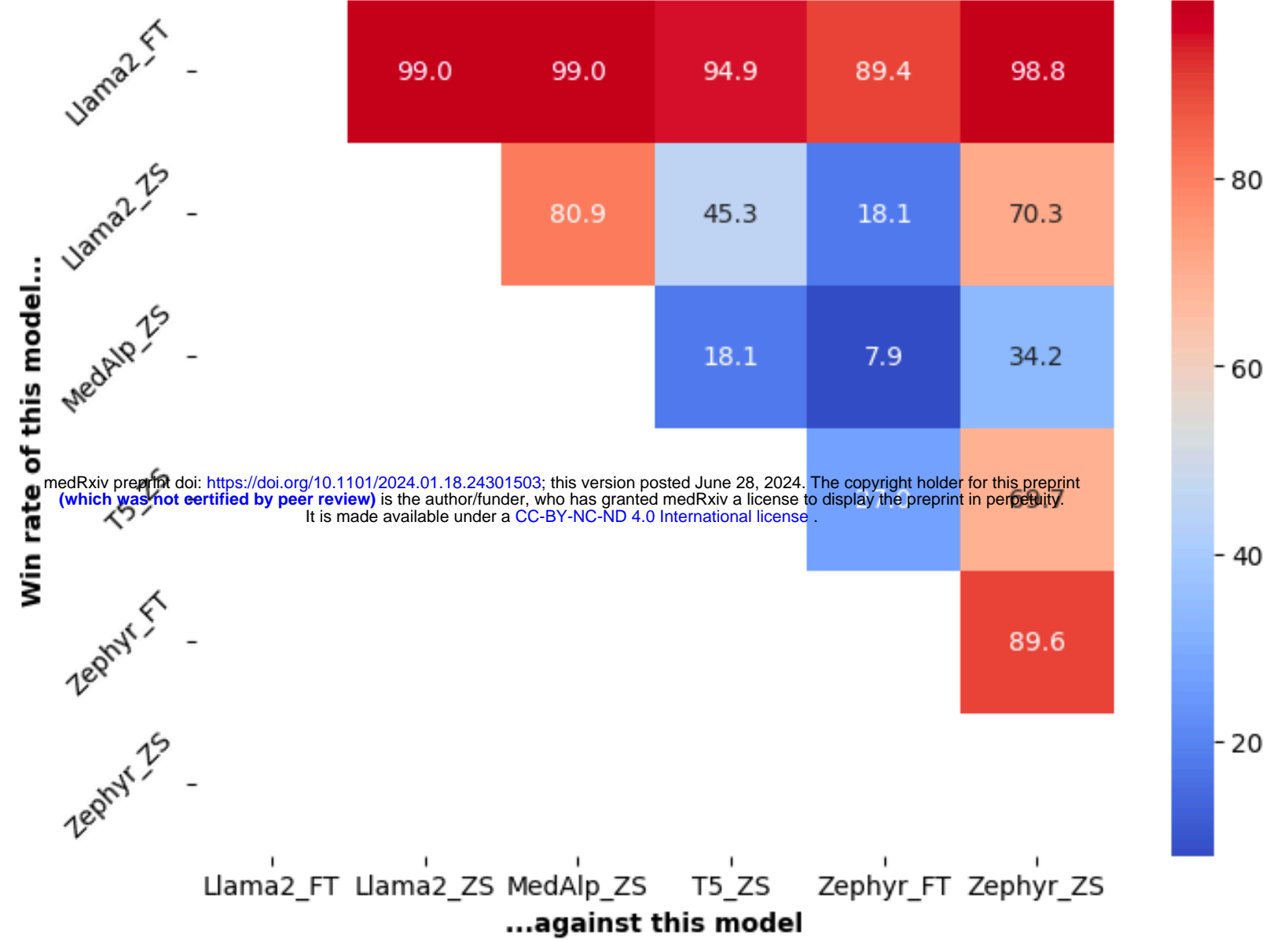
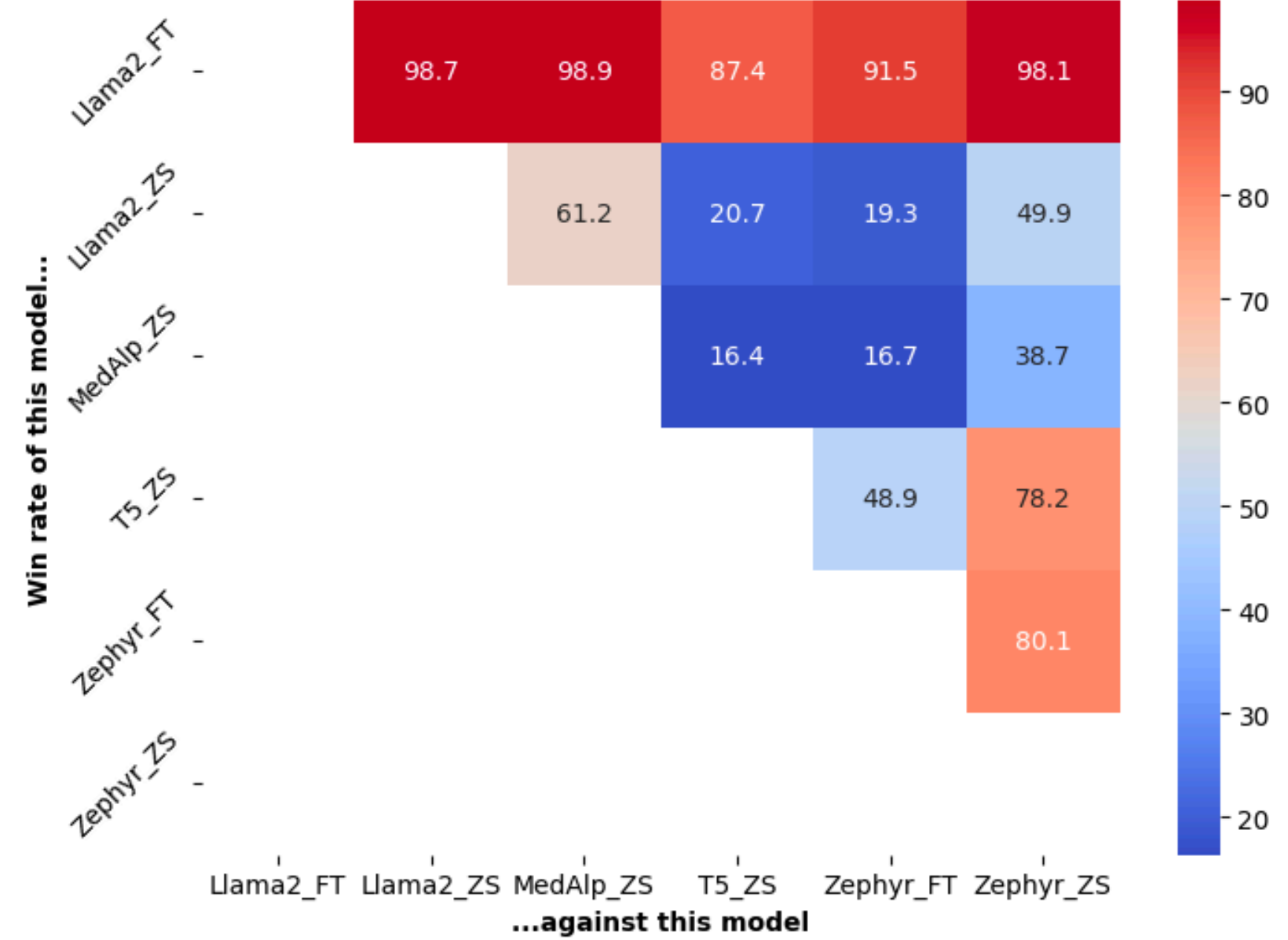
