Using conversant artificial intelligence to improve diagnostic reasoning: ready for prime time?

Ian A Scott[1,2]

Tim Miller[1]

Carmel Crock[3]

1 University of Queensland, Brisbane, QLD.

2 Princess Alexandra Hospital, Brisbane, QLD.

3 Royal Victorian Eye and Ear Hospital, Melbourne, VIC.

i.scott@uq.edu.au

doi: 10.5694/mja2.52401

An estimated 10–15% of diagnoses are incorrect and serious patient harm or death from misdiagnosis affects one in 200 patients admitted to hospital.[1] Up to 80% of diagnostic errors are potentially preventable and are mostly due to faults in clinician reasoning related to the gathering of relevant background information and integrating symptoms, signs and situational factors in generating an appropriate differential diagnosis.[2]

Experience with digital symptom checkers,[3] electronic differential diagnosis generators[4] and electronic medical record (EMR) screening for missed diagnoses[5] has shown minimal impact, in part due to poor integration into clinical workflows and negative clinician perceptions.[6] In this perspective article, we consider how artificial intelligence (AI) may assist clinicians in diagnosing complex cases at the bedside or in the clinic.

Advent of AI-assisted diagnosis

Machine learning prediction models applied to imaging data have shown promise in diagnosing pneumothoraces from chest radiographs,[7] diabetic retinopathy from fundal images[8] or skin cancer from dermatoscopic photographs.[9] Randomised trials confirm superior AI-assisted clinician performance in diagnosing diabetic retinopathy,[10] detecting adenomas on colonoscopy,[11] and identifying impaired cardiac function from electrocardiographs.[12]

To date, most diagnostic machine learning models input images or structured data from EMRs or investigations and generate single disease probabilities or disease present/not present predictions. Moving upstream and using machine learning tools to assist bedside clinicians in more complex reasoning tasks requires integration of relevant clinical information (history from medical records, presenting complaint and findings from physical examination) and formulation of a differential diagnosis containing the correct diagnosis.

To achieve this aim, AI tools should work in ways that align with how clinicians reason in terms of System 1 (intuitive) and System 2 (analytical) thinking (two modes of cognitive processing introduced by Daniel Kahneman in *Thinking, fast and slow*).[13] For common clinical presentations, such as crushing central chest pain or sudden onset hemiplegia, intuitive reasoning often suffices in arriving quickly at the correct diagnosis. For more complex and undifferentiated cases, such as fever, weight loss and generalised bruising in an older patient, analytical reasoning is required. This is where AI may help clinicians generate and reason through a differential diagnosis, detailing the key pros and cons of each diagnosis from the clinician or AI.

Using ChatGPT and related technologies to assist with diagnostic reasoning

Large language models (LLMs), such as the general purpose generative pretrained transformer (GPT) series of models, embodied in the chatbot ChatGPT, use natural language processing to learn and generate human-like text content in response to text-based prompts (Box). Studies of LLM-assisted diagnostic reasoning have used GPT-3.5 or GPT-4. Applied to EMRs and other source documents, these LLMs can generate concise summaries of patients' active diagnoses and past medical history (thus saving interview time and effort),[14,15] suggest differential diagnoses surpassing previous differential diagnosis generators,[16,17] detect diagnostic uncertainty in clinical documentation[18] and solve complex diagnostic problems.[19,20] Furthermore, LLMs can perform multistep reasoning and provide rationales for the links between each step, known as chain-of-thought reasoning.[21] LLMs can thereby function as conversant sounding boards against which clinicians iteratively test their diagnostic reasoning.[22] Simply providing a list of differential diagnosis, with no rationales or probability rankings, has no effect on clinician diagnostic accuracy.[23]

Traditional task-specific machine learning prediction models generate a single set of diagnostic predictions in response to a fixed, one-off input of pre-processed data with predictions pre-analysed by domain experts. These predictions rely on extracting and presenting, as explanation, key features learnt from being trained on a circumscribed, domain-specific dataset.[24] In contrast, LLMs are pretrained and fine-tuned on a large, heterogenous dataset of clinical knowledge, and can discern complex relationships and variations within the data, beyond the limits of human cognition. In response to text prompts relating to a diagnostic case, LLMs can generate a list of plausible alternatives, and, when provided with further information (eg, revised history or physical signs, clinician insight, simple bedside test results), they can re-evaluate and re-order their differential diagnosis. The more narrative text a LLM has as input, rather than only key clinical features, the better its diagnostic performance.[25] The generated diagnostic rationales, highlighting relevant patient data, provide a reasoning path towards the final diagnosis.[26] This "reasoning aware" approach, using chain-of-thought, prompt-based learning, allows the LLM to use rationales as part of its input, further improving the diagnostic outputs and even correcting its own misdiagnoses arising, in part, from inaccurate training data.[27]

