COMMENTARY

Jiawen Deng, MS-2
Temerty Faculty of Medicine,
University of Toronto, Toronto, ON, Canada

Kiyan Heybati, MS-3, MSc(c)
Mayo Clinic Alix School of Medicine,
Jacksonville, FL

Ye-Jean Park, MS-2
Temerty Faculty of Medicine,
University of Toronto, Toronto, ON, Canada

Fangwen Zhou, MSc(c)
Faculty of Health Sciences and Faculty of Engineering, McMaster University,
Hamilton, ON, Canada

Anthony Bozzo, MD, MSc, FRCSC
Orthopedic Oncology, McGill University,
Montréal, QC, Canada

Artificial intelligence in clinical practice: A look at ChatGPT

In the rapidly evolving landscape of healthcare technologies, the integration of artificial intelligence in clinical practice is increasingly gaining attention. Recent advancements in large language models (LLMs), such as ChatGPT (Chat Generative Pre-trained Transformer), seem to herald a future where artificial intelligence-powered platforms will significantly enhance clinician workflow by reasoning through patient cases to provide differential diagnoses and treatment recommendations and by alleviating administrative burdens.

However, the prospect of using general-purpose LLMs like ChatGPT in clinical applications also raises several pertinent considerations. Can they deliver factual information with the accuracy and reliability required for patient care? Are they transparent enough to allow for the practice of evidence-based medicine? And how well do optimistic findings from published studies of LLMs translate to actual practice? Through this commentary, we aim to demystify the role of ChatGPT and similar technologies in clinical settings, highlighting their limitations and current applications and discussing future directions in research and development.

WHAT IS CHATGPT?

ChatGPT (https://chat.openai.com/) is a software platform developed by artificial intelligence research company OpenAI to produce conversational responses to user inputs. ChatGPT can process free-form prompts, which are inputs that do not follow a strict format or structure, much like a normal conversation with friends and colleagues. Since its popularization in 2022, ChatGPT has been tested and used across diverse domains, such as providing customer support, script editing, and computer coding, due to its ability to hold natural conversations and synthesize text. Recently, there has also been a growing interest in ChatGPT's potential applications in clinical settings, owing to its ability to answer medical questions and assess patient cases.

doi:10.3949/ccjm.91a.23070

HOW DOES IT WORK?

ChatGPT is a type of machine learning model. These models are programs that learn to associate specific patterns in data with specific outputs, similar to how clinicians learn to associate clusters of signs and symptoms with specific diagnoses during their training. Many currently trialed clinical machine learning models are designed to recognize patterns in numeric data, such as predicting deterioration of patients with COVID-19 using vital signs and laboratory values,[1] or to recognize patterns in images, such as identifying tumors on computed tomography or magnetic resonance imaging.[2] Unlike these clinical models, ChatGPT belongs to a category of machine learning models called LLMs, which are designed to recognize and predict patterns in text. When given an incomplete sentence, for example, LLMs can recognize the context of existing words in the sentence and then fill in the blank using its predictions, not unlike the autocorrect or autocomplete functions on phones.

ChatGPT was first trained to learn textual patterns using millions of sentences from a large collection of books and websites. During the training process, words were removed from the end of the sample sentences and the ChatGPT model was tasked with trying to predict the missing words, 1 word at a time, based on the available context. When the model predicts incorrectly, it tries to learn from its mistake to improve future predictions (**Figure 1A**). This process is reminiscent of how medical students hone their skills by predicting diagnoses using simulated patient cases, comparing their predictions with the answer and learning from the experience. In an additional step of the ChatGPT training process, human annotators review the model's

