UCSF
UC San Francisco Previously Published Works

Title

ChatGPT Helped Me Write This Talk Title, but Can It Read an Echocardiogram?

Permalink

https://escholarship.org/uc/item/6j464766

Journal

Journal of the American Society of Echocardiography, 36(10)

Author

Arnaout, Rima

Publication Date

2023-10-01

DOI

10.1016/j.echo.2023.07.007

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Am Soc Echocardiogr. Author manuscript; available in PMC 2024 March 05.

Published in final edited form as:

J Am Soc Echocardiogr. 2023 October ; 36(10): 1021–1026. doi:10.1016/j.echo.2023.07.007.

ChatGPT helped me write this talk title, but can it read an echocardiogram?

Rima Arnaout, MD[a,*]

[a]Departments of Medicine, Radiology, and Pediatrics; Bakar Computational Health Sciences Institute; UCSF UC Berkeley Joint Program in Computational Precision Health, University of California, San Francisco, San Francisco, CA, USA

Keywords

machine learning; artificial intelligence; deep learning; echocardiography; foundation models; large language models; Feigenbaum Lecture

I'm not going to talk about something new today.

Multi-disciplinary collaboration and innovation in echocardiography is not new(1). Since the days of Edler and Hertz, clinicians and engineers have worked to develop ultrasound technology in service of patient care.

The push to scale and standardize echocardiography is also not new. Through collaborative guidelines and practice statements as well as accreditation and board certification programs, professional organizations such as the American Society of Echocardiography are striving to standardize and scale delivery of quality echocardiography around the world (Figure 1A). What *has* become more prominent in recent years is the idea that machine learning, a type of computer algorithm that learns patterns from data, can be used to further these goals(2–4). Ultimately, we hope that machine learning will help us achieve accurate, reproducible, expert ultrasound for every patient, in every hospital and clinic worldwide.

AI for medical imaging: progress and open questions

Over the past several years, AI algorithms have been tested echocardiography in myriad use cases for echocardiography, including detection of congenital heart disease, quantification of LVEF, and analysis of HFpEF and diastolic function, to name just a few(5–11). But like all science, the more we study machine learning, especially as applied to echocardiography and medical imaging, the more questions arise.

One open question is how to free machine learning from human labeling. Remember that in supervised machine learning, algorithms learn from the data and corresponding ground truth labels. These ground truth labels may be measurements traced from image or interpretations

*Corresponding Author.

Competing interests: None

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of a certain image or study. When relying on human-annotated labels for training or testing, however, we must remember there are limits to the accuracy and reproducibility of clinical measurements and interpretations, especially in busy, real-world clinical settings (Figure 1B–E). Training AI algorithms with sub-optimal labels runs the risk of teaching a model to re-create human error and variability, while evaluating AI algorithms against flawed labels makes determining their true performance difficult.

Another open question in AI for medical imaging is how to make AI models generalize better to datasets outside those upon which they were trained. While a trained clinician can interpret an imaging study no matter what hospital it came from, for example, current AI algorithms can struggle in this task(12,13), overfitting on image features that are specific to a certain dataset but don't port well to a new dataset because they are not clinically relevant (Figure 1F–I).

Rise of large foundational models

Given current shortcomings in AI, researchers have been hard at work. Consequently, this year has seen the rise of Large Language Models (LLMs)(14,15). As their name suggests, these AI algorithms are designed to model human language. Different LLMs can model language-to-language tasks (Figure 2A), or language-to-imaging tasks (Figure 2B).

In addition to modeling language tasks, some recent AI algorithms can even model image-to-image tasks. For example, the Segment Anything Model (SAM) aims to trace out shapes in any image, aided again by user prompts (Figure 2C)(16). Researchers are already attempting to use this model for medical imaging(17–19).

For simplicity, one can refer to all of the models described above as 'foundation models,' because whether they model language, imaging, or both, they aspire to be models upon which several different tasks can be accomplished.

While foundation models have existed for several years, two main features have improved their utility. First, improved user interfaces allow a user to input a text-based question or request—a user prompt—and receive a response from the model (Figure 2). Second, increased size of these models has translated to more realistic results, at least for many creative and/or non-medical tasks (Figure 3A). These advancements have caused a lot of excitement, investment, hope, and hype(20–22) (Figure 3B).