Such evaluative LLMs could provide clinicians with a second opinion in real time, share uncertainty, deal with limited or noisy data, and defer appropriately to

Brief overview of how large language models (LLMs) work

- LLMs are sequence-to-sequence learners. They create output by answering a prompt. Given this sequence of words (eg, a clinical vignette), LLMs generate a highly likely sequence of words that would occur after this, based on the training data.
- LLMs are trained using machine learning techniques, typically based on huge corpora of text. ChatGPT, for example, is trained on a large amount of publicly available internet data.
- Unlike humans, LLMs have no abstract or causal models of the world, and they operate only when prompted.
- Given these points, LLMs can give drastically wrong answers, make up information, and give very different answers to the same query posed twice.

clinician expertise and judgement.[28] Studies show the accuracy of clinicians' diagnoses across multiple cases improves markedly if the diagnoses are discussed with one other colleague, more so with two or more.[29,30] The benefit of this collective intelligence can feasibly be replicated using LLMs. This "machine-in-the-loop" approach better leverages clinician expertise in hypothesis-driven decision making, mitigates over- and under-reliance on machine learning decision support, and builds clinician trust and control.[31] This contrasts with the more conventional "human-in-the-loop" approach where the role of the clinician is relegated to accepting or rejecting AI outputs that are unaccompanied by any reasoning chain, leading to clinician resistance to and disuse of LLMs.[22]

Experimental studies of LLMs in diagnostic reasoning

ChatGPT does not appear to significantly enhance the differential diagnosis of clinicians for common clinical presentations.[32-34] In contrast, in a study comparing GPT-4 with a simulated population of 10 000 online medical-journal-reading clinicians in solving 38 challenging cases, the March 2023 version of GPT-4 correctly diagnosed a mean of 22 cases (57%) versus 14 cases (36%) for the clinicians.[35] In a vignette study comparing GPT-4 with 553 clinicians, GPT-4 more accurately estimated pre-test probability of the disease in all five cases, and post-test probability in all cases after a negative test result, and in four cases after a positive test.[36] In a randomised study of 20 experienced clinicians diagnosing 302 difficult real-world cases, those assigned to assistance from Med-PaLM-2, an LLM trained on biomedical texts such as PubMed abstracts, compared with those assigned to more traditional decision support (search engines, online resources) generated higher quality differential diagnosis containing the correct diagnosis (top-10 accuracy, 52% *v* 44%) and demonstrated higher accuracy for the final diagnosis (59% *v* 34%).[37] In a randomised crossover simulation study, 20 standardised patients were subjected to text-based consultations with an LLM (Articulate Medical Intelligence Explorer) or face-to-face consultations with 20 primary care clinicians across 149 clinical scenarios, with responses assessed by 23 specialists.[38] The LLM showed significantly higher top-10 diagnostic accuracy than clinicians (93% *v* 83%). Both specialists and patients rated the LLM superior in communication, reasoning and empathy.

But LLMs have limitations. ChatGPT-3.5 demonstrated an 83% error rate when applied to 100 challenging paediatric cases, underscoring the need to avoid unrepresentative training datasets.[39] In another study, the differential diagnosis created by GPT-4 across 18 standardised clinical vignettes were more likely to include diagnoses that stereotyped certain races, ethnicities and genders.[40] ChatGPT is also often inaccurate when used by patients to self-diagnose and self-triage,[41,42] suggesting research should, for the moment, remain focused on clinician-facing applications.

Future directions

Several innovations will likely move LLM-assisted diagnosis towards prime-time use. Biomedically trained LLMs, such as Med-PaLM-2, augmented with real-time access to additional, up-to-date medical information, semantic knowledge graphs, reinforcement learning with human feedback and optimised prompt engineering, will develop accuracy superior to models such as GPT-4 trained on internet data of variable quality.[27,43] Multimodal LLMs are emerging that can process not only text but also numerical, image, video and audio data, further enhancing performance. For example, an LLM trained on both text and images (GPT-4 with Vision version), when compared with human respondents across 934 cases from the *New England Journal of Medicine* Image Challenge and 69 clinicopathological conferences (*New England Journal of Medicine*), achieved an overall diagnostic accuracy of 61% versus 49%, with longer, more informative captions increasing performance.[44] Another study found a diagnostic LLM trained on multimodal data from real-world EMRs outperformed text-based models.[45]