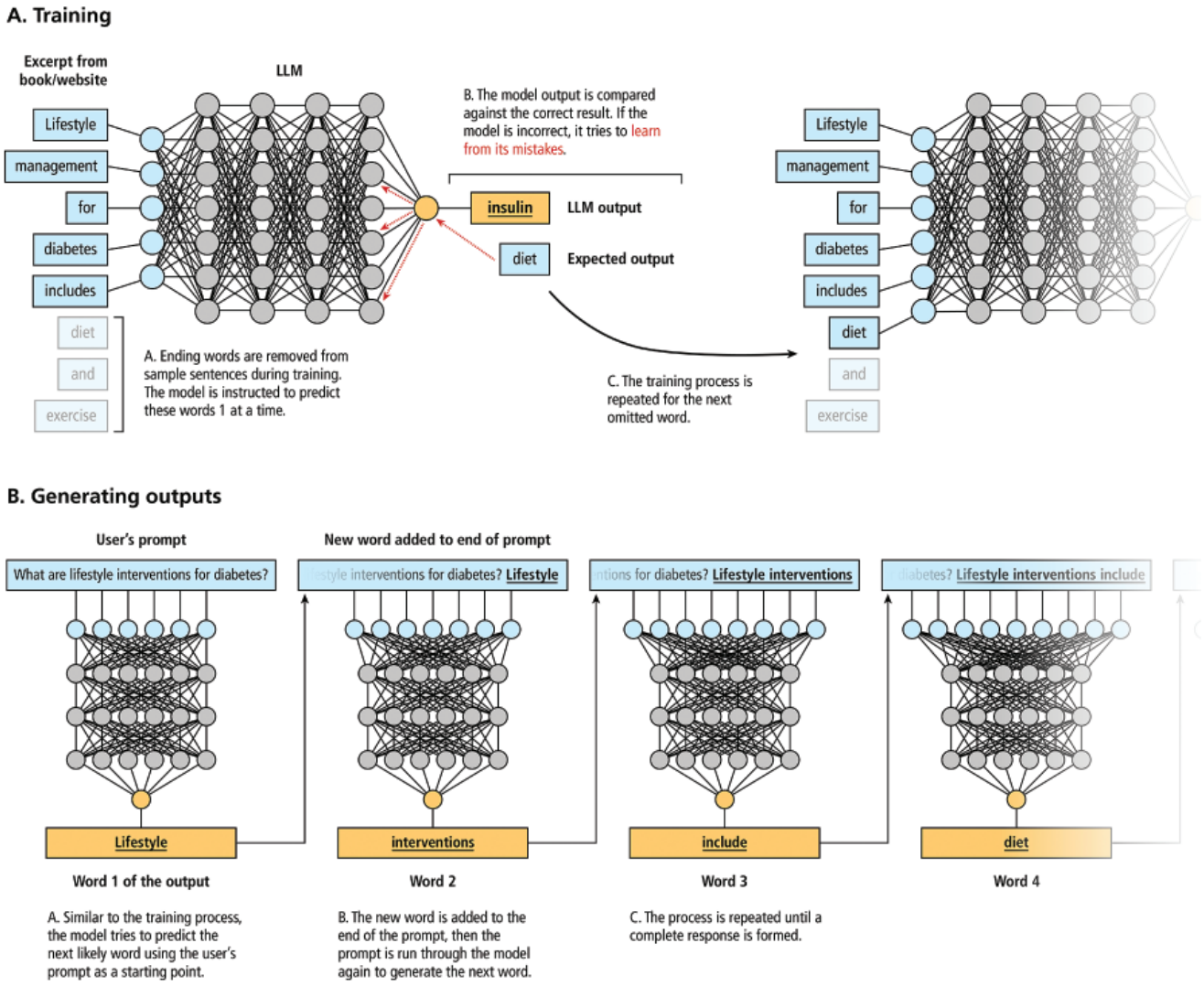

Figure 1. (A) Training and (B) output-generation processes of typical general-purpose large language models (LLMs) like ChatGPT.

output and provide feedback to guide the model to produce more conversational responses. When generating an output, ChatGPT uses the user's input as a starting point and repeatedly predicts which word is likely to come next, just as it did during its training, until a complete answer is formed (**Figure 1B**).[3]