In this setting, it is natural to ask whether these newer foundation models may help address gaps and open questions about AI for medical imaging, and especially for echocardiography. With respect to freeing AI from human labeling, for example, researchers hope that fine-tuning a foundation model for a certain task may require less domain-specific, clinician-labeled data than if they had to train that task from scratch. With respect to generalization, some feel that improved performance will always correlated with model size. Some even feel that foundation models could learn to perform reasoning tasks(23), while others think that evidence of reasoning capabilities is spurious(24); assessing foundation models' true capabilities is currently an active area of study.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Therefore, while foundation models have certainly shown impressive performance so far on many tasks, applying these models to echocardiography and other high-stakes clinical tasks currently still requires caution and further research.

Caveats and failure modes of foundation models to date

Labeling and prompting.

One drawback of current foundation models is a continued need for time-consuming manual human input in the form of user prompts. Prompt engineering, as it is called, runs the risk of being variable and extremely laborious for a human user(25). In the course of preparing examples for this talk, each example required several text prompt attempts before generating an acceptable result. In several cases, no acceptable result was possible despite numerous prompts.

Examples of imaging prompts fared similarly. The Segment Anything Model may trace shapes from an image, but it still relies on user prompts both to achieve better results and to obtain semantic information (i.e., which of the many segmented shapes is relevant). A brief comparison between SAM and a completely human-label free approach designed for echocardiographic images(26) suggests that smaller, task-specific models still have a role in AI for medical imaging, especially when those required no manual labels (Figure 4). (In contrast, in addition to user prompts at the point of care, SAM also required over 1 billion masks during its training(16).)

Diversity.

Other concerns about current foundation models pertain to diversity. As with all AI models, foundation models represent any bias or error in the data they trained on. If that data lacks important types of diversity, so could the model itself, as well as any AI algorithms one may build from that foundation(27,28). While diversity in imaging datasets is important to measure(29), foundation models such as GPT-4 did not provide details on its dataset construction or algorithm(14).

In addition, exclusive reliance on a few, large, foundation models has the potential to affect the diversity of the AI research community. Training a very large model requires computing infrastructure and resources on a scale that few researchers have access to. Task-specific AI algorithms all built on the same handful of foundational models may reduce diversity of AI algorithm development. Furthermore, models such as ChatGPT, began as open-source but did not remain so(30). Fortunately, it appears that language models both large and small will continue to be part of the AI landscape, and several are arising from the open-source domain(31).

Performance for high-stakes tasks.

Current foundation models appear to excel in two areas: (i) tasks utilizing everyday text and images, of the type one presumes were abundant during training, and (ii) creative tasks for which a good result will be realistic and/or plausible but does not have to be measured against a specific ground truth (Figure 2B). Foundation models often generate

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

realistic-seeming results that are factually incorrect, irrelevant, or physically impossible (Figure 5). This behavior has been termed “hallucination,” and it is clearly problematic when using these AI algorithms for medicine(32).

Hallucination occurs because, as AI pioneer and Meta’s chief data scientist Yann LeCun says, they “have no idea of the underlying reality that language describes… Those systems generate text that sounds fine, grammatically, semantically, but they don’t really have some sort of objective other than just satisfying statistical consistency with the prompt.”(33) In fact, inaccurate and hallucinatory information coming from foundation models has caused thousands in the field to sign onto an open letter calling for a 6-month moratorium on developing them further(34), while others feel this is an overreaction.

Research, as well as spirited debate, regarding large foundation models will almost certainly continue as scientists strive to resolve outstanding problems and develop best practices for their use. With luck, foundation models will soon overcome current limitations and be ready to aid in medical imaging tasks.

We as clinicians need to be ready for that day. Whether or not they are actively involved in technical AI research, sonographers, echocardiographers, medical imaging companies, clinical IT staff, and hospital administrators all have roles to play in the responsible adoption of AI for clinical use, across imaging modality and at several stages in the imaging pipeline(35). These include readying imaging data formats and infrastructure for AI(36), understanding the potential and pitfalls of AI models, understanding how data is used for AI, advocating for patients as needed, and testing AI-enabled products(2).

After several prompt attempts, the title for this talk was inspired, but not actually generated, by ChatGPT. While foundation models have potential, they are not quite ready for clinical use. However, given the pace of AI research, that may change quickly—especially if, in the tradition of echocardiography, clinicians, computer scientists, and engineers come together to push this new technology forward.

