Techniques in Coloproctology (2023) 27:609–614
https://doi.org/10.1007/s10151-023-02837-8

EDITORIAL

How large language models including generative pre-trained transformer (GPT) 3 and 4 will impact medicine and surgery

S. B. Atallah[1] · N. R. Banda[2] · A. Banda[2] · N. A. Roeck[3]

Received: 12 June 2023 / Accepted: 12 June 2023 / Published online: 15 July 2023

Introduction

Recently, a promising autoregressive large language model (LLM), *G*enerative *P*re-trained *T*ransformer (GPT)-3 trained with 175 billion parameters via cloud computing [1] has been made available to the public online (released by OpenAI on November 30, 2022; https://chat.openai.com/). Its size makes it one of the largest deep learning models ever created [2–5]. ChatGPT's global uptake has been exponential: 40 days post launch, GPT-3, had 10 million daily users, surpassing social media giant Instagram in daily users [6] and becoming an overnight "cultural sensation" [7]. GPT-4 was released on March 14, 2023, and it is capable of performing better than humans on high-level professional school exams, and it is perceived as a general purpose artificial intelligence (AI) that is suitable for multiple economic sectors, including healthcare.

ChatGPT (GPT-3/4) is able to generate de novo textual outputs that are grammatically and semantically fluent. Interestingly, human performance for a task does not define the upper bound of LLMs [8–10]. GPT-3 and higher iterations are able to write computer code, compose poetry, generate unique musical composition, and even create cooking recipes. Importantly, ChatGPT can be used for a multitude of healthcare-relevant scenarios.

S. B. Atallah
atallah@post.harvard.edu

N. R. Banda
nickbanda12@gmail.com

A. Banda
Archanab@me.com

N. A. Roeck
nar18d@fsu.edu

1 HCA Healthcare Oviedo, Adventheath Orlando, University of Central Florida, College of Medicine, Orlando, FL, USA

2 University of Central Florida, Orlando, FL, USA

3 Florida State University, Tallahassee, FL, USA

Uses in medicine pertaining to clinical care and education should be considered as target applications for this disruptive technology. Herein, we explore how LLMs can be applied to health literacy, decision-making, and written task formation. Potential pitfalls of LLMs are also discussed.

Understanding large language models (LLMs)

LLMs are a specific application of Natural Language Processing (NLP), which is a subfield of AI that focuses on the interaction between computers and human language. It involves processing and analyzing natural language data (text or speech) to enable machines to understand, generate, and respond to human language [11].

In the field of AI, LLMs are a subset of deep learning that overlaps with *generative AI*. Generative AI is capable of producing text, images, audio, and synthetic data. Current LLMs are most useful for four types of outputs: (a) text classification; (b) question answering; (c) document summarization (including sentiment analysis); and (d) text generation.

While GPT-4 is the most advanced iteration at the time of this writing (2023), there are other LLMs including PaLM (Pathways Language Model), LaMDA, Microsoft Bing, Google Bard, BERT, and T5. To be capable of advanced textual outputs, all of these models are based on *transformer* neural networks [12], as opposed to *convolutional* neural networks. The latter defines the architecture underlying modern computer vision [13].

One important aspect of transformers is that they are *highly parallelizable* [1]. By using thousands of graphics processing units (GPUs) in parallel, GPT-3 was able to be trained in just 1 month. Without being parallelizable, this would have required 355 years using a single GPU. GPUs were initially developed for video gaming because graphic outputs requires parallelized computation. Transformers allow for rapid scale-up compared to previous

AI for system training [14]. OpenAI's original ChatGPT (2018) [15] and ChatGPT-2 (2019) [16, 17] were LLMs developed a few years prior to GPT-3 and 4, but did not possess the power of the current iterations.