While LLMs do not keep copies of the documents that they were trained on, they may retain knowledge and facts in the form of patterns they notice and learn during their training. For instance, if the LLM's training data contain sentences that include the keywords diet, exercise, and diabetes in close proximity, the model may learn to generate sentences that offer diet and exercise as interventions for patients with diabetes in its output. This characteristic leads to the belief that LLMs can encode medical knowledge, although it is widely debated whether LLMs actually understand what diabetes is, how it affects the body, and why certain diets and exercises are beneficial.[4]

Similarly, ChatGPT has no mechanism to learn from user inputs or feedback "on the fly." It does not improve itself incrementally. Instead, ChatGPT's developer periodically retrains the model from the ground up to incorporate some chat transcripts and user feedback. In these cases, the users serve a similar role as the human annotators described above, as they can rate thumbs up or thumbs down to ChatGPT responses (on the ChatGPT website) to guide the model to produce

more conversational and relevant passages. But once the training is finished and the model is released, the model will not change or improve itself until it undergoes a manual update again.

WHAT DOES THE EVIDENCE SAY?

Numerous recent studies have presented positive findings on the clinical utility of general-purpose LLMs such as ChatGPT. A well-known study by Kung et al[5] illustrated ChatGPT's ability to perform at or near the passing threshold of the United States Medical Licensing Examination. A study by Yeo et al[6] reported that ChatGPT could correctly answer questions relating to cirrhosis 79.1% of the time, while Rao et al[7] reported a 88.9% accuracy rate for its recommendations on breast cancer screening. Levine et al[8] concluded that GPT's ability to triage primary care case vignettes is close to that of physicians. More recently, the authors of a study that used ChatGPT to generate recommendations in response to clinical decision support system alerts described the tool's responses as offering "unique perspectives" while being "highly understandable and relevant."[9]

Do these results translate to real patients?

These results should be interpreted within the context of the limitations of the studies that produced them. For instance, current studies of ChatGPT have relied heavily on question banks and standardized case vignettes, which are easy to acquire but do not capture the complexity of real-life cases. In particular, questions from test banks such as the United States Medical Licensing Examination are usually based on common signs and symptoms, vetted for clarity, and written in a multiple-choice format. Thus, while these studies show that ChatGPT could recognize textbook descriptions of medical conditions and provide standard management recommendations, it is unclear how it would perform in actual clinical practice. After all, real patients present and describe their complaints variably, have different backgrounds and needs, and do not come with multiple-choice options.

An example of this limitation can be seen in the study by Rao et al[7] where ChatGPT was used to provide recommendations for breast cancer screening in those with breast pain. While breast pain is an uncommon symptom of breast cancer, an experienced clinician may recognize or reason that certain types of focal, persistent pain can be suggestive of malignancy, or ask questions about constitutional symptoms to further clarify the diagnosis. However, ChatGPT's responses varied from recommending unnecessary mammograms for diffuse and cyclical breast pain, to not recommending imaging for focal pain in high-risk populations. The accuracy of ChatGPT recommendations was only 58.3% when limited to cases involving breast pain, a large difference from the 88.9% accuracy it achieved on prompts without breast pain.[7] These findings illustrate the uncertainties surrounding ChatGPT's ability to analyze atypical or granular presenting symptoms, much like medical students who can score well on standardized tests but lack the clinical experience needed to deal with the complexity of actual patient presentations.

How useful are the responses?

In addition, it is unclear how clinically useful responses from ChatGPT are, even when they are technically correct. Liu et al,[9] for example, found that ChatGPT-generated recommendations were rated by expert human clinician reviewers as significantly less useful than human-generated recommendations. General-purpose LLMs, such as ChatGPT, may produce generic responses that are ambiguous or lack details, making it hard for clinicians to act on them. As there are currently no standardized methods for assessing the "usefulness" of LLM outputs, this aspect of their performance is often undertested.

Is ChatGPT more empathetic than physicians?