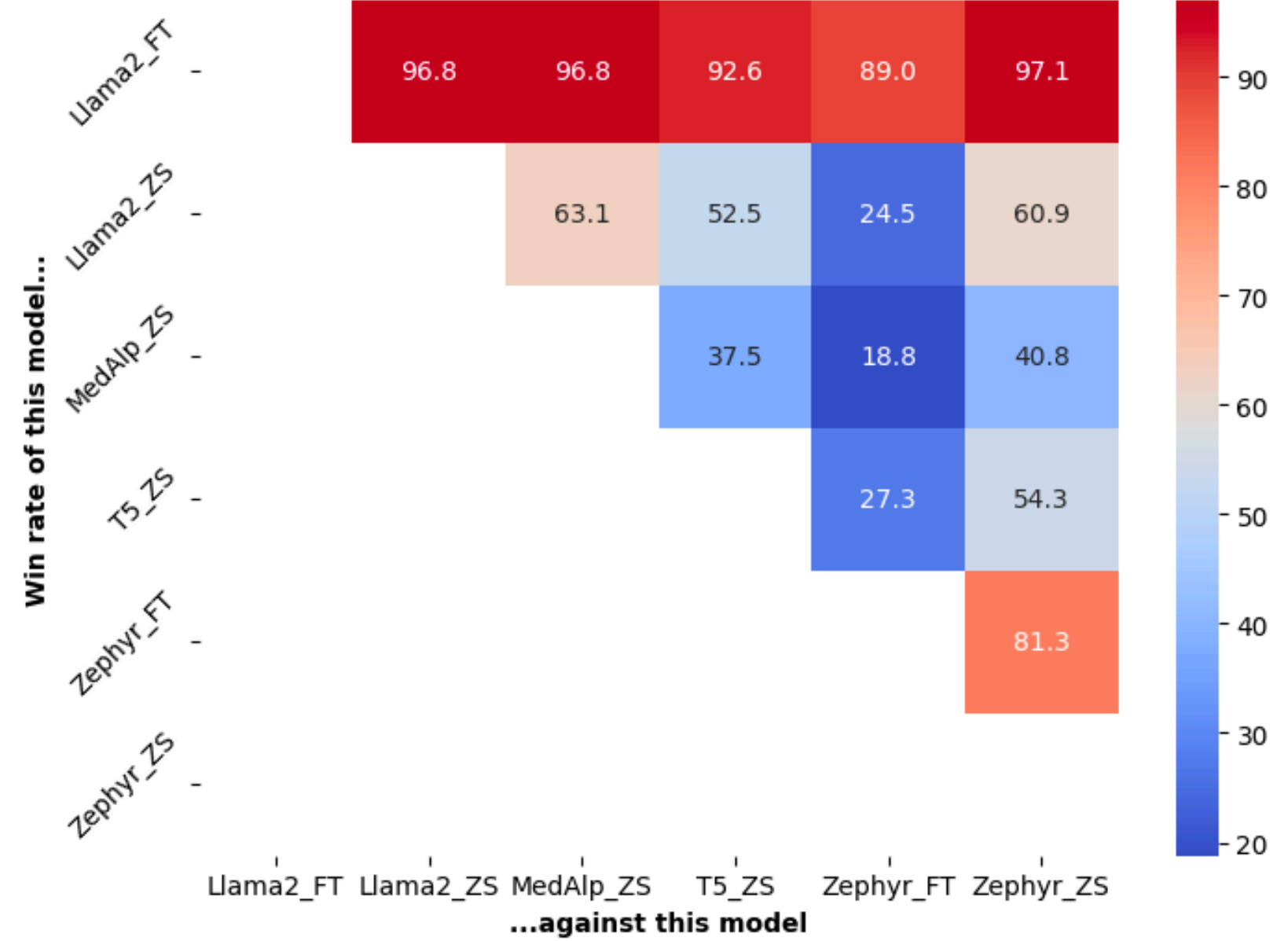
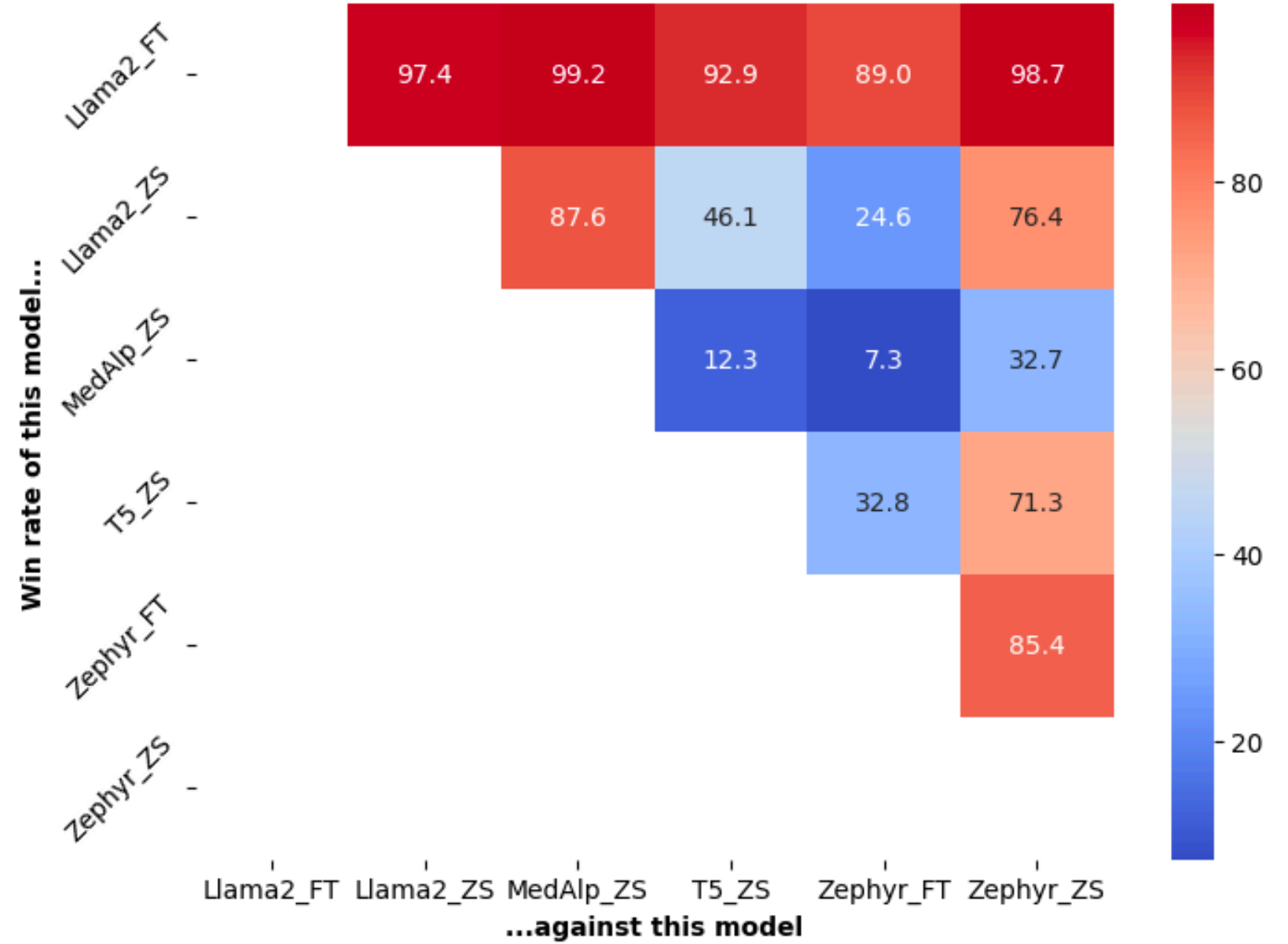