Acknowledgements:

R.A. is funded by the National Institutes of Health, Department of Defense, Moore Foundation, and McGovern Foundation

References

1. Singh S, Goyal A. The origin of echocardiography: a tribute to Inge Edler. Tex Heart Inst J. 2007;34(4):431–8. [PubMed: 18172524]
2. Quer G, Arnaout R, Henne M, et al. Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review. Journal of the American College of Cardiology. 2021 Jan 26;77(3):300–13. [PubMed: 33478654]
3. Sengupta PP, Shrestha S, Berthon B, et al. Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): A Checklist: Reviewed by the American College of Cardiology Healthcare Innovation Council. JACC Cardiovasc Imaging. 2020 Sep;13(9):2017–35. [PubMed: 32912474]
4. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020 Sep;26(9):1320–4. [PubMed: 32908275]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Arnaout R, Curran L, Zhao Y, et al. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nat Med. 2021 May;27(5):882–91. [PubMed: 33990806]
6. Athalye C, Nisselrooij A van, Rizvi S, et al. Deep learning model for prenatal congenital heart disease (CHD) screening generalizes to the community setting and outperforms clinical detection [Internet]. medRxiv; 2023 [cited 2023 Mar 21]. p. 2023.03.10.23287134. Available from: 10.1101/2023.03.10.23287134v1
7. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature. 2020 Apr;580(7802):252–6. [PubMed: 32269341]
8. Hathaway QA, Yanamala N, Siva NK, et al. Ultrasonic Texture Features for Assessing Cardiac Remodeling and Dysfunction. J Am Coll Cardiol. 2022 Dec 6;80(23):2187–201. [PubMed: 36456049]
9. Chiou YA, Hung CL, Lin SF. AI-Assisted Echocardiographic Prescreening of Heart Failure With Preserved Ejection Fraction on the Basis of Intrabeat Dynamics. JACC: Cardiovascular Imaging. 2021 Nov;14(11):2091–104. [PubMed: 34147456]
10. Arnaout R. Can Machine Learning Help Simplify the Measurement of Diastolic Function in Echocardiography? JACC Cardiovasc Imaging. 2021 Nov;14(11):2105–6. [PubMed: 34274276]
11. Akerman A, Porumb M, Beqiri A, et al. COMPARISON OF CLINICAL ALGORITHMS AND ARTIFICIAL INTELLIGENCE APPLIED TO AN ECHOCARDIOGRAM TO CATEGORIZE RISK OF HEART FAILURE WITH PRESERVED EJECTION FRACTION (HFPEF). Journal of the American College of Cardiology. 2023 Mar 7;81(8, Supplement):360.
12. HealthITAnalyyics. HealthITAnalytics. 2018 [cited 2023 Jul 10]. Deep Learning for Medical Imaging Fares Poorly on External Data. Available from:https://healthitanalytics.com/news/deep-learning-for-medical-imaging-fares-poorly-on-external-data
13. Pfau J, Young AT, Wei ML, et al. Global Saliency: Aggregating Saliency Maps to Assess Dataset Artefact Bias [Internet]. arXiv; 2019 [cited 2023 Jul 10]. Available from: http://arxiv.org/abs/1910.07604
14. OpenAI. GPT-4 Technical Report [Internet]. arXiv; 2023 [cited 2023 Jul 10]. Available from: http://arxiv.org/abs/2303.08774
15. DALL·E: Creating images from text [Internet]. [cited 2023 Jul 10]. Available from: https://openai.com/research/dall-e
16. Kirillov A, Mintun E, Ravi N, et al. Segment Anything [Internet]. arXiv; 2023 [cited 2023 Jun 12]. Available from: http://arxiv.org/abs/2304.02643
17. Hu C, Xia T, Ju S, et al. When SAM Meets Medical Images: An Investigation of Segment Anything Model (SAM) on Multi-phase Liver Tumor Segmentation [Internet]. arXiv; 2023 [cited 2023 Jun 15]. Available from: http://arxiv.org/abs/2304.08506
18. Mazurowski MA, Dong H, Gu H, et al. Segment Anything Model for Medical Image Analysis: an Experimental Study [Internet]. arXiv; 2023 [cited 2023 Apr 21]. Available from: http://arxiv.org/abs/2304.10517
19. Huang Y, Yang X, Liu L, et al. Segment Anything Model for Medical Images? [Internet]. arXiv; 2023 [cited 2023 Jul 10]. Available from: http://arxiv.org/abs/2304.14660
20. ai Q. Forbes. [cited 2023 Jul 11]. Microsoft Confirms Its $10 Billion Investment Into ChatGPT, Changing How Microsoft Competes With Google, Apple And Other Tech Giants. Available from: https://www.forbes.com/sites/qai/2023/01/27/microsoft-confirms-its-10-billion-investment-into-chatgpt-changing-how-microsoft-competes-with-google-apple-and-other-tech-giants/
21. Hardian Health [Internet]. [cited 2023 Jun 15]. How to get ChatGPT regulatory approved as a medical device. Available from: https://www.hardianhealth.com/blog/how-to-get-regulatory-approval-for-medical-large-language-models
22. Felten E, Raj M, Seamans R. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. Strategic Management Journal. 2021;42(12):2195–217.
23. Kojima T, Gu SS, Reid M, et al. Large Language Models are Zero-Shot Reasoners [Internet]. arXiv; 2023 [cited 2023 Jul 11]. Available from: http://arxiv.org/abs/2205.11916