Another recent point of controversy regarding the clinical utility of ChatGPT was introduced by Ayers et al,[10] who showed that ChatGPT responses to patient inquiries were rated significantly higher for empathy than responses written by physicians. Results from the study were widely reported by news and social media outlets, giving the impression that ChatGPT may have better bedside manner than physicians.[11]

However, it is essential to consider the limitations of the study by Ayers et al,[10] the most prominent being that the physicians assessed were off-duty and were answering questions on Internet forums, which hardly reflects their clinical performance. At the same time, this study raises a fundamental question: can a text-based entity like ChatGPT truly provide empathetic care? The relationship between physicians and patients is multifaceted and built on trust, and relies on nonverbal cues, subtle signs, and rapport. LLMs, being restricted to text-based communication, inevitably have limitations in this regard.

Current discussions on empathy notwithstanding, it is also worth examining why perceived empathy seems lacking among healthcare workers. Administrative burden, which often leads to burnout and empathy fatigue, is a significant contributor. US physicians,

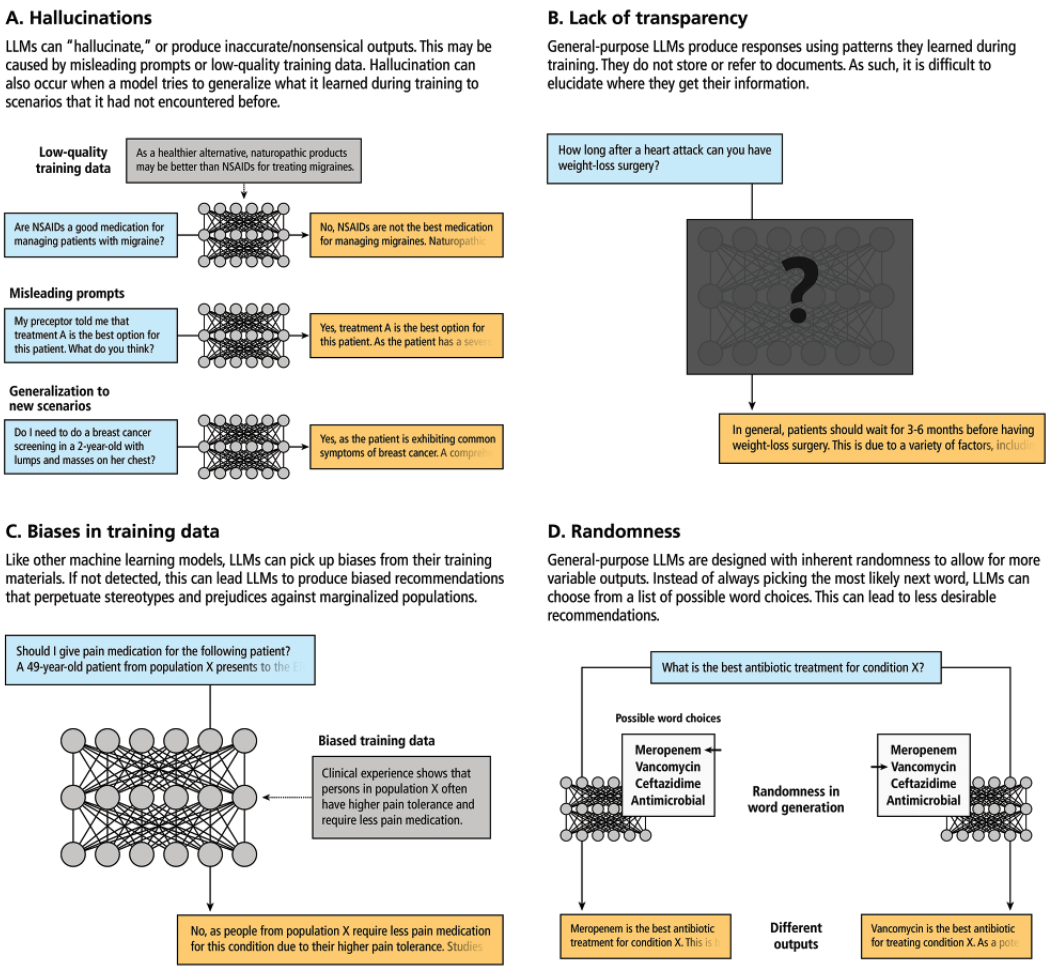

Figure 2. Common technical limitations of current general-purpose large language models (LLMs) like ChatGPT include (A) hallucinations, (B) lack of transparency, (C) biases in training data, and (D) randomness. Prompts and responses shown are for illustrative purposes only and do not represent actual output from LLMs.

