An Opinion on ChatGPT in Health Care—Written by Humans Only

Jens Kleesiek[1], Yonghui Wu[2], Gregor Stiglic[3], Jan Egger[1], and Jiang Bian[2]

[1]Institute for AI in Medicine, University Medicine Essen, Essen, Germany; [2]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida; and [3]Faculty of Health Sciences, University of Maribor, Maribor, Slovenia

ChatGPT, created by OpenAI, has taken the world by storm, and its user base is growing even faster than the previous record held by TikTok, reaching 100 million users in just 2 mo after it launched. Textual context, presentations, and even source code are already being generated using ChatGPT. Many publications have been issued, and meanwhile, ChatGPT has been banned as an author by many publishing companies for several different reasons, such as plagiarism, incorrect information, or inaccurate information (*1,2*), whereas others argue its benefits, such as the ability to write more coherent sentences than nonnative speakers (*3*). But that does not stop people from all walks of health care from using it.

ChatGPT is powered by a generative pretrained transformer (GPT-3.5), which is a large language model (LLM) trained with 175 billion parameters (*4*). LLMs originate in natural language processing to formulate the probability distribution of a sequence of words or the next word in a sequence. Recent studies report that LLMs are foundation models in which a single model can be adapted to solve a wide range of different natural language-processing tasks because of few-shot learning, zero-shot learning, and transfer learning ability (*5*). The conversational artificial intelligence (AI) ability is achieved using LLM-based prompt learning (*6*). To alleviate the toxic responses and integrate human ethics, ChatGPT applied a strategy of reinforcement learning from human feedback to align LLMs to follow human instructions (*7*). These breakthroughs in natural language processing empower ChatGPT with conversational AI ability so good it has surprised the world. Even within OpenAI, ChatGPT has been a surprise. AI chatbots are not a new thing, but many previous attempts have not achieved the sensation that ChatGPT achieved. Meta's BlenderBot was a disappointment. What may be different for ChatGPT, beyond the unknown technologies, is OpenAI's goal of creating artificial general intelligence to match human-level intellect (*8*). ChatGPT certainly is not an artificial general intelligence, but it sure looks like one because of the breadth and depth of the knowledge it demonstrates through conversations.

Even though many are excited by its first use, disillusionment often sets in over time, for several reasons. On the one hand, ChatGPT gives wrong answers and is prone to confabulation ("a memory error defined as the production of fabricated, distorted, or misinterpreted memories about oneself or the world" (*9*)). This is exacerbated by the fact that we set different standards for communication among humans and between humans and computers. The belief is that a computer will not make mistakes. Moreover, many users' expectations are wrong, especially for medical interactions. The program was trained and designed for conversation, not diagnostic support or treatment recommendations. Yet, questions arise as to whether ChatGPT is a medical product and who is liable, even though ChatGPT always generates a disclaimer that it is not a health-care professional licensed to give medical advice. This is a typical case of intended use versus actual use as described in the medical device regulation. We argue that there is a difference between general-purpose conversational AI—in which the focus is the conversational ability such as readability—and medical AI—in which the focus is the health facts about flesh-and-blood humans. Speaking a fake fact using elegant words is amusing (that is why many ChatGPT users are tricking this conversational AI), but providing a wrong fact in medical AI is dangerous—indeed, making ChatGPT a medical device if it should turn out that doctors are actually using it to diagnose and treat their patients. Nevertheless, philosophically, asking ChatGPT for health-related information (to inform health decision making) is not much different from asking Dr. Google, which has long been criticized for not just giving but spreading medical misinformation (*10*). Nevertheless, this is again not only the gap between intended use versus actual use but also the consistent push and pull between the expectations of the developers versus the end users.

As always with any potentially disruptive technologies, such use can be seen as either a threat or an opportunity. Many articles are optimistic, pointing to the potential symbiosis, the modern centaur, a combination of humans and computers leading to a beneficial augmentation of our capabilities. But pessimistic views also need to be discussed. Take the global positioning system, for example. Because of this technology, many young people are no longer able to navigate with a compass and map. Of course, one could argue that use of a map is not required as a basic skill anymore. But that is certainly not the case with language. If we as humans lose the ability to communicate, debate, and think critically, then we are taking a step backward, leading to devolution.