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

24. Schaeffer R, Miranda B, Koyejo S. Are Emergent Abilities of Large Language Models a Mirage? [Internet]. arXiv; 2023 [cited 2023 Jul 11]. Available from: http://arxiv.org/abs/2304.15004
25. AI Prompt Engineering Isn't the Future [Internet]. [cited 2023 Jul 11]. Available from: https://hbr.org/2023/06/ai-prompt-engineering-isnt-the-future
26. Ferreira DL, Salaymang Z, Arnaout R. Label-free segmentation from cardiac ultrasound using self-supervised learning [Internet]. arXiv; 2022 [cited 2022 Dec 30]. Available from: http://arxiv.org/abs/2210.04979
27. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models [Internet]. arXiv; 2022 [cited 2023 Jul 11]. Available from: http://arxiv.org/abs/2108.07258
28. Glocker B, Jones C, Bernhardt M, et al. Risk of Bias in Chest X-ray Foundation Models [Internet]. arXiv; 2022 [cited 2023 Jul 11]. Available from: http://arxiv.org/abs/2209.02965
29. Chinn E, Arora R, Arnaout R, et al. ENRICHing Medical Imaging Training Sets Enables More Efficient Machine Learning [Internet]. medRxiv; 2023 [cited 2023 Mar 21]. p. 2021.05.22.21257645. Available from: 10.1101/2021.05.22.21257645v3
30. https://github.com/factoidforrest. OpenAI Vendor Lock-in: The Ironic Story of How OpenAI Went from Open Source to "Open Your Wallet" | LunaTrace [Internet] 2023 [cited 2023 Jul 11]. Available from: https://www.lunasec.io/docs/blog/openai-not-so-open/
31. Patel D. Google "We Have No Moat, And Neither Does OpenAI" [Internet]. [cited 2023 Jul 11] . Available from: https://www.semianalysis.com/p/google-we-have-no-moat-and-neither
32. Simonite T AI Has a Hallucination Problem That's Proving Tough to Fix. Wired [Internet]. [cited 2023 Jul 11]; Available from: https://www.wired.com/story/ai-has-a-hallucination-problem-thats-proving-tough-to-fix/
33. Eye On A.I.: Yann LeCun: Filling the Gap in Large Language Models [Internet]. [cited 2023 Jul 12]. Available from: https://aneyeonai.libsyn.com/yann-lecun
34. Pause Giant AI Experiments: An Open Letter [Internet]. Future of Life Institute. [cited 2023 Jun 15]. Available from: https://futureoflife.org/open-letter/pause-giant-ai-experiments/
35. Dey D, Arnaout R, Antani S, et al. Proceedings of the NHLBI Workshop on Artificial Intelligence in Cardiovascular Imaging. JACC: Cardiovascular Imaging.
36. Arnaout R, Hahn RT, Hung JW, et al. The (Heart and) Soul of a Human Creation: Designing Echocardiography for the Big Data Age. J Am Soc Echocardiogr. 2023 May 16;S0894-7317(23)00208-0.
37. Introducing Segment Anything: Working toward the first foundation model for image segmentation [Internet]. [cited 2023 Jul 12]. Available from: https://ai.facebook.com/blog/segment-anything-foundation-model-image-segmentation/
38. Johri AM, Picard MH, Newell J, et al. Can a teaching intervention reduce interobserver variability in LVEF assessment: a quality control exercise in the echocardiography lab. JACC Cardiovasc Imaging. 2011 Aug;4(8):821–9. [PubMed: 21835373]
39. Thavendiranathan P, Popović ZB, Flamm SD, et al. Improved interobserver variability and accuracy of echocardiographic visual left ventricular ejection fraction assessment through a self-directed learning program using cardiac magnetic resonance images. J Am Soc Echocardiogr. 2013 Nov;26(11):1267–73. [PubMed: 23993695]
40. Kopecna D, Briongos S, Castillo H, et al. Interobserver reliability of echocardiography for prognostication of normotensive patients with pulmonary embolism. Cardiovascular Ultrasound. 2014 Aug 4;12(1):29. [PubMed: 25092465]
41. Leischik R, Dworrak B, Hensel K. Intraobserver and interobserver reproducibility for radial, circumferential and longitudinal strain echocardiography. Open Cardiovasc Med J. 2014;8:102–9. [PubMed: 25356089]
42. Cole GD, Dhutia NM, Shun-Shin MJ, et al. Defining the real-world reproducibility of visual grading of left ventricular function and visual estimation of left ventricular ejection fraction: impact of image quality, experience and accreditation. Int J Cardiovasc Imaging. 2015 Oct;31(7):1303–14. [PubMed: 26141526]
43. Simon J. Large Language Models: A New Moore's Law? [Internet]. Medium. 2021 [cited 2023 Jul 12]. Available from: https://julsimon.medium.com/large-language-models-a-new-moores-law-66623de5631b