LLM architecture is based on being able to *probabilistically* predict the next word in a sentence: For example, the colors of the US flag are red, white, and _____. Mathematically, this can be expressed as $P(X_n|X_{n-1})$, whereby the probability P of the next word X_n is based on the word appearing immediately before it in the sentence, X_{n-1}. But the probability of the next word can also be based on more than the *immediately* prior word, thus more generally:

$$P(X_n|X_{n-1}, X_{n-2}, X_{n-3}, X_{n-4}, \ldots)$$

While transformer neural networks are highly complex, a critical part of their operation includes assigning *weights* to the words in the sentence because, generally speaking, certain words in a sentence *are* more important than others. A model can assign parameters, φ, to maximize the probability that X_n will be accurate and that it will make grammatical sense. Thus transformer LLMs generally use this formula:

$$P_{\varphi}(X_n|X_{n-1}, X_{n-2}, X_{n-3}, X_{n-4}, \ldots)$$

This is a simplified mathematical representation of how transformer LLMs are able to determine the next word (responses) to text queries, and how they are able to perform generative AI functions. LLMs do not actually "think", but rather the model generates textual outputs based on next word probability *and* by "*paying attention*" to key words in text [12]. In computer science, this is termed, *autoregression*, where a weighted sample of past data (text) is used to predict *future* results or textual outputs.

The architecture of prior iterations of ChatGPT and current iterations (GPT-3/GPT-4) remained fundamentally the same; however, the build of the initial *pre-training* process was much smaller for GPT-1 and GPT-2, with approximately 20,000 times more computation used for the current model's training [18]. In computing, what is known as *scaling law* allows LLMs to become exponentially more intelligent by feeding them more data [19, 20]. Importantly, GPT-3/4 are capable of so-called *few shot* or even *zero shot* learning, whereby learning is achieved from just a few or no examples. In addition, GPT-3/4 have the ability to perform *chain-of-thought prompting* to demonstrate reasoning, e.g., showing the steps for solving a complicated problem in calculus or physics [8, 21]. According to S. Bowman, experts are not yet able to interpret the inner workings of LLMs, because the LLMs generate outputs based on the vast textual data they are fed, and we have no satisfactory method to know "what kinds of knowledge, reasoning, or goals a model is using when it produces some output" [8].

LLMs in healthcare

One of the most intriguing ideas behind GPT-3, and higher versions, is that it can not only analyze voluminous amounts of textual data but that it can also *compose* it, making it a functioning generative AI model [2, 8, 22]. A potential use of GPT-3/4 by healthcare providers is implementation as an informatics support system with the objective of reducing staff workload and patient wait times. It can rapidly synthesize high volume, complex patient data and generate summative reports effortlessly. Theoretically, healthcare IT can integrate LLMs into the electronic medical record (EMR) so as to use the generative AI capability to write discharge summaries [23], assist with data entry, and optimize patient check-in for visits (e.g., by assimilating necessary patient data prior to consultation and treatment). Through EMR–LLM integration and, in the future, App-embedding, physicians and medical staff can conserve time, reallocating it to more patient-centric tasks which mandate interpersonal interactions—including face-to-face consultation and personalized treatment. In this manner, LLMs could have a transformative impact on productivity and well-being, ultimately decreasing provider burnout rates in the medical field [24, 25] while enriching patient experience. The advantages and limitations of LLMs in healthcare are summarized in Table 1.

Real-world example

In the USA, it is not infrequent that medical insurers deny payment for procedures and services rendered to patients, such as the "off-label" use of medications [26]. LLMs can be tooled to sift through patient data from EMRs and thus generate articulate medical insurance appeal letters written with proper prose; this is one example of how GPT-3/4 can be used to offset a growing clerical and logistical burden that ultimately results in delayed delivery of healthcare. This reallocated time could, in turn, be used for direct patient care. In surgery and research, it appears GPT-3 has a role in autogenous tasks [27], including creating grant proposals [28], and generating procedure-specific consent forms [29, 30].