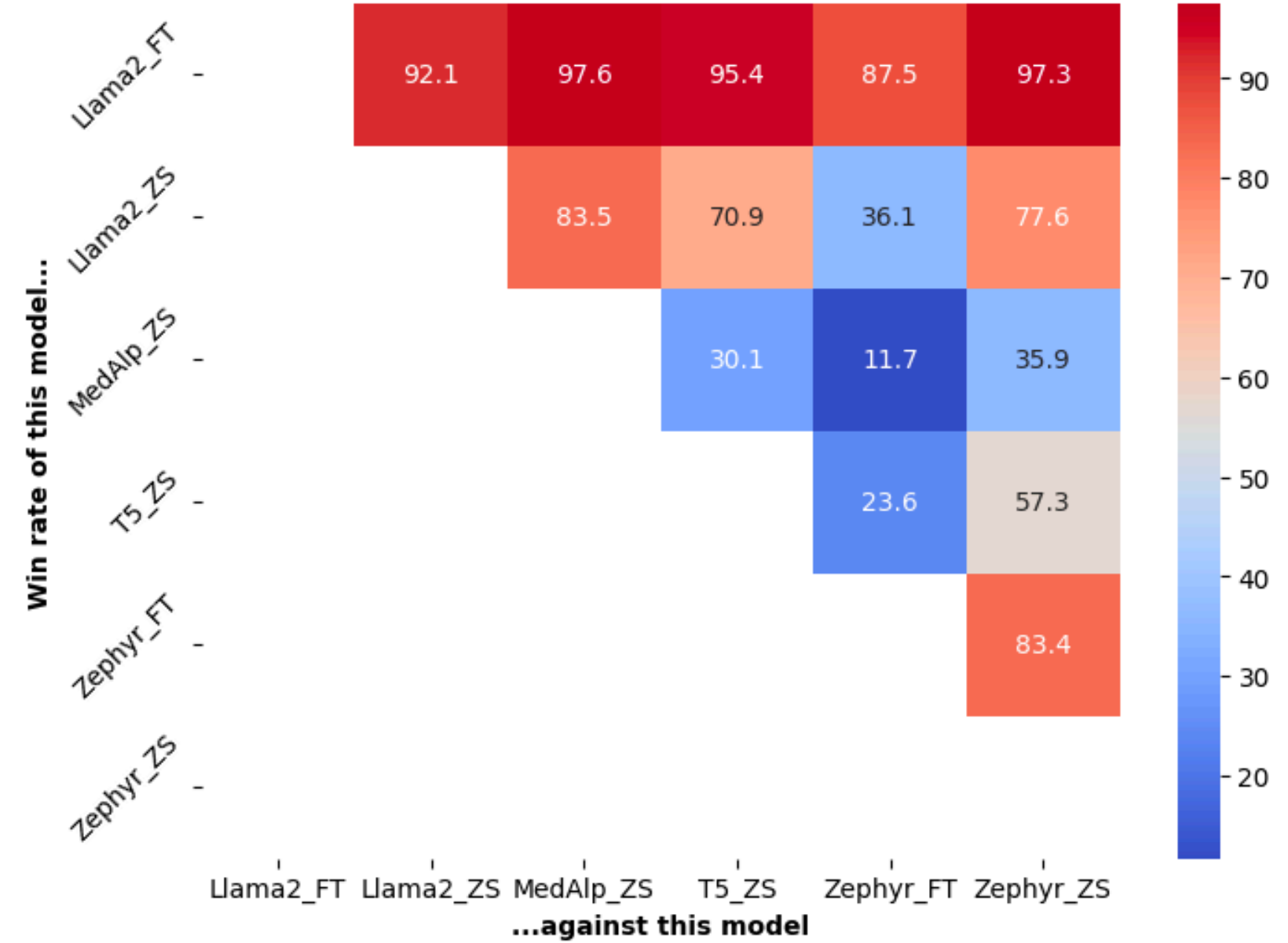
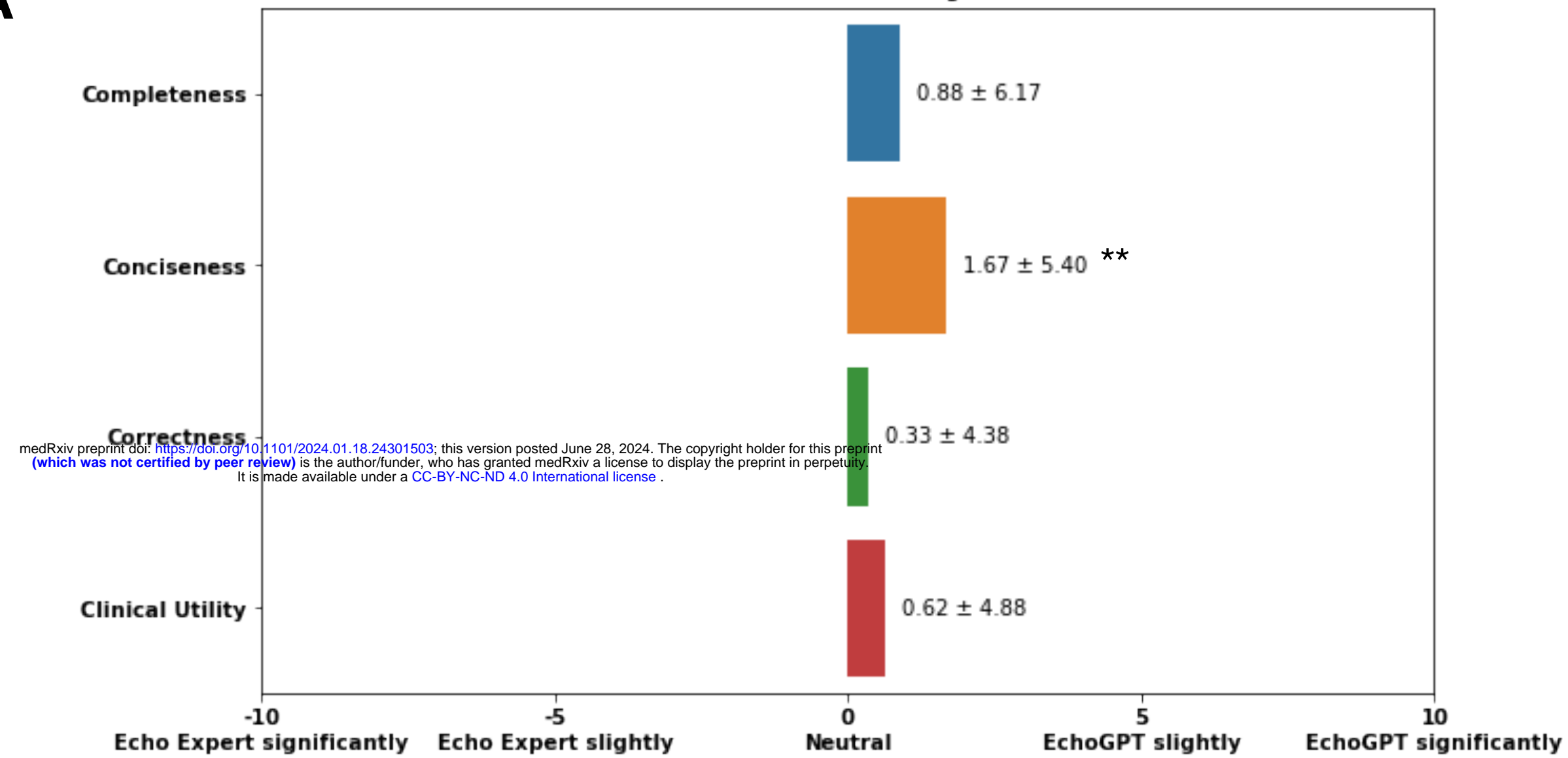


