European Society of Cardiology
European Heart Journal - Digital Health (2025) **6**, 5–6
https://doi.org/10.1093/ehjdh/ztae079

EDITORIAL

Fine-tuned large language models can generate expert-level echocardiography reports

Achille Sowa[1,2,3] and Robert Avram[1,2,3,*]

[1]Department of Medicine, Université de Montréal, 5000 Bélanger Street, Montreal, Québec H3T 1J4, Canada; [2]Heartwise (heartwise.ai), Montreal Heart Institute, 5000 Bélanger Street, Montreal, Québec H1T 1C8, Canada; and [3]Department of Medicine, Montreal Heart Institute, Université de Montréal, 2900 Edouard Montpetit Blvd, Montreal, Québec H1T 1C8, Canada

Online publish-ahead-of-print 24 October 2024

Introduction

Trans-thoracic echocardiograms (TTEs) are one of the most requested non-invasive cardiac imaging tests worldwide.[1,2] It consists of 30–60 videos that physicians interpret and synthesize into comprehensive reports. These reports typically include detailed 'Findings' sections, which list all observations and measurements, and concise 'Impressions' sections that summarize the most clinically relevant information. This process is not only time-consuming but also requires significant expertise to ensure accuracy and clinical relevance.[3] The increasing demand for TTEs coupled with long waiting times[4] has created a pressing need for more efficient reporting methods. Artificial intelligence (AI), particularly large language models (LLMs), offers a promising solution to partially automate and streamline the trans-thoracic echocardiograms report creation, potentially enhancing both the speed and consistency of echocardiography reporting.

In the present issue of the *European Heart Journal – Digital Health*, Chao *et al.*[5] report their findings on the development and evaluation of EchoGPT, a LLM fine-tuned for echocardiography report summarization. They trained EchoGPT, based on the LLaMA-2-7 Billion model[6] on a substantial dataset of 97 506 echocardiogram reports from Mayo Clinic to generate the 'Impressions' section, from the 'Findings' section.

Due to patient privacy policy regulations, EchoGPT was compared exclusively against open-source baseline models. These included two general-purpose LLMs: Llama-2-7b-chat[6] and Zephyr-7b,[7] one medical-specific LLM: Med-Alpaca[8] and a sequence-to-sequence model: Flan-T5. The study employed several common natural language processing (NLP) metrics, each providing a different perspective on the quality of the generated text. BLEU (Bilingual Evaluation Understudy) measures n-gram precision, METEOR (Metric for Evaluation of Translation with Explicit ORdering) considers synonyms and word order, ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence) focuses on the longest common subsequence, and BERT (Bidirectional encoder representation from transformers) Score evaluates semantic similarity using contextual embeddings. The RadGraph-F1 metric,[9] a medical-specific measure for assessing the accuracy of structured data representation in medical text, was used as the primary evaluation criterion for ranking different models based on the factual correctness of the generated clinical content. These metrics range from 0 to 100, with higher scores indicating better performance. The results showed that EchoGPT achieved significant improvements over baseline models across all metrics: BLEU (45.9 ± 18.9, $P < 0.0001$) suggesting moderate similarity to reference texts, METEOR (62.4 ± 18.0, $P < 0.0001$) indicating good capture of meaning and structure, ROUGE-L (55.7 ± 17.8, $P < 0.0001$) demonstrating moderate to good summarization quality, BERT Score (91.6 ± 3.0, $P < 0.0001$) revealing high semantic alignment with human-written reports, and RadGraph-F1[9] (47.7 ± 14.9, $P < 0.0001$), a medical specific NLP metric, revealing moderate success in accurately extracting and identifying medical entities from the original reports.

EchoGPT was compared against human expert evaluation. A panel of cardiologists evaluated the generated summaries of 30 randomly selected cases based on four criteria (4C): completeness, measuring whether all relevant details were included; conciseness, evaluating clarity and brevity; correctness, assessing accuracy; and clinical utility, determining usefulness for clinical decision-making. Scores range from 0 to 1, with higher scores indicating better performance. The results were as follows: correctness (0.17) suggesting slight agreement, conciseness (0.22), clinical utility (0.34) indicating fair agreement, and completeness (0.49) revealing moderate agreement. Additionally, cardiologist compared original and generated reports on the 4C metrics with scores ranging from −10 to 10, where positive values indicate EchoGPT outperformed human experts. EchoGPT significantly outperformed human experts on conciseness (1.67 ± 5.40, $P < 0.001$), and showed no significant differences in the other criteria.

To assess the sensitivity of automatic metrics to clinically significant errors, the researchers replaced actual measurements in the reports with random numbers. Surprisingly, this manipulation, which rendered the reports medically meaningless, only slightly lowered the automatic metric scores ($P < 0.0001$). This demonstrates that common NLP metrics are inadequate for assessing the clinical relevance and potential impact of errors in medical reporting.

Discussion

Chao *et al*. have made a significant contribution to the field of AI in cardiology by developing and evaluating EchoGPT, an LLM fine-tuned for

The opinions expressed in this article are not necessarily those of the Editors of the *European Heart Journal – Digital Health* or of the European Society of Cardiology.