for example, spend twice as much time on paperwork as they do with patients.[12] While LLMs might mimic empathy, their true value could lie in alleviating this administrative burden, potentially giving healthcare professionals more time for genuine patient interactions (as we discuss in later sections).

LIMITATIONS OF CURRENT LARGE LANGUAGE MODELS

In addition to limitations highlighted in studies assessing general-purpose LLMs such as ChatGPT, there are several technical limitations in the current design of these models.

Hallucinations

A key limitation is ChatGPT's tendency to "hallucinate," a phenomenon where the model generates factually incorrect or nonsensical outputs or fabricates information (**Figure 2A**). This behavior stems from ChatGPT's reliance on word patterns learned during training to generate responses, and reflects the fact that the model is not designed to function like a search engine or database. While this design allows ChatGPT to respond to scenarios that it had not encountered during training by generalizing its word associations—without adhering to a fixed knowledge set as Google or PubMed do—it can sometimes lead to inaccuracies. In Rao et al,[7] for instance, ChatGPT insisted on providing breast cancer screening for many cases where imaging would be futile or where the patient was at low risk, contrary to prevailing guidelines. It is likely that ChatGPT learned to associate keywords on breast cancer symptoms with screening recommendations,

which it generalized to all patient cases containing these keywords without considering other variables such as prognosis or risk factors.

Hallucination can also arise from training on low-quality or erroneous datasets, leading to incorrect word associations. ChatGPT does not evaluate the credibility of its sources during training, which compounds this problem. For example, in the study by Liu et al,[9] ChatGPT suggested using a nonexistent medication called "etanerfigut," an error that could have resulted from typos in the training material. Yeo et al[6] found that while ChatGPT correctly suggested using mean Model for End-Stage Liver Disease-Na scores for liver transplantation evaluation, it provided inaccurate cutoff values, which may be attributable to incorrect values in the original training dataset.

Additionally, because ChatGPT utilizes the user's inputs as a starting point to generate its response, it can be influenced by misleading prompts (eg, prompts that "hint" the desired answer to the LLM). An example can be seen in a study that assessed ChatGPT's ability to screen article abstracts for inclusion in clinical reviews.[13] After ChatGPT was given screening decisions from expert reviewers, it changed its answer to match the human decisions without trying to defend its original position. When prompted to explain the change, it gave nonspecific rationales (eg, "The study does not meet any of the inclusion criteria."). While technically not an example of hallucination, this study shows how if prompted with a preconceived notion, ChatGPT may contribute to confirmation bias rather than provide the correct information.

Lack of transparency

Given the tendency of LLMs to hallucinate, it is critical to verify and improve their responses by identifying the rationale behind their recommendations. However, the complexity of current general-purpose LLMs makes it difficult to elucidate how these models function. And because ChatGPT does not store or refer to documents from its training, it cannot provide references as other platforms such as UpToDate can (**Figure 2B**).

In fact, when asked to generate a list of citations, LLMs such as ChatGPT often "hallucinate" fake references that seem authentic at first glance. In a study assessing the accuracy of ChatGPT in providing clinical radiological information, only 124 references out of 343 references ChatGPT provided were real and accessible, and 47 references were actually relevant.[14] This demonstrates that ChatGPT has limited transparency and accountability, qualities that are often relied upon in the practice of evidence-based medicine.