But there are challenges. Studies of diagnostic LLMs have involved laboratory-based vignettes that may not represent usual clinical practice where diagnoses unfold temporally with recursive question-answering interactions involving clinicians and patients. Variations in patient populations, clinical settings and data quality may degrade model performance. Embedding cognitive bias mitigations into the design of LLM applications and their user interfaces, and implementing LLMs in ways that blend with clinician workflows are required.[46] Randomised trials involving clinicians diagnosing acute clinical scenarios in real-world settings, with and without LLM assistance, are needed. Clinicians will still need to critically appraise the differential diagnosis and associated rationales in terms of their consistency with the clinical data, their correctness and level of relevant detail (ie, specificity), their usefulness in pointing towards the correct diagnosis, and their similarity to the way humans think.[26] Clinicians will also have to validate LLM performance on local datasets, use prompts with LLMs correctly,[47] and avoid over-reliance on model outputs. Regulatory approval and monitoring of LLM quality management systems will be required to preserve data privacy, ensure transparency and fairness, determine medical liability for harm and guarantee LLMs remain effective and safe over their life cycle.[48] The Therapeutic Goods Administration of Australia has stated that LLM developers must understand and demonstrate the sources and quality of text inputs used to train and test the model, in addition to showing how the data are relevant and appropriate for use on Australian populations.[49] Although LLM-assisted diagnosis is not yet ready for prime time use, it may not be far off.

Open access: Open access publishing facilitated by The University of Queensland, as part of the Wiley - The University of Queensland agreement via the Council of Australian University Librarians.

Competing interests: No relevant disclosures.

Provenance: Not commissioned; externally peer reviewed. ■

1 Scott IA, Crock C. Diagnostic error: incidence, impacts, causes and preventive strategies. *Med J Aust* 2020; 213: 302-305. https://www.mja.com.au/journal/2020/213/7/diagnostic-error-incidence-impacts-causes-and-preventive-strategies

2 Graber ML. The incidence of diagnostic error in medicine. *BMJ Qual Saf* 2013; 22 Suppl 2: ii21-ii27.

3 Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 2019; 9: e027743.

4 Riches N, Panagioti M, Alam R, et al. The effectiveness of electronic differential diagnosis (DDX) generators: a systematic review and meta-analysis. *PLoS One* 2016; 11: e0148991.

5 Mahajan P, Pai C-W, Cosby KS, et al. Identifying trigger concepts to screen emergency department visits for diagnostic errors. *Diagnosis (Berl)* 2020; 8: 340-346.

6 Scott IA. Using information technology to reduce diagnostic error – still a bridge too far? *Intern Med J* 2022; 52: 908-911.

7 Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: a retrospective study. *PloS Med* 2018; 15: e1002697.

8 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; 316: 2402-2410.

9 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542: 115-118.

10 Abramoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 2018; 1: 39.

11 Kamba S, Tamai N, Saitoh I, et al. Reducing adenoma miss rate of colonoscopy assisted by artificial intelligence: a multicenter randomized controlled trial. *J Gastroenterol* 2021; 56: 746-757.

12 Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nat Med* 2021; 27: 815-819.

13 Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Adv Health Sci Educ Theory Pract* 2009; 14 Suppl 1: 27-35.

14 Gao Y, Dligach D, Miller T, et al. Overview of the problem list summarization (ProbSum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes. *ArXiv* 2306.05270v1 [cs.CL]. https://arxiv.org/abs/2306.05270 (viewed June 2023).

15 Li C, Zhang Y, Weng Y, et al. Natural language processing applications for computer-aided diagnosis in oncology. *Diagnostics (Basel)* 2023; 13: 286.

16 Kottlors J, Bratke G, Rauen P, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology* 2023; 308: e231167.

17 Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator. *JFO Open Opthalmol* 2023; 1: 100005.

18 Marshall TL, Nickels LC, Brady PW, et al. Developing a machine learning model to detect diagnostic uncertainty in clinical documentation. *J Hosp Med* 2023; 18: 405-412.

19 Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023; 330: 78-79.

20 Mehnen L, Gruarin S, Vasileva M, Knapp B. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases [preprint]. *medRxiv* 2023.04.20.23288859; 27 Apr 2023. https://doi.org/10.1101/2023.04.20.23288859

21 Wu C-K, Chen W-L, Chen H-H. Large language models perform diagnostic reasoning. *ArXiv* 2307.08922v1 [cs.CL]. https://arxiv.org/abs/2307.08922 (viewed July 2023).

22 Miller T. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support. In: FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; Chicago (USA), Jun 12–15, 2023. New York, NY: Association for Computing Machinery, 2023; pp. 333-342. https://doi.org/10.1145/3593013.3594001

23 Harada Y, Katsukura S, Kawamura R, Shimizu T. Effects of a differential diagnosis list of artificial intelligence on differential diagnoses by physicians: an exploratory analysis of data from a randomized controlled study. *Int J Environ Res Public Health* 2021; 18: 1-8.