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

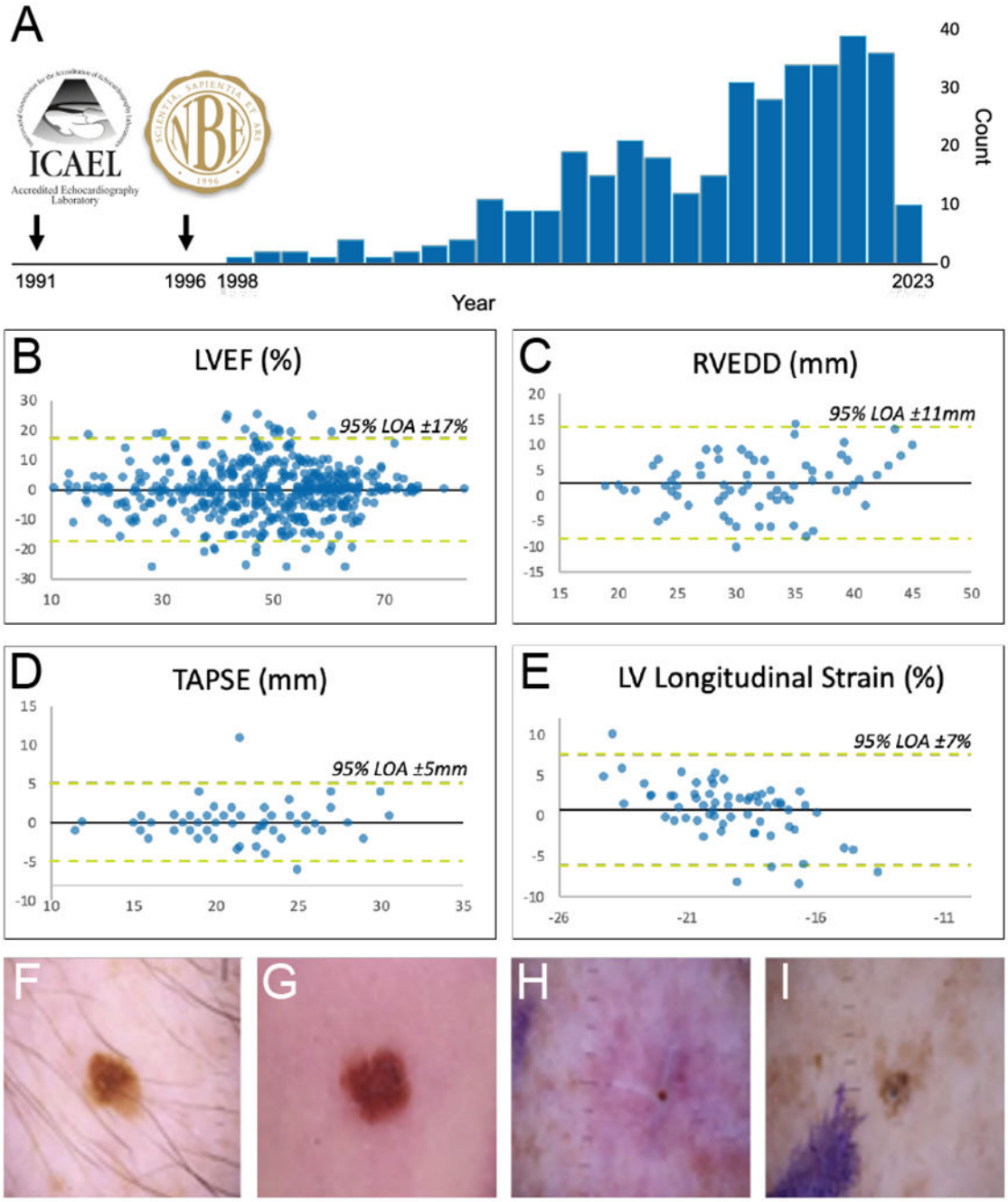

Figure 1.

(A) A histogram of guidelines documents per year produced by the American Society of Echocardiography and partner institutions. Arrows depict the start of the Intersocietal Accreditation Commission for Echocardiography and the American National Board of Echocardiography. (B)-(E) Bland-Altman limits of agreement on several common echocardiography measurements, demonstrates that there is still significant measurement variability. Adapted with permission from . Examples of benign skin lesions (F)-(G) compared to malignant ones (H)-(I), where a model attempting to diagnose a malignant