The question remains: what is the actual use of ChatGPT, despite all the hype during the last few months? Of course, it can be used to generate simple text and to produce code snippets (but often with errors). It can even quickly analyze a research topic and generate an academic paper—again, with frequent errors that may go unnoticed even by reviewers and editors of scientific journals (*11*).

Received Mar. 7, 2023; revision accepted Mar. 14, 2023.
For correspondence or reprints, contact Jens Kleesiek (jens.kleesiek@uk-essen.de).
Published online Apr. 13, 2023.

DOI: 10.2967/jnumed.123.265687

This application may be helpful for student assignments but will not be of much use for learning, in which an individual must come up with a solution through a step-by-step thought process. And when the teacher turns to an oral exam at the end of the semester, irresponsible students who have used the ChatGPT approach will most likely fail, as they might later in their actual work life.

In a health-care setting, one cannot afford to stay on the surface. ChatGPT produces false information that requires checking and correcting of every sentence. In addition to the significant time investment, which calls into question the efficiency of this approach, the real danger comes from false information that goes undetected by the human corrector.

We argue that this technology is worth building toward a clinical knowledge system that can provide health and clinical decision support and enable better self-care and patient care in an era of skill shortages. Hundreds of medical articles are published every day, and it is impossible for humans to cope with this flood of information. In particular, we must keep in mind that information is not the same as knowledge. Filtering information and extracting knowledge from it have enormous potential. Nevertheless, from a health behavior perspective, knowledge is only a small component of healthful behavior and decision making (*12*). Other aspects, such as beliefs, feelings, norms, and the importance of healthful behavior, are equally important. In our view, developing ChatGPT into a medical product such as a clinical decision support system needs to be considered in a broader context with a wider range of other aspects (e.g., reliability, ethics, and fairness) than just model performance, and like any other AI system, humans must be in the loop (*13*).

LLMs have much potential in health care. For example, text-to-text generation may help autocomplete the sentences and paragraphs of a clinical document (e.g., a progress report) based on short phrases provided by a human clinician, thus reducing the documentation burden (*14*). When used for the generation of clinical documents, LLMs also have the potential to integrate the observations of clinicians and knowledge about clinical guidelines, thus reflecting real-world diagnosis and treatment patterns and subsequently being helpful for compiling a differential diagnosis and composing treatment plans. But much more research and development are needed to achieve this goal. We recently developed the first (to our knowledge) clinical LLM, GatorTron (8.9 billion parameters) using over 90 billion words of text (including 82 billion words of clinical text) and demonstrated its power in clinical natural language processing (*15*). We also examined the text generation ability of SynGatorTron (5 billion and 20 billion parameters), a generative clinical LLM based on the GPT-3 architecture (*16*). There is ongoing research on LLMs addressing fundamental issues, such as incorporating chains of reasoning through selection-inference and chain-of-thought prompting (*17–19*). Once current limitations have been addressed, many applications are conceivable using ChatGPT and the next generation of LLMs (Table 1).

More than a decade ago, self-driving cars were heralded as a disruptive technology. As with many technologies, the last 20% of development takes 80% of the total time. The same may be happening with ChatGPT, and additional development time will be needed for productive use in health care. Like the Human Genome Project, in which mapping out the base pairs in the human DNA is not the end but only the start of the genetic revolution, Chat GPT's eventual possibilities are certainly more exciting than the current hype. Whether this technology is a revolution or just an evolution remains to be seen. What is certain in any case is that there will be no more stepping back. As a society, we have the responsibility to shape its future development.