Figure 3

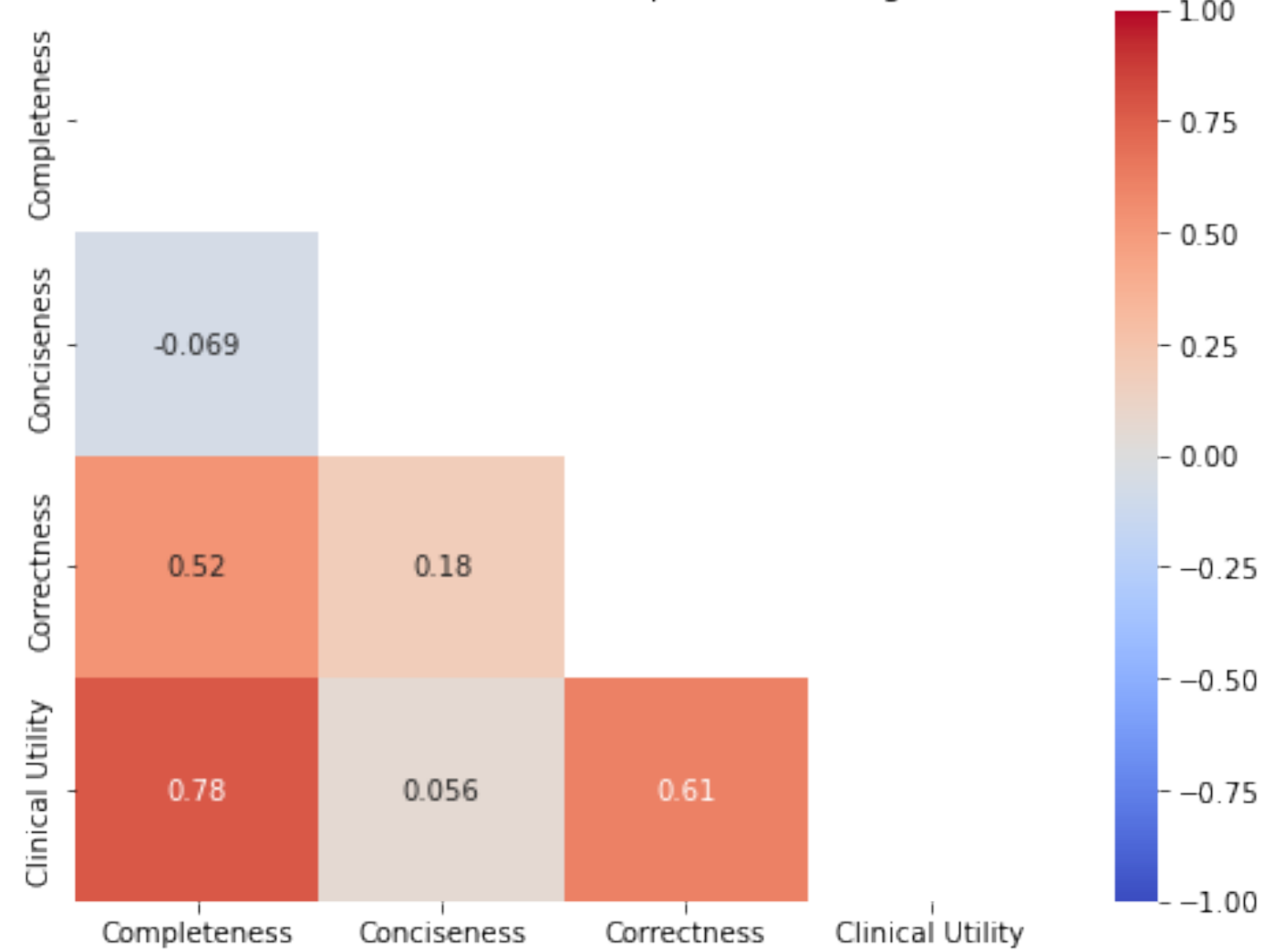
A**BLEU Score Win Rate Comparisons****B****METEOR Score Win Rate Comparisons****C****ROUGE-L Score Win Rate Comparisons****D****BERT Score Win Rate Comparisons****E****RadGraph F1 Win Rate Comparisons****Figure 4 - win rate**

A

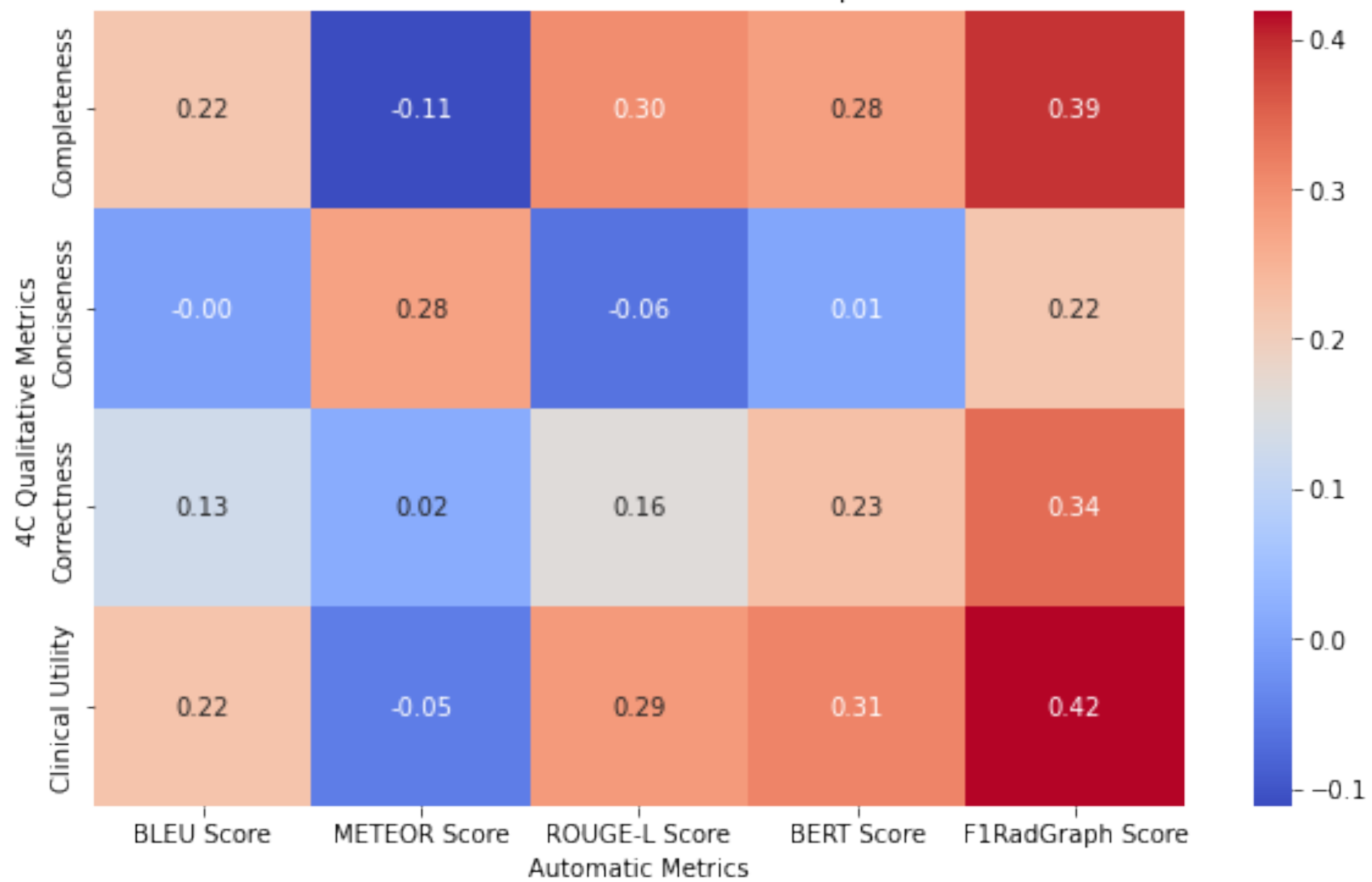
Mean Values for Categories

**B**

Pearson Correlation Heatmap (Lower Triangle)