* Corresponding author. Tel: 514-376-3330, Ext. 7825, Email: robert.avram.md@gmail.com

TTE report summarization. Their study demonstrates that EchoGPT not only outperforms other open-source LLMs in this task but also produces reports comparable to those written by cardiologists in terms of completeness, correctness, conciseness, and clinical utility. Notably, EchoGPT's outputs were significantly preferred for their conciseness, suggesting potential improvements in reporting efficiency. Furthermore, the authors are to be commended for their commitment to open science and research reproducibility. While patient privacy concerns prevent the release of the primary Mayo Clinic dataset, the publicly available MIMIC-EchoNotes[10] dataset provides a valuable resource for other researchers to conduct comparative studies and validate findings. Moreover, by releasing a checkpoint of their models, and statistical analysis code on GitHub (https://github.com/chiehjuchao/EchoGPT.git), the authors have set a commendable example of transparency in AI research. This approach not only enhances the credibility of their work but also facilitates further advancements in the field by enabling other researchers to build upon and potentially improve their model. Such openness is crucial for the responsible development and evaluation of AI in medical imaging, striking a balance between data privacy and scientific collaboration.[11]

Although EchoGPT has demonstrated promising results in summarizing echocardiography reports, this study highlights the need for more robust and scalable evaluation methods tailored specifically for medical AI applications. The medical-specific metric RadGraph-F1[9] proved more suitable than traditional NLP metrics as it was the only automatic metric that exhibited moderate correlation with human expert evaluation. Key areas for improvement include enhancing factual accuracy, minimizing hallucinations, improving generalizability across diverse healthcare settings, integrating it with current echocardiography reporting tools, and validating its impact. Additionally, integrating patient history and other relevant clinical data could result in more contextualized and clinically meaningful reports. These advancements, along with rigorous clinical validation and continuous human expert oversight, have the potential to revolutionize echocardiography reporting. However, it is crucial that these developments prioritize patient safety, clinical accuracy, and ethical considerations in AI-assisted medical decision-making.

Funding

Robert Avram is supported by the 'Fonds de la recherche en santé du Québec' (Grant 312758), by the Canadian Institute for Advanced Research (CIFAR), the Montreal Heart Institute Research Centre, the Montreal Heart Institute Foundation and by the Des Groseillers-Bérard Research Chair.

Conflict of interest: none declared.

References

1. Leeson P. Predicting the future with echocardiography: looking outside the heart? *Eur J Prev Cardiol* 2017;**24**:1515–1516.
2. Pelliccia A, Caselli S, Sharma S, Basso C, Bax JJ, Corrado D, *et al.* European Association of Preventive Cardiology (EAPC) and European Association of Cardiovascular Imaging (EACVI) joint position statement: recommendations for the indication and interpretation of cardiovascular imaging in the evaluation of the athlete's heart. *Eur Heart J* 2018;**39**:1949–1969.
3. McAlister NH, McAlister NK, Buttoo K. Understanding cardiac "echo" reports. Practical guide for referring physicians. *Can Fam Physician* 2006;**52**:869–874.
4. Freitas D, Alner S, Demetrescu C, Antonacci G, Proudlove N. Time to be more efficient: reducing wasted transthoracic echocardiography (TTE) diagnostic appointment slots at Guy's and St Thomas' NHS Trust. *BMJ Open Qual* 2023;**12**:e002317.
5. Chao C-J, Banerjee I, Arsanjani R, Ayoub C, Tseng A, Delbrouck J-B, *et al.* Evaluating large language models in Echocardiography reporting: opportunities and challenges. medRxiv 2024.01.18.24301503. https://doi.org/10.1101/2024.01.18.24301503.
6. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. Jul 19 2023. arXiv: arXiv:2307.09288. http://arxiv.org/abs/2307.09288
7. Tunstall L, Beeching E, Lambert N, Rajani N, Rasul K, Belkada Y, *et al.* Zephyr: Direct Distillation of LM Alignment. Oct 25 2023. arXiv: arXiv:2310.16944. http://arxiv.org/abs/2310.16944
8. Han T, Adams LC, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, *et al.* MedAlpaca—An Open-Source Collection of Medical Conversational AI Models and Training Data. Oct. 04, 2023. arXiv: arXiv:2304.08247. http://arxiv.org/abs/2304.08247
9. Jain S, Agrawal A, Saporta A, Truong SQH, Duong DN, Bui T, *et al.* RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. Aug. 29, 2021. arXiv: arXiv:2106.14463. http://arxiv.org/abs/2106.14463
10. Kwak GH, Moukheiber D, Moukheiber M, Moukheiber L, Moukheiber S, Butala N, *et al.* EchoNotes Structured Database derived from MIMIC-III (ECHO-NOTE2NUM). *PhysioNet*. 2024.
11. Engelhardt S. Why thorough open data descriptions matters more than ever in the age of AI: opportunities for cardiovascular research. *Eur Heart J Digit Health* 2024;**5**:507–508.