In the following example GPT-3 was used to generate a response to a hypothetical insurance denial letter. This response was generated within 10 s after online query; the query and response are shown verbatim in Fig. 1.

One can see from this practical example that GPT-3's textual output is in fluent, native-speaker English with appropriate syntax. This is just one example of how ChatGPT can be utilized to improve healthcare delivery in the real world.

Table 1 Summary of large language model applications in healthcare

Advantages	Limitations
Generative AI	*Biased output*
Diagnostic decision tool (e.g., differential diagnosis)	When trained with biased information, it can generate discriminatory responses which perpetuates health disparities
Treatment decision formulation	
Document generation and summative reports (e.g., discharge summary, letters)	
Resource tool	*Ethical considerations*
Increases information accessibility	Risks to privacy and confidentiality (e.g., data security, informed consent)
Educational tool (e.g., generation of synthetic patient data; improved health literacy	Maintaining patient trust
Research tool	
Communication enhancement	*Minimized information capacity*
Virtual health assistant	Dated information; GPT trained on knowledge predating September 2021
Increase awareness and improved healthcare information delivery	Risks involved when inquiring about unfamiliar current information and development
On-demand clinical support	
Staff workload optimization	*Generation of factually incorrect information*
Textual output automation	"Hallucinations"/fictitious textual outputs
Predictive analytics	Reliance on general patterns due to lack of context awareness
Assistance with triage	Propagation of misinformation
Data organization (e.g., patient health informatics)	
Telemedicine enhancement	*Legal considerations*
Information gateway for providers and patients	Ownership of LLM output, fair use violations, generation of unauthorized, copyright content
Improved delivery of healthcare in underserved regions	Liability: responsibility of provider, user, or AI developers?
	Mitigated by professional judgment
	Potential for misuse and malpractice
	Over reliance on LLM recommendations

AI artificial intelligence, *GPT* Generative Pre-trained Transformer, *LLM* large language model

Educational resource for patients and providers

The ability for GPT-3 (and higher iterations) to abstract large volumes of data and present it succinctly makes its output more useful than classical search engines; and this is a key reason why the technology has gained rapid usership. ChatGPT could serve as a tool to enhance health literacy, i.e., the capacity to seek, understand, and act on health information [31]. With the advent of this technology, patients have a way to probe an intelligent chatbot for healthcare knowledge. Since ChatGPT often provides in-depth and concise responses to health inquiries, individuals interested in furthering their knowledge of disease and health have the potential to ask questions and obtain responses that can be easily understood by laypersons. Whether advisable or not, patients will learn from ChatGPT's responses, shaping their knowledge base and allowing users an "invisible hand" in the algorithm of healthcare [32], since they are able to ask specific questions and receive immediate personalized responses. One can predict that LLMs will eventually dethrone "Dr. Google" [33] to become the first choice as a source of reliable healthcare information.

Limitations and future considerations

When obtaining knowledge from ChatGPT it is critical for individuals to also be aware of the limitations [2] of the model such as its ability to generate factually incorrect information and produce potentially harmful or biased content. At least currently, GPT-3/4 lack training on events and developments which are current or those which occurred within the recent past (3 years). Despite these limitations, it serves as a dynamic tool patients can use to learn from.

Query: my patient's insurance company is denying payment for a staging rectal MRI for his new diagnosis of rectal cancer. I want you to write a 600-word letter explaining why this rectal MRI should be approved.

ChatGPT Response:

Dear [Insurance Company],

I am writing to appeal the denial of payment for a staging rectal MRI for my patient who has recently been diagnosed with rectal cancer. As the patient's healthcare provider, I strongly believe that this MRI is a necessary diagnostic tool that will help in accurately staging the patient's cancer and guiding appropriate treatment.

