

CORR Insights®: Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT

Jaret McGraw Karnuta MD, MS¹ 

Where Are We Now?


The emergence of basic copper metallurgy 7000 years ago marked the transition from the Stone Age to the Metal Age. Two hundred years ago or so, the introduction of mechanized manufacturing initiated the Industrial Revolution and ushered humanity into modernity. These days, progress in data processing and information synthesis will characterize the early 21st century as

This CORR Insights® is a commentary on the article “Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT” by Lum available at: DOI: [10.1097/CORR.0000000000002704](https://doi.org/10.1097/CORR.0000000000002704).

The author certifies that there are no funding or commercial associations (consultancies, stock ownership, equity interest, patent/licensing arrangements, etc.) that might pose a conflict of interest in connection with the submitted article related to the author or any immediate family members.

All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.

The opinions expressed are those of the writer, and do not reflect the opinion or policy of CORR® or The Association of Bone and Joint Surgeons®.

J. M. Karnuta , Hospital of the University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104, USA, Email: jaret.karnuta@gmail.com

the dawn of the artificial intelligence (AI) era.

Undoubtedly, the societal transformations spurred by the introduction of metal goods and later the Industrial Revolution played a crucial role in shaping the evolution and practice of medicine. In the same vein, the groundbreaking technology of AI is set to influence—and potentially redefine—the field of medicine.

Early in its introduction to orthopaedics, AI was used to predict relatively simple outcomes using purely numerical inputs, such as length of stay after arthroplasty [11], and to create value-based payment models for reimbursement after hip fracture fixation [6]. Just a few years of development has led to the current state of AI in our field, such as models used to identify arthroplasty component manufacturers using plain radiographs [5] and those that can rival subspecialized physician diagnostic prowess at identifying vertebral fractures on radiographs [7].

However, there are many other (perhaps more revolutionary) AI methods that have thus far not been used in orthopaedics. One such method, which is getting a lot of attention in the media, is the so-called generative stable diffusion model used

by programs such as Dalle-2, which can generate high-quality images from text inputs and generative, pretrained transformer large language models [12] such as GPT4, the underlying technology behind ChatGPT.

In this issue of *Clinical Orthopaedics and Related Research*®, Lum [8] highlights an important application of language-based AI to orthopaedics and its current ability to digest and synthesize orthopaedic knowledge. As shown here, although the AI model was able to quickly recall facts, it was largely unable to synthesize aggregate information to answer complex questions. Thus, we can breathe a collective sigh of relief. At least, for now, our jobs are safe [2]; the program described here could not pass the American Board of Orthopaedic Surgery Part I examination.

Where Do We Need To Go?

Because AI technology is so revolutionary, it is difficult to even fathom the question of “Where do we need to go?” Just as the Stone Age toolmaker couldn’t have envisioned the benefits of copper, just as the Copper Age craftsman wouldn’t have contemplated the advantages of iron, we are riding a wave of history, and at best can try to steer just a little bit.

This current study [8] highlights the current state of how natural language

¹Resident Physician, Hospital of the University of Pennsylvania, Philadelphia, PA, USA

models perform while synthesizing highly domain-specific knowledge: poorly. However, it must be noted that ChatGPT was trained by scanning a large corpus of general, English-language texts (basically, combing the internet), rather than by reading specialized orthopaedic texts and manuscripts. But just imagine if the system had read those instead!

If we had an AI system that was specifically trained for orthopaedic knowledge, I'd imagine that it would have passed the exam. With that, we'd have a true paradigm shift in how we teach and test budding orthopaedic surgeons and how these surgeons would care for their patients. Beyond the addition of a second (and in many ways, better), ectopic brain to aid the surgeon in diagnosis and management of patient disease, such a system could help improve a patient's satisfaction with their surgeon in the clinic and ultimately lead to happier patients [3].

Of course, such a system suggests an obvious question: How can we incorporate AI into patient visits ethically and effectively to provide unbiased, personalized care? All trained models are susceptible to the "garbage in, garbage out" rule: When poor-quality or flawed data are used as input for a computer system or algorithm, the resulting information or conclusions drawn will also be of poor quality or flawed. Additionally, such models act as a "black box," meaning their outputs cannot be reliably audited. Indeed, with edge cases, these systems may confabulate predictions and provide spurious information to the end user [10]. It's impossible (thus far) to ask the AI, "how do you know that you know that?", for even a reference to a valid citation raises the question, "Why did you choose that citation?", or deeper still, "Why did

you choose that particular search strategy?" This process of asking questions can be repeated recursively ad infinitum. The combination of these flaws of AI models means that bias can be perpetuated throughout output text [1] and we cannot perform a digital autopsy to understand why such biases exist [4].