Biases

It is well-documented that machine learning models can often produce results that are systemically prejudiced.[15] These biases are usually caused by biases in the models' training data. For instance, insurance models trained on data that associate lower healthcare costs with Black patients may allocate less care to Black patients.[15] The model analyzes the data at its face value, without considering the impact of socioeconomic status, unequal access to care, and other factors that lead to decreased costs in this population.[16] In a similar vein, ChatGPT can exhibit biases that are reflective of its training materials, which may include many unvalidated text sources from webpages with problematic characteristics (**Figure 2C**). ChatGPT had been shown to regurgitate many racial and sexist stereotypes that may harm marginalized communities and even affirm suicidal ideations.[17] If unmitigated, it is possible for ChatGPT to cause patient harm by producing biased recommendations.

Randomness

Lastly, general-purpose LLMs such as ChatGPT are designed to have inherent randomness in their outputs. As such, ChatGPT does not always choose the most likely next word when generating its responses, but rather selects from a list of possible options. As a result, running the same prompt through ChatGPT multiple times would likely yield different outputs (**Figure 2D**).

This design characteristic is useful for engaging users in a chatbot setting (ie, to make ChatGPT responses less predictable and more interesting) or for creative purposes such as brainstorming writing prompts. However, in medical practice, where there is often a limited number of optimal diagnoses or management strategies, this variability can lead to erroneous or less desirable outputs. For instance, in a study that assessed ChatGPT responses to questions about bariatric surgery, the model recommended waiting 6 to 12 months when asked the question, "How long after a heart attack can you have weight loss surgery?"[18] On a second run with the same prompt, however, ChatGPT recommended waiting 3 to 6 months. The 2014 American College of Cardiology/American Heart Association guidelines actually recommend waiting at least 2 months after an acute myocardial infarction before having major surgery,[19] making this a good example of hallucinations as well as the inherent randomness of ChatGPT. And because it is impossible to track down the source of the numbers in ChatGPT recommendations, this example also shows how its lack of transparency can be problematic.

TABLE 1
Resources for clinicians to learn more about large language models and machine learning research

Newsletters	
Doctor Penguin	doctorpenguin.com
NEJM AI Email Newsletter	store.nejm.org/signup/ai/newsletter
Podcasts	
Medicine and the Machine by Medscape	medscape.com/features/public/machine
NEJM AI Grand Rounds by NEJM Group	ai-podcast.nejm.org
The AI Health Podcast	podbay.fm/p/the-ai-health-podcast/about
Journals	
NEJM AI	ai.nejm.org
The Lancet Digital Health	thelancet.com/journals/landig/home
npj Digital Medicine	nature.com/npjdigitalmed/
Journal of Medical Internet Research and related journals	jmir.org

While some response variations only impact the syntax or structure of ChatGPT responses, recent studies have found that around 10% to 20% of prompts presented a second time would incur a substantial change to the content of the responses.[6,18] Not only is this unacceptable for clinical applications, but it also means that the performance and behavior observed in research studies of ChatGPT may not translate to actual practice.

A REALISTIC VIEW OF ChatGPT'S CLINICAL APPLICATIONS

Ultimately, general-purpose LLMs like ChatGPT are not designed for use in clinical settings. When used to provide factual information or reason through clinical cases, ChatGPT lacks the accuracy, reliability, and transparency needed for patient care. However, there are still ways for clinicians to make use of ChatGPT's ability to rapidly interpret and synthesize textual data.

One way to improve ChatGPT's performance is by providing it with the knowledge needed to answer the question in the user's prompt, an approach called "context injection." This works because ChatGPT generates responses using the user's input as a starting point, and thus can extract the needed information from the prompt rather than relying on its word associations, reducing the risk of hallucinations. An example of context injection is to provide ChatGPT with passages from the latest clinical practice guidelines or clinical trial publications, and then ask questions relating to the passages or ask ChatGPT to summarize the passages. This can make it easier for busy clinicians to stay up to date with the latest research or for journals to rapidly produce succinct summaries.