Rectal cancer is a serious and potentially life-threatening condition that requires prompt and accurate diagnosis and treatment. Staging is a critical component of the diagnostic process, as it allows for the determination of the extent of the cancer and the appropriate course of treatment. While other imaging modalities such as CT scans and ultrasounds can be useful, MRI is considered the gold standard for evaluating the extent of rectal cancer due to its superior soft tissue resolution and ability to detect subtle changes in the anatomy and surrounding tissues.

Furthermore, the patient's specific case warrants the use of an MRI due to the potential for a more advanced stage of cancer. This patient has already undergone a biopsy that confirmed the presence of cancer in the rectum, and additional imaging is necessary to determine the extent of the cancer and plan appropriate treatment. Given the high incidence of lymph node involvement in rectal cancer, an MRI is necessary to accurately evaluate the regional lymph nodes and determine if additional treatment such as chemotherapy or radiation is necessary.

In addition, an MRI can also detect any potential spread of the cancer to other organs such as the liver or lungs, which can significantly impact the course of treatment. Early detection of metastasis can prevent the need for more invasive treatments and increase the patient's chances of survival.

It is also important to note that an MRI is a non-invasive and safe imaging modality with minimal risk to the patient. While there are some contraindications for MRI such as pacemakers or other implanted devices, the patient in question has been cleared for this imaging modality and the benefits of the test far outweigh any potential risks.

Lastly, denying payment for this necessary diagnostic test can lead to significant delays in the patient's treatment, which can ultimately result in poorer outcomes and increased healthcare costs. Delayed diagnosis and treatment can allow the cancer to progress, requiring more extensive and costly treatments such as surgery or chemotherapy. Furthermore, delays in treatment can have a significant impact on the patient's quality of life and well-being, which is something that should be avoided whenever possible.

In conclusion, I strongly urge you to reconsider your decision to deny payment for a staging rectal MRI for my patient with rectal cancer. This test is a critical component of the diagnostic process and will allow for accurate staging and appropriate treatment planning. The safety and effectiveness of MRI, coupled with the potential for improved patient outcomes and decreased healthcare costs, make this test a necessary and valuable component of the patient's care.

Thank you for your attention to this matter.

Sincerely,

[Your Name]

Fig. 1 ChatGPT-3's rebuttal letter for denial of services. Both query and response are unedited. *GTP-3* Generative Pre-trained Transformer-3, *MRI* Magnetic Resonance Imaging, *CT* Computed Tomography

Individuals who reside in remote locales often face a significant barrier: lack of easy access to medical professionals and health resources. This health disparity has a substantial impact on an individual's health awareness, and GPT-3/4 will benefit persons residing in such underserved communities.

It does not take much to imagine an amalgam of technology that includes speech recognition, recent developments of human-expressive robotics [34], and generative AI via LLMs. Combining those elements appears to be a natural next step and a future in which a patient can be evaluated and triaged by a speaking intelligent humanoid robot in not inconceivable. Once the stuff of science fiction, such a construct could become a reality in the near future.

Ethics and safety governing LLMs remain challenging, especially because ChatGPT is growing at an exponential rate and also because it is definitely prone to misuse [2, 8]. Through proper human oversight, future iterations aim to minimize bias and potential harm to humans. Current renditions have proven to be suboptimal in certain settings and along certain chat streams. It has been shown in simulation, for example, that GPT-3 was capable of encouraging suicidal ideation in a mock patient [35], raising serious concern for public health rendered via GPT's unregulated and free access.

GPT-3 delivered a functional AI to the masses, which is in itself an historic achievement. We must be prepared to shepherd its use to prevent malicious application and/or harmful outputs. Isaac Asimov famously set forth the cornerstone principles fundamental to all artificially intelligent systems. While intended for robots specifically, the *Three Laws of Asimov* [36] can be broadened to be inclusive of all AI—including GPT and other LLMs. Expanding this, Asimov's Laws are as follows:

1st law	A robot/AI may not injure a human being or, through inaction, allow a human being to come to harm.
2nd law	A robot/AI must obey the orders given it by human beings except where such orders would conflict with the 1st law.
3rd law	A robot/AI must protect its own existence as long as such protection does not conflict with the 1st or 2nd law.