**C**

Pearson Correlation Heatmap

**Figure 5- Human eval**

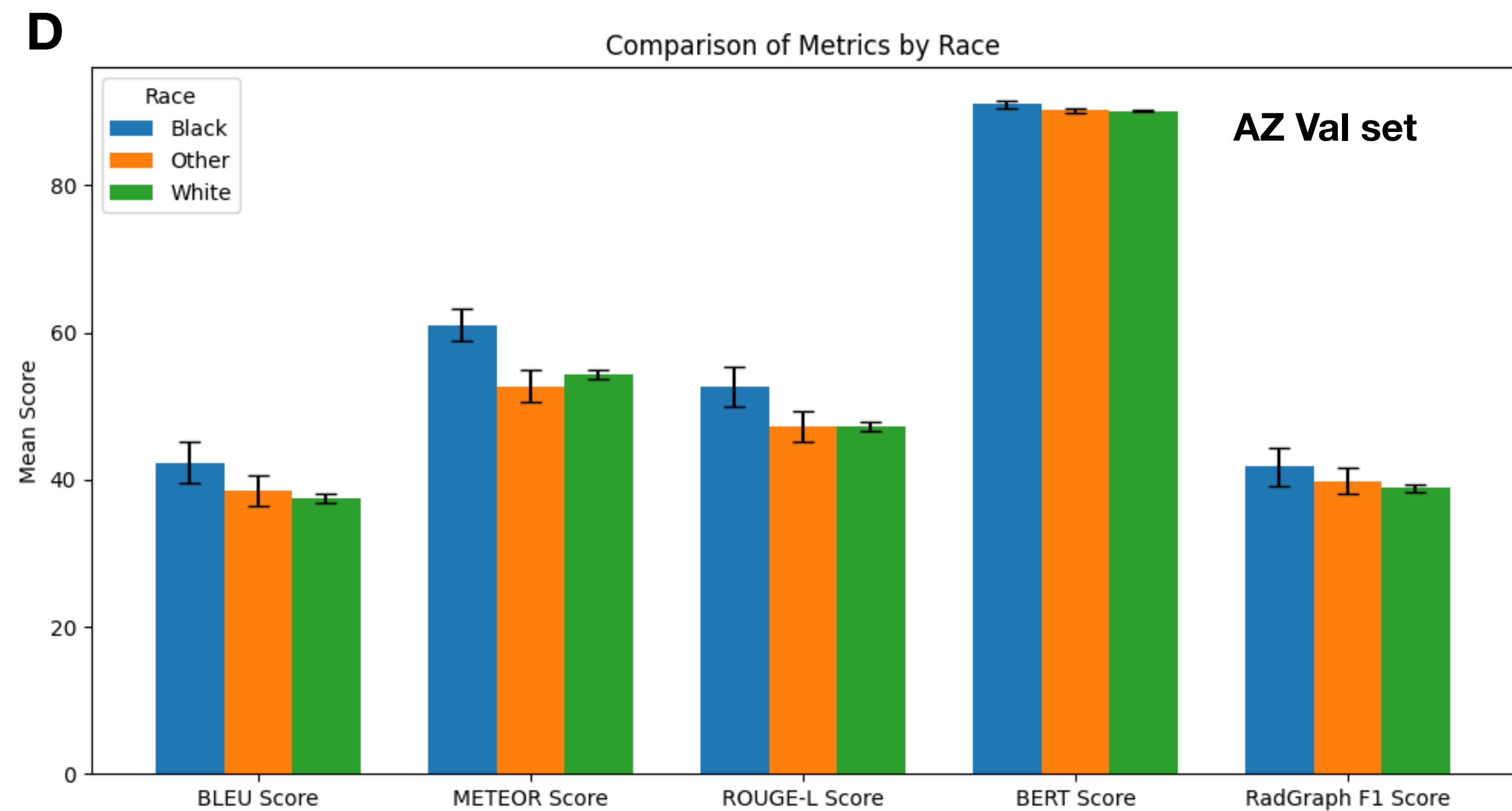
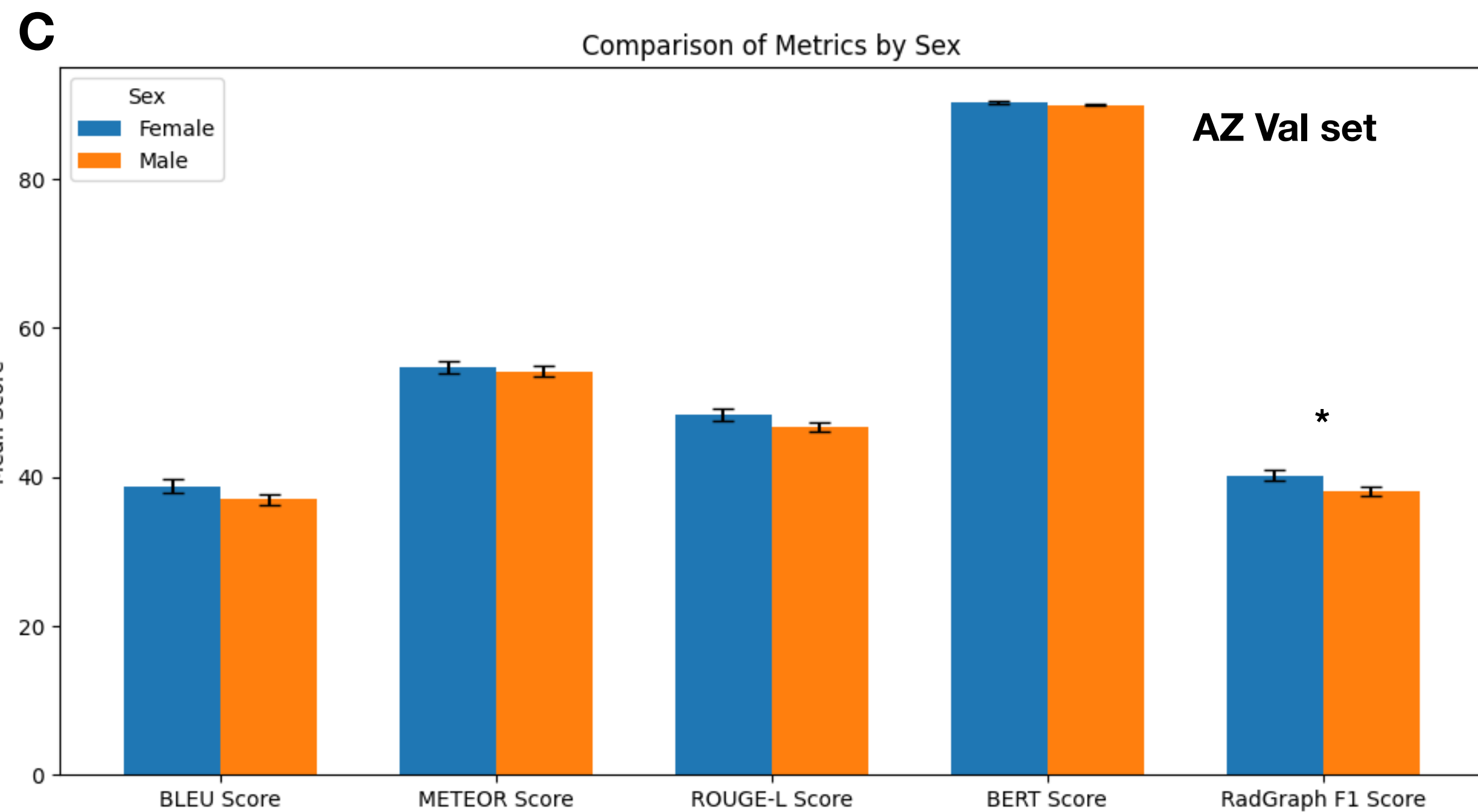
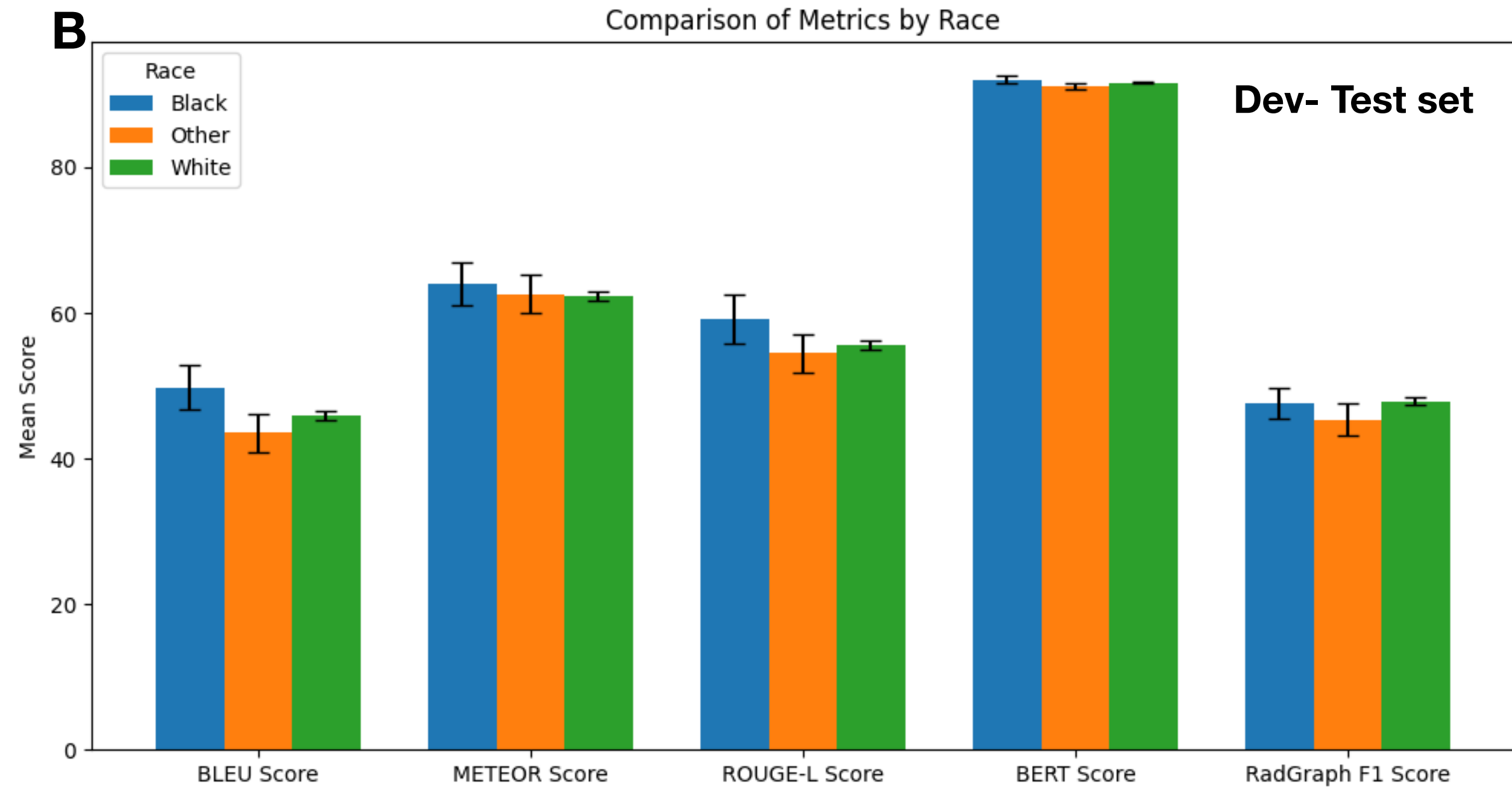
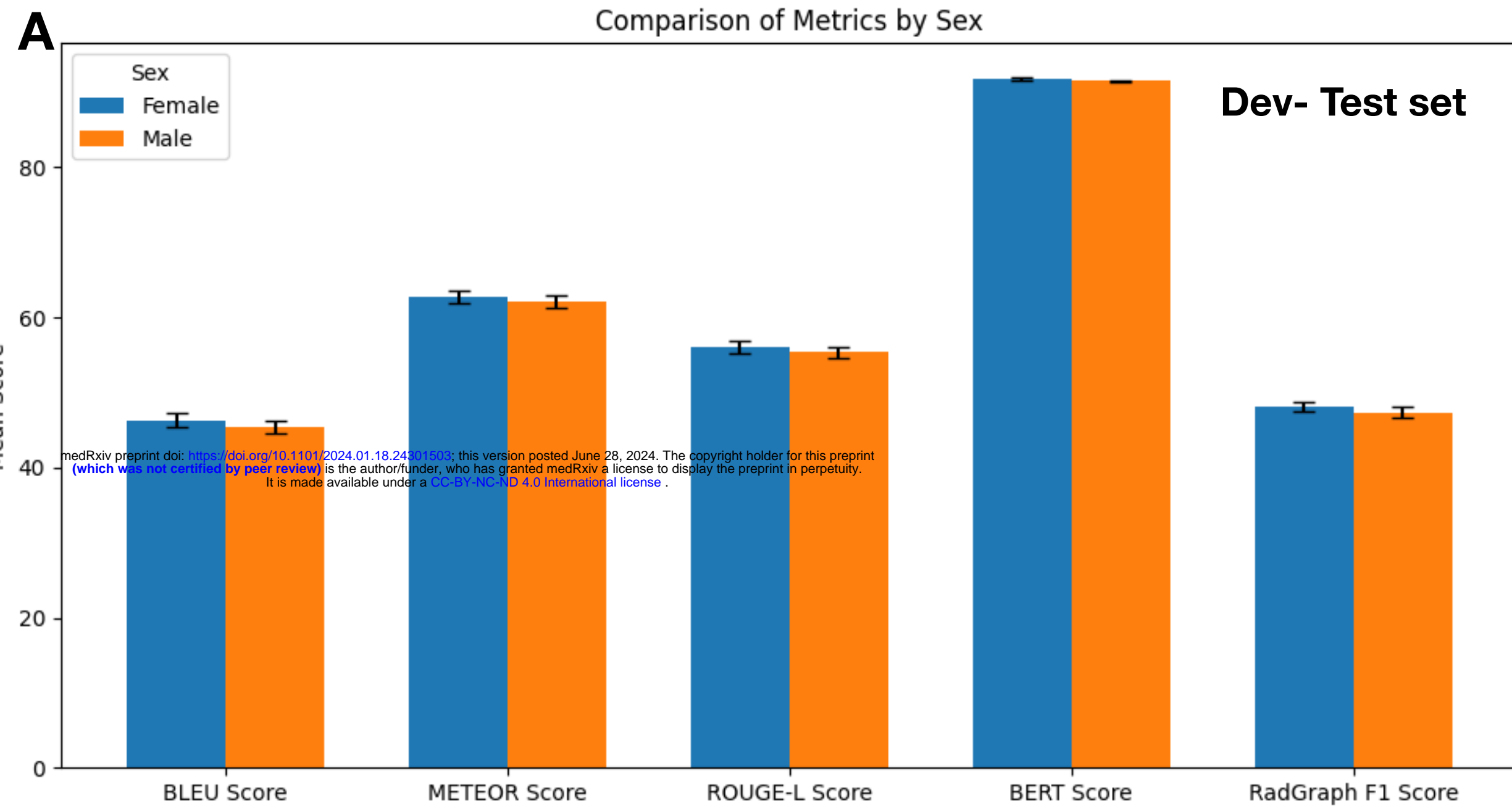
Findings: no intracardiac mass or thrombus, but the left atrial appendage cannot be visualized adequately with transthoracic echo to exclude thrombus in this location. normal left ventricular wall thickness. colorflow and spectral doppler were performed to assess valvular heart disease. echocardiographic 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%). * 

