





DISCUSSION/COMMENTARY/PERSPECTIVES

The Role of Explainable AI and Evaluation Frameworks for Safe and Effective Integration of Large Language Models in Healthcare

Sandeep Reddy, MBBS, MSc, PhD¹; Alexandre Lebrun, MS, MSEE²; Adam Chee, PhD³; and Dimitrios Kalogeropoulos, PhD^{4,5}

¹Director, Deakin School of Medicine, Victoria, Australia; ²Co-Founder and CEO, Nabla, Paris, France; ³Associate Professor, Saw Swee Hock School of Public Health, National University of Singapore, Singapore; ⁴Founder and Chief Executive, Global Health & Digital Innovation Foundation, London, UK; ⁵UCL Global Business School for Health, London, UK

Corresponding Author: Sandeep Reddy; Email: sandeep.reddy@deakin.edu.au

DOI: <https://doi.org/10.30953/thmt.v9.485>

Keywords: AI, artificial intelligence, Converge2Xcelerate, healthcare, large language models, LLM

Submitted: February 8, 2024; Accepted: April 2, 2024; Published: April 30, 2024

The integration of artificial intelligence (AI), specifically large language models (LLMs: a specialized type of AI trained on vast amounts of text to understand existing content and generate original content) into healthcare continues to accelerate, necessitating thoughtful evaluation and oversight to ensure safe, ethical, and effective deployment. During a panel discussion at the 2023 Converge2Xcelerate event, which was organized by this journal's publisher, Partners in Digital Health, experts in AI shared salient perspectives on opportunities and challenges in thoughtfully translating innovations like LLM into healthcare.¹ A summary of key perspectives from the panel conversation is presented here.

Key topics covered:

- The potential of explainable AI to facilitate transparency and trust;
- Challenges in aligning AI with variable global healthcare protocols;
- The importance of evaluation via translational and governance frameworks tailored to healthcare contexts;
- Skepticism around overly expansive uses of LLMs for conversational interfaces;
- The need to judiciously validate LLMs, considering risk levels.

In addition, the discussion highlighted explainability (i.e., the concept that a machine learning model and its

output can be explained and “makes sense” to a human being at an acceptable level), evaluation, and careful deliberation with healthcare professionals as pivotal to realizing benefits while proactively addressing risks of larger AI adoption in medicine. Other discussion themes centered on critical evaluation practices for AI in medicine, the necessity and limitations of explainable AI, and deliberations healthcare leaders must undertake before deployment.

The adoption of AI techniques, especially LLMs like GPT-4 (Generative Pre-trained Transformer 4), into healthcare administration and clinical practice is accelerating rapidly.² While holding the promise to augment human capabilities and improve access, quality, and efficiency, AI integration introduces complex technical, ethical, and regulatory considerations regarding transparency, in addition to accountability and impact on patients and healthcare professionals.^{3,4}

Explainable AI: Building Trust and Transparency

A focal point of conversation was the role of explainable AI techniques in establishing confidence in AI systems. Participants concurred that while LLMs can provide useful outputs, they inherently function as “black boxes,” obscuring the reasoning behind conclusions. This opacity becomes concerning in high-stakes healthcare contexts.

The panelists emphasized explainable AI as an active area of research to address these transparency issues.

It facilitates accurate and repeatable results while clarifying the connections between inputs and outputs.⁵ However, realizing comprehensive explainability with large neural networks remains challenging. The complex, multivariate, and dynamic nature of clinical environments may further confound explanation approaches tuned to more constrained settings.⁶

Evaluation: Ensuring Effective and Ethical Translation

These observations underscore the necessity of rigorous evaluation tailored to healthcare applications throughout AI system design, deployment, and operation. The panelists advocated expansive, continuous evaluation frameworks, such as TEHAI (Translational Evaluation of Healthcare AI),⁷ spanning integration with healthcare IT systems, clinical adoption factors, updated performance monitoring, and governance components,⁸ like accountability and ethics. The panelists also noted distinctions from narrower regulatory assessments of safety and harm alone. Multidimensional evaluation can illuminate strengths, weaknesses, and appropriate use cases to guide AI adoption responsibly.

Automating Care Processes, Not Conversations

An additional consideration raised during the panel discussion was the increasing propensity to anthropomorphize AI systems, especially conversational interfaces. The panelists cautioned that excessive personification leads to impractical expectations around completeness and perfection. Yes, LLMs can help automate key care processes, but currently, fully automated diagnosis or treatment remains unreliable.

Furthermore, the panelists suggested focusing automation on augmenting clinicians with documentation, coding, and care coordination support. Artificial intelligence explanations serve to build appropriate mental models for its capabilities and limitations more than humanizing it.⁹

Progress Via Prudence

In summary, the panel participants highlighted cautious, collaborative,⁸ and evaluative approaches as imperative to productively applying AI advances in clinical practice. Rather than replacement, emphasis lays on starting with supportive, lower-risk roles and transparently assessing benefits and shortcomings revealed through real-world deployment.⁷

Continued progress relies on open and honest appraisals by cross-disciplinary leaders to find the most constructive niches for emerging innovations like LLMs in actual care delivery. Explainable evaluation frameworks facilitating transparency can play integral roles in advancing AI integration responsibly.

Funding

No external funding was utilized in the preparation of the manuscript.

Financial and non-Financial Relationships and Activities

Sandeep Reddy chairs and holds directorship with Healea Services and the Centre for Advancement of Translational AI in Medicine.

Contributors

Sandeep Reddy was responsible for drafting and finalizing the manuscript. All authors contributed to the manuscript's content.

Data Availability Statement (DAS), Data Sharing, Reproducibility, and Data Repositories

No original data were used in the preparation of this article.

Application of AI-Generated Text or Related Technology

AI technology was used to generate the transcript of the panel discussion and assist with grammar in preparing the manuscript.

Acknowledgments

The authors acknowledge Partners in Digital Health for hosting the event and panel discussion.

References

1. Telehealth and Medicine Today. ConV2X symposium [Internet]. [cited 2024 Feb 1]. Available from: <https://telehealth.conv2xsymposium.com>
2. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial Intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health*. 2018;3:e000798. <https://doi.org/10.1136/bmjgh-2018-000798>
3. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. 2018;378:981–3. <https://doi.org/10.1056/NEJMp1714229>
4. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17. <https://doi.org/10.1186/s12916-019-1426-2>
5. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–60. <https://doi.org/10.1109/ACCESS.2018.2870052>
6. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1:206–15. <https://doi.org/10.1038/s42256-019-0048-x>
7. Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform*. 2021;28(1):e100444. <https://doi.org/10.1136/bmjhci-2021-100444>

8. Kalogeropoulos D, Barach P. Telehealth's role enabling sustainable innovation and circular economies in health. *Telehealth Med Today*. 2023;8(1). <https://doi.org/10.30953/thmt.v8.409>
9. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1–38. <https://doi.org/10.1016/j.artint.2018.07.007>

Copyright Ownership: This is an open-access article distributed in accordance with the Creative Commons Attribution Non-Commercial (CC BY-NC 4.0) license, which permits others to distribute, adapt, enhance this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, and the use is non-commercial. See <http://creativecommons.org/licenses/by-nc/4.0>.