Recall that computers have been making critical decisions in healthcare for decades. Perhaps the best example of this is the computerized human-*independent* algorithms that control the automated external defibrillator (AED), first developed by Paul Zoll at Harvard Medical School in the 1950s [37]; AED implementation has been crucial not just because of the access of the device in public areas, but the ability for it to be operated *without* the need for medical expertise and intentionally without human input [38].

While humans have been accustomed to specific computerized applications in medicine such as AEDs, generative AI models are far more complex and may not be as easily adapted. Modern LLMs do not always get it right and on occasion produce fluent and grammatically correct textual outputs that are categorically false, and in some instances fictitious—a phenomenon known as "hallucination" [39].

Experts, including OpenAI CEO Sam Altman and others, have voiced concern about the potential misuse of LLMs to write computer code for the purpose of carrying out cyberattacks, as well as GPT's potential to propagate biased information and misinformation. Computer scientists have suggested that LLMs can manipulate humans to acquire power. The majority of 700 plus computer science researchers believe there is more than a 10% probability that humans will not be able to control further advancements in AI leading to "human extinction" [8, 40]. The gravity of these statements suggests that while LLMs are likely here to stay, human oversight and governance will be absolutely crucial.

Data availability All data, analytic methods, and study materials used to conduct the research will be made available to any researcher from the corresponding author upon request.

Declarations

Conflict of interest There are no conflicts of interest to declare.

Ethical approval This is an editorial and ethical approval is not applicable.

Informed consent This publication does not involve patient care, therefore informed consent is not applicable.

References

1. Atallah AB, Atallah S (2021) Cloud computing for robotics and surgery. Digital Surg. https://doi.org/10.1007/978-3-030-49100-0_4
2. Korngiebel DM, Mooney SD (2021) Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digital Med 4(1):93
3. Sallam M (2023) ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. InHealthcare 11(6):887
4. Li J, Dada A, Kleesiek J, Egger J (2023) ChatGPT in Healthcare: a taxonomy and systematic review. medRxiv. https://doi.org/10.1101/2023.03.30.23287899
5. George AS, George AH (2023) A review of ChatGPT AI's impact on several business sectors. Partners Univers Int Innov J 1(1):9–23