lesion may overfit on image features such as ink marks, rulers, and scar rather than on clinically relevant features. (Images from DuckDuckGo search for images free to share, modify, and use.)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Figure 2.

(A) Example text prompt and text response from ChatGPT large language model. (B) Example text prompt and image response from DALL-E model, via Microsoft Bing. (C) Example image with segmentation results from Meta's (formerly Facebook) Segment Anything Model, adapted from (37–42). The white cursor represents a user who may prompt the model by choosing an automatic segment and/or indicating an object in the image to receive segmentation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

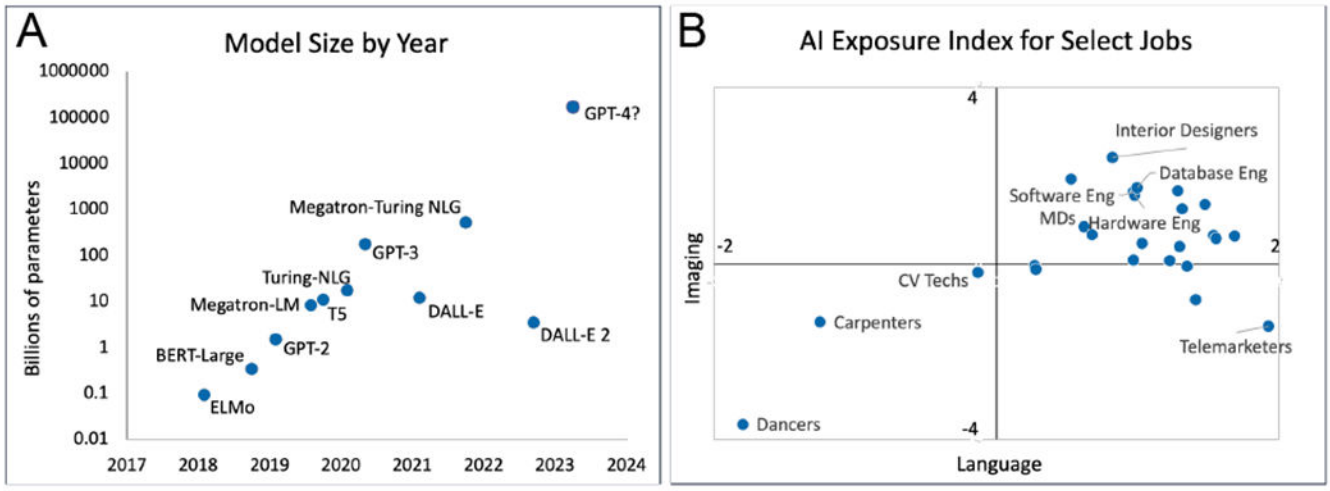

Figure 3.