Other possible applications follow similar approaches of prompting LLMs with the information necessary for completing the requested task, such as using ChatGPT to quickly translate patient education materials to different reading levels or asking it to proofread email communications. Ali et al[20] used ChatGPT to rewrite surgical consent forms at a sixth-grade reading level and found that the model was able to preserve clinical details and increase clarity, as judged by expert subspecialty surgeon review. Lyu et al[21] used ChatGPT to translate radiology reports into plain language summaries for patients and determined that the reports were concise, clear, and comprehensive. Another study[22] used ChatGPT to summarize dictated transcripts of physician-patient encounters and found that it was able to produce high-quality notes in well-known formats. These preliminary investigations demonstrate that general-purpose LLMs can reduce the amount of time that healthcare professionals spend on documentation and other administrative duties, enabling them to spend more time with patients.[22]

It should be noted, however, that ChatGPT by default is not considered compliant with Health Insurance Portability and Accountability Act regulations. Thus, protected health information should not be entered into the platform. Compliant variations of the platform are available via enterprise solutions such as

CompliantChatGPT (https://compliantchatgpt.com) or BastionGPT (https://bastiongpt.com).

WHAT DOES THE FUTURE HOLD?

While many current research studies on the medical use of LLMs are directed toward ChatGPT, the clinical application of these general-purpose models is likely limited to the use cases we described here. After all, models like ChatGPT are designed to interpret and generate text across a wide range of topics and disciplines beyond medicine, with little to no consideration for consistency and transparency in its outputs. These characteristics make general-purpose LLMs poorly equipped for fulfilling clinical decision support roles.

However, several advancements are being made to tackle the technical limitations we identified, with the goal of developing LLM systems designed specifically for clinician use. BioGPT[23] and neuroGPT-X,[24] for instance, are LLMs trained on academic articles with the aim of reducing the risk of hallucinations. HippoAI (https://pendium.health/) and Glass AI (https://glass.health) are both clinician-focused LLM platforms that implement this concept, providing recommendations and diagnoses based on peer-reviewed clinical guidelines and medical databases. Platforms such as Perplexity.ai (https://www.perplexity.ai/) use LLMs to summarize results from search engines, allowing the platform to interact with users conversationally while remaining transparent by providing links to its sources. And a medicine-specific LLM from Google called Med-PaLM attempted to improve consistency in its answers by running a prompt through the model multiple times, surveying the results, and responding with the most commonly produced output.[25]

The field of clinical machine learning systems is evolving rapidly. **Table 1** lists some useful resources for clinicians to keep up to date with the latest advancements in LLMs.

With these developments, it is easy to imagine a future where specially designed LLMs power clinical decision support systems to provide clinicians with treatment recommendations, assist with differential diagnoses, and further integrate themselves into administrative roles. But for now, clinicians should exercise caution when interpreting optimistic results from studies involving general-purpose platforms like ChatGPT, and should remain cognizant of the limitations of ChatGPT.

DISCLOSURES

The authors report no relevant financial relationships which, in the context of their contributions, could be perceived as a potential conflict of interest.

Acknowledgments: *We thank Qi Kang Zuo, MS-1, and Kate Kim, MS-1, both affiliated with the UBC School of Medicine at the University of British Columbia, and Shubh Patel, MS-1, affiliated with the Temerty Faculty of Medicine at the University of Toronto, for providing internal reviews of this commentary.*