24 Caruccio L, Cirillo S, Polese G, et al. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Exp Syst Applic* 2024; 235: 121186.

25 Reese JT, Danis D, Caulfield JH, et al. On the limitations of large language models in clinical diagnosis [preprint]. *medRxiv* 2023.07.13.23292613; 14 Jul 2023. https://doi.org/10.1101/2023.07.13.23292613

26 Kwon T, Ong KT, Kang D, et al. Large language models are clinical reasoners: reasoning-aware diagnosis framework with prompt-generated rationales. *ArXiv* 2312.07399v1 [cs.CL]. https://arxiv.org/abs/2312.07399 (viewed Dec 2023).

27 Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023; 29: 1930-1940.

28 Dvijotham K, Winkens J, Barsbey M, et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. *Nat Med* 2023; 29: 1814-1820.

29 Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA Netw Open* 2019; 2: e190096.

30 Khoong EC, Nouri SS, Tuot DS, et al. Comparison of diagnostic recommendations from individual physicians versus the collective intelligence of multiple physicians in ambulatory cases referred for specialist consultation. *Med Decis Making* 2022; 42: 293-302.

31 Zhang S, Yu J, Xu X, et al. Rethinking human-AI collaboration in complex medical decision making: a case study in sepsis diagnosis. *ArXiv* 2309.12368v2 [cs.HC]. https://arxiv.org/abs/2309.12368 (viewed Feb 2024).

32 Hirosawa T, Harada Y, Yokose M, et al. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health* 2023; 20: 3378.

33 Harada Y, Katsukura S, Kawamura R, Shimizu T. Efficacy of artificial-intelligence-driven differential-diagnosis list on the

diagnostic accuracy of physicians: an open-label randomized controlled study. *Int J Environ Res Public Health* 2021; 18: 2086.

34 Berg HT, van Bakel B, van de Wouw L, et al. ChatGPT and generating a differential diagnosis early in an emergency department presentation. *Ann Emerg Med* 2024; 83: 83-86.

35 Eriksen AV, Moller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI* 2023; 1: https://doi.org/10.1056/AIp2300031

36 Rodman A, Buckley TA, Manrai AK, Morgan DJ. Artificial intelligence vs clinician performance in estimating probabilities of diagnoses before and after testing. *JAMA Netw Open* 2023; 6: e2347075.

37 McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. *ArXiv* 2312.00164v1 [cs.CY]. https://arxiv.org/abs/2312.00164 (viewed Nov 2023).

38 Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic AI. *ArXiv* 2401.05654v1 [cs.AI]. https://arxiv.org/abs/2401.05654 (viewed Jan 2024).

39 Barile J, Margolis A, Cason G, et al. Diagnostic accuracy of a large language model in pediatric case studies. *JAMA Pediatr* 2024; 178: 313-315.

40 Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health* 2024; 6: e12-22.

41 Fraser H, Crossland D, Bacher I, et al. Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. *JMIR Mhealth Uhealth* 2023; 11: e49995.

42 Levine D, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model [preprint]. *medRxiv* 2023.01.30.23285067; 1 Feb 2023. https://doi.org/10.1101/2023.01.30.23285067 (viewed July 2024).

43 Gao Y, Li R, Caskey J, et al. Leveraging a medical knowledge graph into large language models for diagnostic prediction. *ArXiv* 2308.14321v1 [cs.CL]. https://arxiv.org/abs/2308.14321 (viewed Aug 2023).

44 Buckley T, Diao JA, Rodman A, Manrai AK. Accuracy of a vision-language model on challenging medical cases. *ArXiv* 2311.05591v1 [cs.CL]. https://arxiv.org/abs/2311.05591 (viewed Nov 2023).

45 Niu S, Ma J, Bai L, et al. EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Inform Fusion* 2024; 102: 102069.

46 Sibbald M, Zwaan L, Yilmaz Y, Lal S. Incorporating artificial intelligence in medical diagnosis: a case for an invisible and (un) disruptive approach. *J Eval Clin Pract* 2024; 30: 3-8.

47 Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: Qualitative study of popular large language models. *JMIR Med Educ* 2024; 10: e51391.

48 Gilbert S, Harvey H, Melvin T, et al. Large language model AI chatbots require approval as medical devices. *Nat Med* 2023; 29: 2396-2398.

49 Therapeutic Goods Administration. Regulation of software based medical devices. Canberra: Department of Health and Aged Care, May 2024. https://www.tga.gov.au/how-we-regulate/manufacturing/medical-devices/manufacturer-guidance-specific-types-medical-devices/regulation-software-based-medical-devices (viewed Mar 2024). ■