(A) Model size by year for selected foundation models. Adapted with permission from (43). (B) Selected job categories along an index of how those jobs may be affected (positively or negatively) by AI imaging models (y-axis) and AI language models (x-axis). Adapted from (22).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

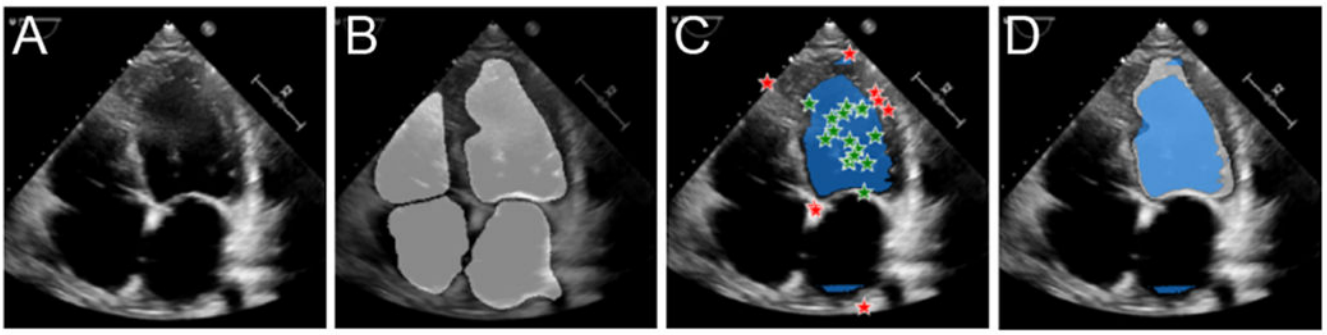

Figure 4.

(A) An example apical 4-chamber ultrasound image, with (B) all four chambers segmented by an ML model that required no user labels(26). (C) An attempt to use the Segment Anything Model (SAM) for ultrasound chamber segmentation required several user clicks (green and red stars) to segment the leq ventricle alone (D); this process would need to be repeated to get the rest of the chambers. (SAM without user prompts fared poorly, not shown.)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

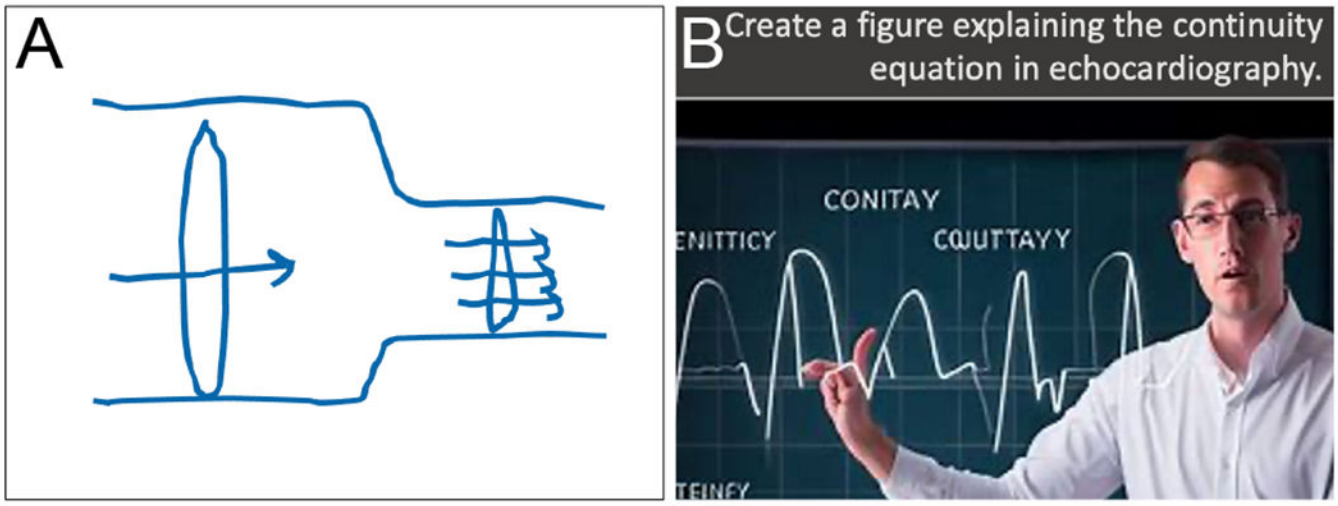

Figure 5.
(A) Human-drawn response to a Pictionary task, “draw the continuity equation,” holds physical and clinical meaning. (B) the same prompt to a foundation model (shown here, DALL-E via Microsoft Bing) produced an image that lacked meaning.