medRxiv preprint doi: <https://doi.org/10.1101/2024.01.18.24301503>; this version posted June 28, 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](https://creativecommons.org/licenses/by-nc-nd/4.0/).

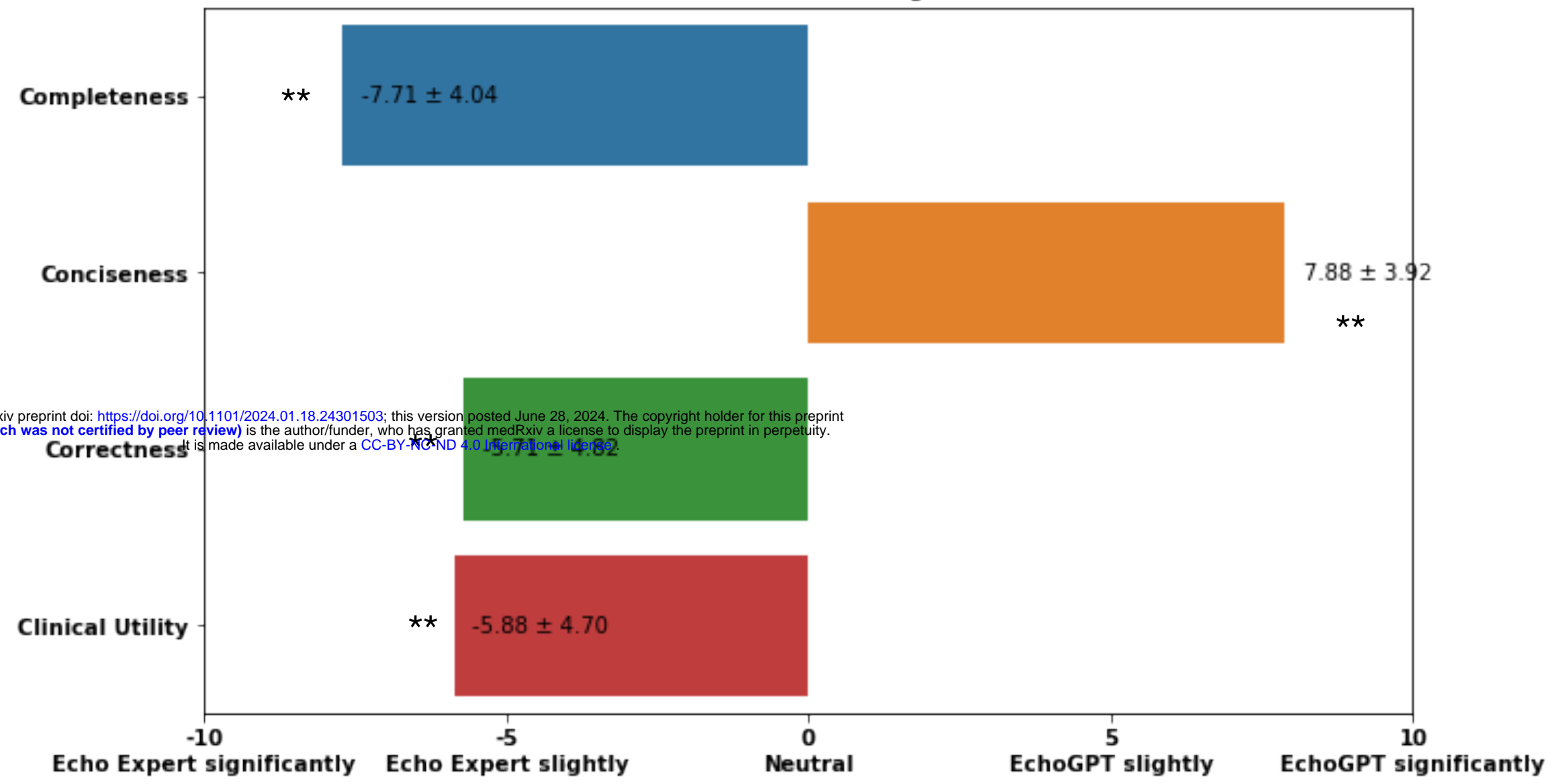
| | Summary A | Summary A slightly | Neutral | Summary B slightly | Summary B |
|------------------|-----------------------|----------------------------------|----------------------------------|----------------------------------|-----------------------|
| Completeness | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input checked="" type="radio"/> | <input type="radio"/> |
| Conciseness | <input type="radio"/> | <input type="radio"/> | <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Correctness | <input type="radio"/> | <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Clinical Utility | <input type="radio"/> | <input type="radio"/> | <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> |