6. Liquid Ocelot. https://medium.com/inkwater-atlas/chatgpt-surpasses-instagram-with-10-million-daily-users-in-just-40-days-580944badd9e. Accessed 28 May 2023
7. Thorp HH (2023) ChatGPT is fun, but not an author. Science 379(6630):313–313
8. Bowman SR (2023) Eight things to know about large language models. arXiv preprint arXiv:2304.00612
9. Shlegeris B, Roger F, Chan L (2022) Language models seem to be much better than humans at next-token prediction. Alignm Forum. https://www.alignmentforum.org/posts/htrZrxduciZ5QaCjw/language-models-seem-to-be-much-better-than-humans-at-next
10. Stiennon N, Ouyang L, Wu J et al (2020) Learning to summarize with human feedback. Adv Neural Inf Process Syst 33:3008–3021
11. Hirschberg J, Manning CD (2015) Advances in natural language processing. Science 349(6245):261–266
12. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inform Process Syst 30. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
13 Atallah S (2021) Artificial intelligence and computer vision. Digital Surg. https://doi.org/10.1007/978-3-030-49100-0_31
14. Hoffmann J, Borgeaud S, Mensch A et al (2022) Training compute-optimal large language models. arXiv preprint arXiv:2203.15556
15. Radford A, Narasimhan K, Salimans T et al (2018) Improving language understanding by generative pre-training. OpenAI blog. https://openai.com/research/language-unsupervised
16. Budzianowski, Paweł, Ivan Vulić (2019) Hello, it's GPT-2--how can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. arXiv preprint arXiv:1907.05774
17. Ham D, Lee JG, Jang Y, Kim KE (2020) End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. InProceedings of the 58th annual meeting of the association for computational linguistics (pp. 583–592)
18. Sevilla J, Heim L, Ho A, Besiroglu T, Hobbhahn M, Villalobos P (2022) Compute trends across three eras of machine learning. In 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE
19. S. Hoffmann J, Borgeaud S, Mensch A et al (2022) An empirical analysis of compute-optimal large language model training. In: Oh AH, Agarwal A, Belgrave D, Cho K (eds) Advances in Neural Information Processing Systems, 2022. https://openreview.net/forum?id=iBBcRUlOAPR
20. Kaplan S, McCandlish J, Henighan T et al (2020) Scaling laws for neural language models. arXiv preprint 2001.08361
21. Wei J, Wang X, Schuurmans D et al (2022) Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903
22. Dwivedi YK, Kshetri N, Hughes L et al (2023) "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage 1(71):102642
23. Patel SB, Lam K (2023) ChatGPT: the future of discharge summaries? Lancet Digital Health 5(3):e107–e108
24. Woolhandler S, Himmelstein DU (2014) Administrative work consumes one-sixth of US physicians' working hours and lowers their career satisfaction. Int J Health Serv 44(4):635–642
25. Patel RS, Bachu R, Adikey A, Malik M, Shah M (2018) Factors related to physician burnout and its consequences: a review. Behav Sci 8(11):98
26. Kahn SA, Bousvaros A (2023) Topic of the month: How to write an effective letter of medical necessity. J Pediatr Gastroenterol Nutr 3:10–97
27. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS (2023) Using ChatGPT to write patient clinic letters. The Lancet Digital Health 5(4):e179–e181
28 Janssen BV, Kazemier G, Besselink MG (2023) The use of ChatGPT and other large language models in surgical science. BJS open. https://doi.org/10.1093/bjsopen/zrad032
29. Tel A, Parodi PC, Robiony M, Zanotti B, Zingaretti N (2023) Could ChatGPT improve knowledge in surgery? Ann Surg Oncol 18:1–2
30. Hassan AM, Nelson JA, Coert JH, Mehrara BJ, Selber JC (2023) Exploring the potential of artificial intelligence in surgery: insights from a conversation with ChatGPT. Ann Surg Oncol 5:1–4
31. Paterick TE, Patel N, Tajik AJ, Chandrasekaran K (2017) Improving health outcomes through patient education and partnerships with patients. Proc Baylor Univ Med Cent 30(1):112–11
32. Brake DR (2017) The invisible hand of the unaccountable algorithm: how Google, Facebook and other tech companies are changing journalism. Digital Technol J Int Compar Perspect 25–46
33. Lee K, Hoti K, Hughes JD, Emmerton LM (2015) Consumer use of "Dr Google": a survey on health information-seeking behaviors and navigational needs. J Med Internet Res 17(12):e288
34. Faraj Z, Selamet M, Morales C et al (2021) Facially expressive humanoid robotic face. HardwareX 1(9):e00117
35. Daws R (2020) Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. Available at https://artificialintelligencenews.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/
36. Asimov I (1941) Three laws of robotics. Asimov I. Runaround. 2. ISBN: 978-0-385-42304-5
37. Zoll PM (1952) Resuscitation of the heart in ventricular standstill by external electric stimulation. N Engl J Med 247(20):768–771
38. Caffrey SL, Willoughby PJ, Pepe PE, Becker LB (2002) Public use of automated external defibrillators. N Engl J Med 347(16):1242–1247
39. Ji Z, Lee N, Frieske R et al (2023) Survey of hallucination in natural language generation. ACM Comput Surveys 55(12):1–38
40. Stein-Perlman Z, Weinstein-Raun B, Grace K (2022) Expert survey on progress in AI. AI Impacts blog, 2020. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/