TABLE 1
Potential Applications and Areas of Health-Care Research for ChatGPT and Similar LLMs

Area no.	Description
1	Models and applications that can leverage multimodal data such as merging language and imaging, for example, highlighting anomalies in a natural way (with language) when reading PET images
2	Summary of complex medical histories and records
3	Summary of information from medical congresses/clinical trial results
4	Structuring/making information interoperable, for example, during medical documentation (*20*)
5	Facilitating clinical documentation such as writing discharge report; once we have structured information, is there really a need for free text? (facts should be communicated reliably and concisely)
6	Integration with hospital information systems to incorporate patient data, specifications, and requirements (institutional, payer) and resources (staff capacity, provider)
7	Interpretation and explanation of other AI algorithms (*1*)
8	Translation into other languages, with big potential for less frequently used languages for which use of natural language processing was limited in the past
9	Translation into patient-comprehensible language, making medical information communication more consumer-friendly
10	Anamnesis
11	Relief for nursing staff through automized ward communication
12	Medical writing (*21*)
13	Anonymization of clinical text
14	Fairness, bias in LLMs
15	Human-in-loop and human-centered design of LLM applications
16	Chain-of-thought and automated reasoning on LLMs

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology.* January 26, 2023 [Epub ahead of print].
2. Introducing ChatGPT. OpenAI website. https://openai.com/blog/chatgpt. Published November 30, 2022. Accessed March 20, 2023.
3. Ma Y, Liu J, Yi F, et al. AI vs. human: differentiation analysis of scientific content generation. arXiv website. https://arxiv.org/abs/2301.10416. Published January 24, 2023. Revised February 12, 2023. Accessed March 20, 2023.
4. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. *Minds Machines.* 2020;30:681–694.
5. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems.* Vol 33. Curran Associates, Inc.; 2020:1877–1901.
6. Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. ACL Anthology website. https://aclanthology.org/2021.acl-long.295/. Published August 2021. Accessed March 20, 2023.
7. Aligning language models to follow instructions. OpenAI website. https://openai.com/research/instruction-following. Published January 27, 2022. Accessed March 20, 2023.
8. Roose K. How ChatGPT kicked off an A.I. arms race. *The New York Times website.* https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html. Published February 3, 2023. Accessed March 6, 2023.
9. Confabulation. Wikipedia website. https://en.wikipedia.org/wiki/Confabulation. Updated March 16, 2023. Accessed April 4, 2023.
10. Granter SR, Papke DJ. Medical misinformation in the era of Google: computational approaches to a pervasive problem. *Proc Natl Acad Sci USA.* 2018;115:6318–6321.
11. Anderson N, Belavy DL, Perle SM, et al. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in sports & exercise medicine manuscript generation. *BMJ Open Sport Exerc Med.* 2023;9:e001568.
12. Integrated behavior model. Perelman School of Medicine website. https://www.med.upenn.edu/hbhe4/part2-ch4-integrated-behavior-model.shtml. Accessed March 20, 2023.
13. Patel BN, Rosenberg L, Willcox G, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit Med.* 2019;2:111.
14. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Dig Health.* 2023;5:E107–E108.
15. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med.* 2022;5:194.
16. SynGatorTron: a large clinical natural language generation model for synthetic data generation and zero-shot tasks. NVIDIA website. https://www.nvidia.com/en-us/on-demand/session/gtcspring22-s41638/. Published March 2022. Accessed March 20, 2023.
17. Creswell A, Shanahan M, Higgins I. Selection-inference: exploiting large language models for interpretable logical reasoning. arXiv website. https://arxiv.org/abs/2205.09712. Published May 19, 2022. Accessed March 20, 2023.
18. Tafjord O, Mishra BD, Clark P. Entailer: answering questions with faithful and truthful chains of reasoning. arXiv website. https://arxiv.org/abs/2210.12217. Published October 21, 2022. Accessed March 20, 2023.
19. Kazemi SM, Kim N, Bhatia D, Xu X, Ramachandran D. LAMBADA: backward chaining for automated reasoning in natural language. arXiv website. https://arxiv.org/abs/2212.13894. Published December 20, 2022. Accessed March 20, 2023.
20. Fink MA, Kades K, Bischoff A, et al. Deep learning–based assessment of oncologic outcomes from natural language processing of structured radiology reports. *Radiol Artif Intell.* 2022;4:e220055.
21. Biswas S. ChatGPT and the future of medical writing. *Radiology.* February 2, 2023 [Epub ahead of print].