Supplemental Figure 2- bias

A

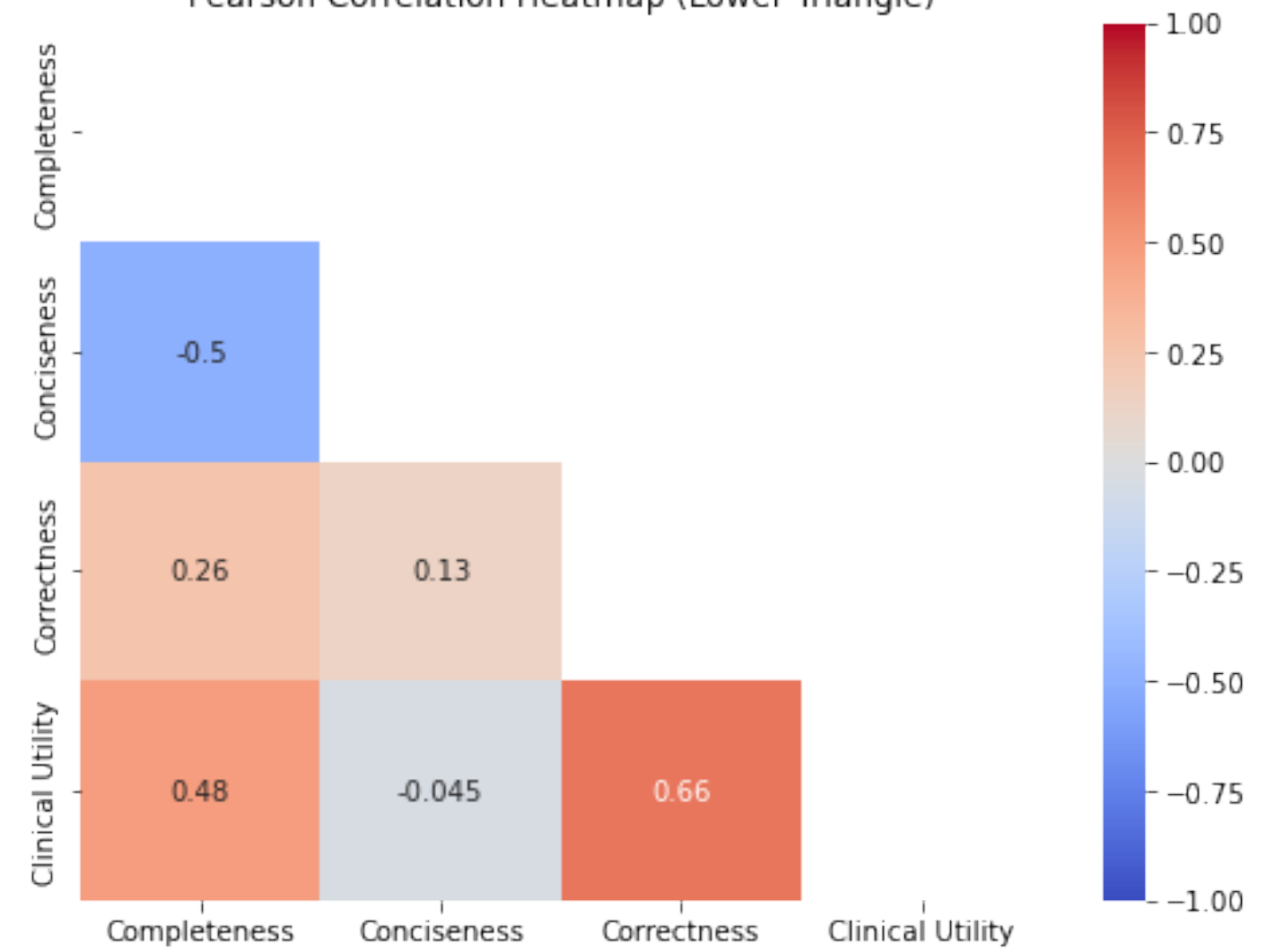
Mean Values for Categories



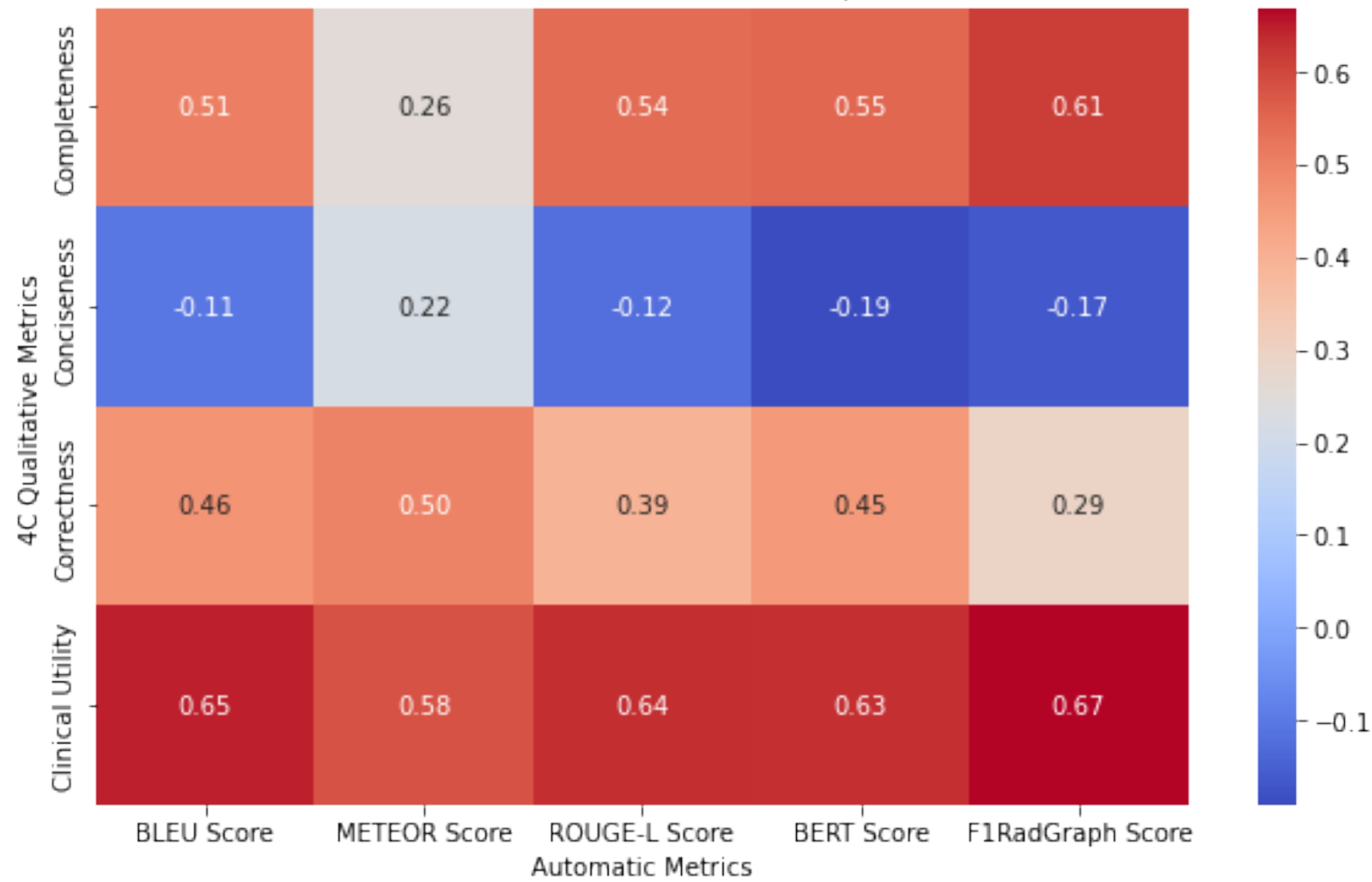
medRxiv preprint doi: <https://doi.org/10.1101/2024.01.18.24301503>; this version posted June 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](#).

B

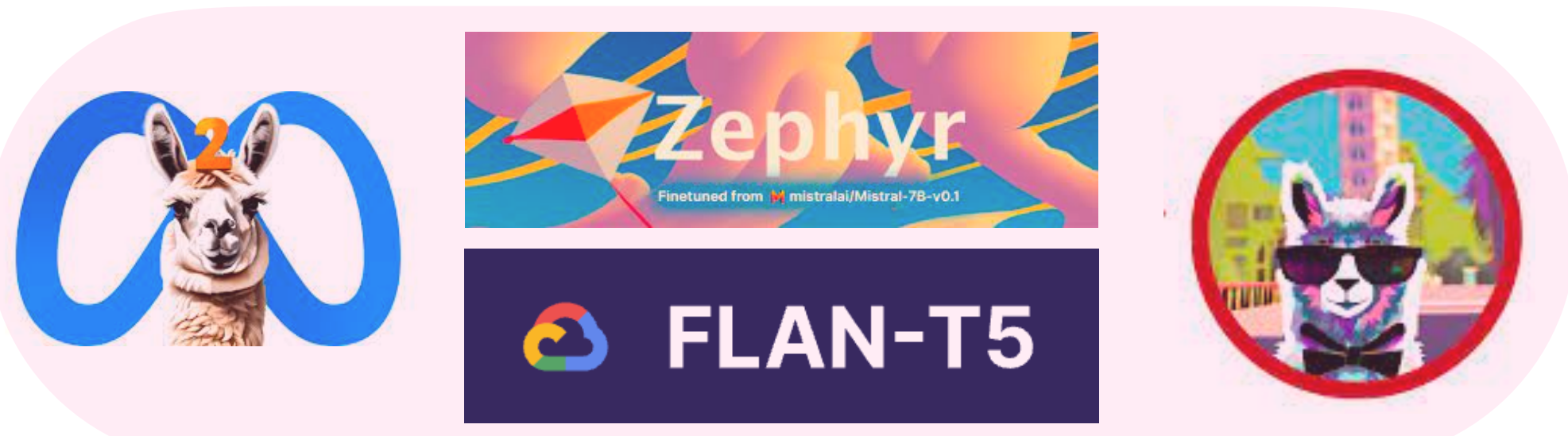
Pearson Correlation Heatmap (Lower Triangle)

**C**

Pearson Correlation Heatmap

**Supplemental Figure 3 - Human eval on MIMIC-Echo**

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



medRxiv preprint doi: <https://doi.org/10.1101/2024.01.18.24301503>; this version posted June 28, 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.

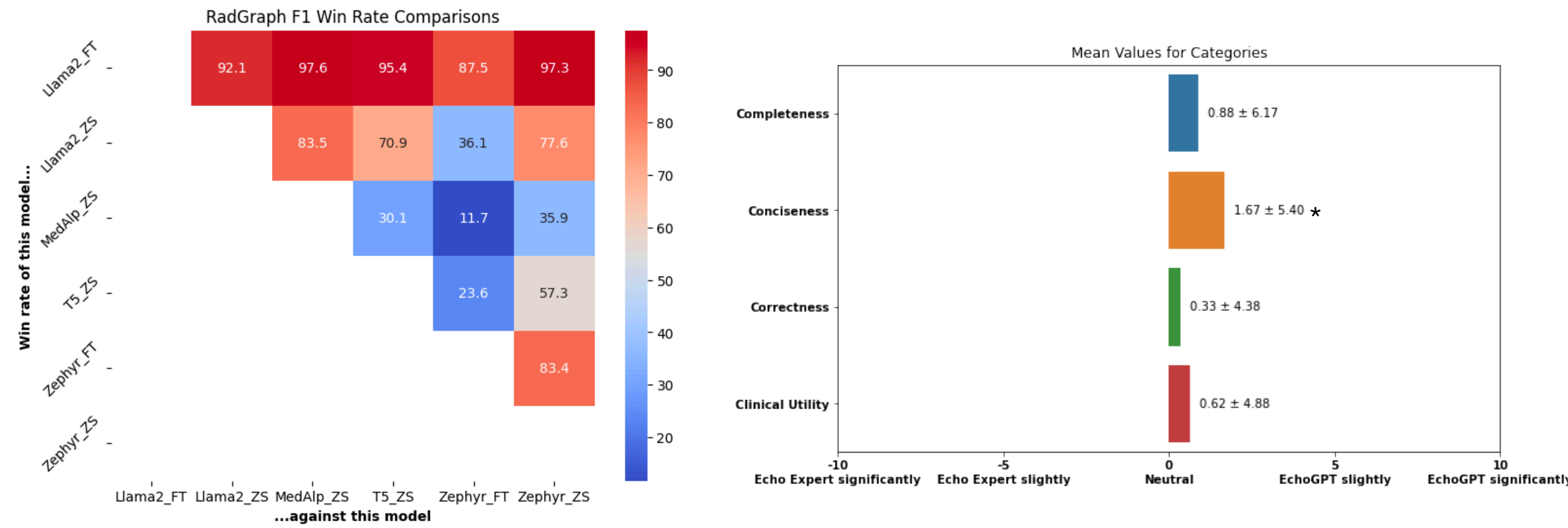
Finding-final impression pairs from Mayo echocardiography reports

Best Model (EchoGPT)

Model Adaptation Methods

- In-Context Learning
- QLoRA Instruction Fine-Tuning

EchoGPT: Outperformed other LLMs, comparable to human reports.



Challenges in Evaluation

- Human Expert Evaluation: Not Scalable
- Automatic Metrics: Not Sensitive to Change in Measurements
- Automatic vs. Human Metrics: Fair-to-modest Correlation