REFERENCES

1. **Campbell TW, Wilson MP, Roder H, et al.** Predicting prognosis in COVID-19 patients using machine learning and readily available clinical data. Int J Med Inform 2021; 155:104594. doi:10.1016/j.ijmedinf.2021.104594
2. **Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL.** Artificial intelligence in radiology. Nat Rev Cancer 2018; 18(8):500–510. doi:10.1038/s41568-018-0016-5
3. **Brown TB, Mann B, Ryder N, et al.** Language models are few-shot learners. Advances in Neural Information Processing Systems 33 (NeurIPS 2020); 33:1877–1901.
4. **Arcas BA.** Do large language models understand us? Daedalus 2022; 151(2):183–197. doi:10.1162/daed_a_01909
5. **Kung TH, Cheatham M, Medenilla A, et al.** Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2(2):e0000198. doi:10.1371/journal.pdig.0000198
6. **Yeo YH, Samaan JS, Ng WH, et al.** Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023; 29(3):721–732. doi:10.3350/cmh.2023.0089
7. **Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD.** Evaluating ChatGPT as an adjunct for radiologic decision-making. J Am Coll Radiol 2023; 20(10):990–997. doi:10.1016/j.jacr.2023.05.003
8. **Levine DM, Tuwani R, Kompa B, et al.** The diagnostic and triage accuracy of the GPT-3 Artificial Intelligence model. Preprint. medRxiv 2023; 2023.01.30.23285067. Published 2023 Feb 1. doi:10.1101/2023.01.30.23285067
9. **Liu S, Wright AP, Patterson BL, et al.** Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023; 30(7):1237–1245. doi:10.1093/jamia/ocad072
10. **Ayers JW, Poliak A, Dredze M, et al.** Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023; 183(6):589–596. doi:10.1001/jamainternmed.2023.1838
11. **McPhillips D.** CNN Health. ChatGPT may have better bedside manner than some doctors, but it lacks some expertise. https://www.cnn.com/2023/04/28/health/chatgpt-patient-advice-study-wellness/index.html. Accessed February 1, 2024.
12. **Herd P, Moynihan D.** Health care administrative burdens: centering patient experiences. Health Serv Res 2021; 56(5):751–754. doi:10.1111/1475-6773.13858
13. **Guo E, Gupta M, Deng J, et al.** Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res Published online September 8, 2023. doi:10.2196/48996
14. **Wagner MW, Ertl-Wagner BB.** Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information [published online ahead of print, 2023 Apr 20]. Can Assoc Radiol J 2023; 8465371231171125. doi:10.1177/08465371231171125
15. **Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G.** Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018; 178(11):1544–1547. doi:10.1001/jamainternmed.2018.3763
16. **Obermeyer Z, Powers B, Vogeli C, Mullainathan S.** Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019; 366(6464):447–453. doi:10.1126/science.aax2342

17. **Sezgin E, Sirrianni J, Linwood SL.** Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR Med Inform 2022; 10(2):e32875. doi:10.2196/32875
18. **Samaan JS, Yeo YH, Rajeev N, et al.** Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023; 33(6):1790–1796. doi:10.1007/s11695-023-06603-5
19. **Fleisher LA, Fleischmann KE, Auerbach AD, et al.** 2014 ACC/AHA guideline on perioperative cardiovascular evaluation and management of patients undergoing noncardiac surgery: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2014; 130(24):e278–e333. doi:10.1161/CIR.0000000000000106
20. **Ali R, Connolly ID, Tang OY, et al.** Bridging the literacy gap for surgical consents: an AI-human expert collaborative approach. Preprint. medRxiv 2023; 05.06.23289615. Published 2023 May 10. doi:10.1101/2023.05.06.23289615
21. **Lyu Q, Tan J, Zapadka ME, et al.** Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023; 6(1):9. doi:10.1186/s42492-023-00136-5
22. **Lee P, Bubeck S, Petro J.** Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023; 388(13):1233–1239. doi:10.1056/NEJMsr2214184
23. **Luo R, Sun L, Xia Y, et al.** BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022; 23(6):bbac409. doi:10.1093/bib/bbac409
24. **Guo E, Gupta M, Sinha S, et al.** neuroGPT-X: toward a clinic-ready large language model [published online ahead of print, 6 Oct 2023]. J Neurosurg 2023; 1–13. doi:10.3171/2023.7.JNS23573
25. **Singhal K, Azizi S, Tu T, et al.** Large language models encode clinical knowledge [published correction appears in Nature 2023]. Nature 2023; 620(7972):172–180. doi:10.1038/s41586-023-06291-2

Address: *Jiawen Deng, MS-2, Temerty Faculty of Medicine, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada; dengj35@mcmaster.ca*