The study in question [8] also raises the issue of how and why standardized testing is performed in our subspecialty. As AI becomes more intelligent, does the utility of a test that is based solely upon a fund of knowledge become less important? Should such tests focus more on complex, less-algorithmic problems that are outside the current scope of AI? Or perhaps highlight issues that arise with patient communication and shared decision-making? These questions will likely become more pressing as AI technologies enter the exam room and operating theater.

How Do We Get There?

On a simple level, this current paper [8] prompts us to examine the issue of assessment. Just as the importance of arithmetic proficiency diminished with the introduction of pocket calculators, the need to memorize facts becomes less critical when the entire world's knowledge is accessible through our smartphones. Consequently, we must reassess the present state of certifying exams. Are we truly serving learners (and their patients) by awarding licenses and certifications based on tests that a poorly trained chatbot can almost pass today? Maybe we are indirectly testing the habits of mind and deferral of gratification that allows a human to pass the test, but if so, we need to be honest about it.

On a more complex level, we must grapple with the appropriate role of AI

in medicine. Microsoft is already pioneering the integration of medical-domain expertise into generative language models [9], such that future tools might be able to diagnose medical conditions and provide consensus-based guidance on basic management in a ChatGPT-like format. Beyond the aforementioned issues regarding bias and factually incorrect outputs, accountability remains a large concern. Who is the key stakeholder in such a system? The developer, physician, or the patient? When the system malfunctions to result in patient harm, who is at fault?

Because big changes are on the horizon, we are in a unique position to shape the trajectory of this technology (although some argue it may already be too late [13]). Early, frequent, and vocal feedback to governing bodies and developers is of paramount importance. Ultimately, the stopgap is us, as physicians and patient advocates. No group is as committed to protecting patients, and it is our duty to shape the regulatory and safety mechanisms surrounding the inevitable introduction of AI in medicine. I fear that, without our summated and coherent voice, AI may be implemented hastily, without thorough examination and dialogue.

References

1. Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. *Nat Mach Intell.* 2021;3:461-463.
2. Bernstein J. Not the last word: ChatGPT can't perform orthopaedic surgery. *Clin Orthop Relat Res.* 2023;481:651-655.
3. Golz A, Kim A, Murphy M, Salazar D. Patient attitudes and preferences for orthopaedic surgeon greetings. *J Am Acad Orthop Surg.* 2021;29:e126-e131.
4. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine.* 2023;90:104512.

5. Karnuta JM, Luu BC, Roth AL, et al. Artificial intelligence to identify arthroplasty implants from radiographs of the knee. *J Arthroplasty*. 2021;36:935-940.
6. Karnuta JM, Navarro SM, Haeberle HS, Billow DG, Krebs VE, Ramkumar PN. Bundled care for hip fractures: a machine-learning approach to an untenable patient-specific payment model. *J Orthop Trauma*. 2019;33:324-330.
7. Li Y-C, Chen H-H, Horng-Shing LH, et al. Can a deep-learning model for the automated detection of vertebral fractures approach the performance level of human subspecialists? *Clin Orthop Relat Res*. 2021;479:1598-1612.
8. Lum GC. Can artificial intelligence pass the American Board of Orthopaedic Surgeons examination? Orthopaedic residents versus ChatGP. *Clin Orthop Relat Res*. 2023;481:1623-1630.
9. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. 2022;23:bbac409.
10. Marcus G. AI platforms like ChatGPT are easy to use but also potentially dangerous. *Scientific American*. Available at: <https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/>. Accessed May 22, 2023.
11. Ramkumar PN, Navarro SM, Haeberle HS, et al. Development and validation of a machine learning algorithm after primary total hip arthroplasty: applications to length of stay and payment models. *J Arthroplasty*. 2019;34:632-637.
12. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Accessed May 9, 2023.
13. Yudkowsky E. Pausing AI developments isn't enough. We need to shut it all down. *Time*. Available at: <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>. Accessed May 12, 2023.