Canadian Journal of Cardiology ■ (2024) 1–9

Review

Revolutionizing Cardiology With Words: Unveiling the Impact of Large Language Models in Medical Science Writing

Abhijit Bhattaru, BA,[a,b] Naveena Yanamala, MS, PhD,[a] and Partho P. Sengupta, MD[a]

[a] *Department of Cardiology, Rutgers Robert Wood Johnson Medical School and Robert Wood Johnson University Hospital, New Brunswick, New Jersey, USA*

[b] *Department of Medicine, Rutgers New Jersey Medical School, Newark, New Jersey, USA*

ABSTRACT

Large language models (LLMs) are a unique form of machine learning that facilitates inputs of unstructured text/numerical information for meaningful interpretation and prediction. Recently, LLMs have become commercialized, allowing the average person to access these incredibly powerful tools. Early adopters focused on LLM use in performing logical tasks, including—but not limited to—generating titles, identifying key words, summarizing text, initial editing of scientific work, improving statistical protocols, and performing statistical analysis. More recently, LLMs have been expanded to clinical practice and academia to perform higher cognitive and creative tasks. LLMs provide personalized assistance in learning, facilitate the management of electronic medical records, and offer valuable insights into clinical decision making in cardiology. They enhance patient education by explaining intricate medical conditions in lay terms, have a vast library

RÉSUMÉ

Les grands modèles de langage (LLM; large language models) représentent une forme unique d'apprentissage automatique qui facilite la saisie de textes non structurés et/ou d'informations numériques en vue d'une interprétation et d'une prédiction significatives. Récemment, les LLM ont été commercialisés, permettant au commun des mortels d'accéder à ces outils incroyablement puissants. Les premiers utilisateurs se sont concentrés sur l'utilisation des LLM pour effectuer des tâches logiques, y compris, mais sans s'y limiter, la génération de titres, l'identification de mots-clés, le résumé de textes, l'édition initiale de travaux scientifiques, l'amélioration de protocoles statistiques et l'exécution d'analyses statistiques. Plus récemment, les LLM ont été étendus à la pratique clinique et au monde universitaire pour effectuer des tâches cognitives et créatives de haut niveau. Les LLM fournissent une aide personnalisée à l'apprentissage, facilitent la

The popularization and commercialization of machine learning methods have increased public awareness, use, and benefits of these powerful algorithms in medicine. One category that has demonstrated great promise is large language models (LLMs). Since their early development, LLMs sought to replicate the human mind's ability to perform logical tasks and comprehend meaningful explanations despite being constructed by aggregated probabilities; this ability to manipulate language was termed "natural language processing (NLP)."[1] In recent history, companies such as OpenAI with ChatGPT and Google with Bard have opened the door for ordinary people to access incredibly powerful LLMs to automate tasks.[2,3]

Given the complexity of medical language, the medical community has adopted LLMs in many tasks. In medical education, LLMs are so computationally advanced that they excel on board examinations such as residency in-training examinations and the United States Medical Licensing Examination (USMLE). Anecdotally, professors at our university have adopted LLMs to provide personalized assistants to break down complex pathways into understandable chunks for students. From a clinical perspective, LLMs are being applied to tackle the abundance of language available in electronic medical record (EMR) systems and perform predictions.

LLMs present unique applications in cardiology, from their patient-side presentations of conditions to furthering academic works. This review will first discuss how LLMs function and how they can perform writing tasks. Next, we will delve into specific applications of LLMs from performing logical tasks to expanding patient education. Finally, we will focus on important limitations of LLMs and considerations before clinicians use LLMs in day-to-day work. Although the literature is limited on the specific applications of LLMs in medical science writing in cardiology, we performed an extensive search to find the best articles to filter for inclusion

Received for publication February 17, 2024. Accepted May 24, 2024.

Corresponding author: Dr Partho Sengupta, Rutgers Robert Wood Johnson Medical School and Robert Wood Johnson University Hospital, 1 Robert Wood Johnson Place, New Brunswick, New Jersey 08901, USA. Tel.: +1-732-828-3000.

E-mail: partho.sengupta@rutgers.edu

See page 8 for disclosure information.

https://doi.org/10.1016/j.cjca.2024.05.022

of knowledge to help clinicians expedite administrative tasks, provide useful feedback regarding content of scientific writing, and assist in the peer-review process. Despite their impressive capabilities, LLMs are not without limitations. They are susceptible to generating incorrect or plagiarized content, face challenges in handling tasks without detailed prompts, and lack originality. These limitations underscore the importance of human oversight in using LLMs in medical science and clinical practice. As LLMs continue to evolve, addressing these challenges will be crucial in maximizing their potential benefits while mitigating risks. This review explores the functions, opportunities, and constraints of LLMs, with a focus on their impact on cardiology, illustrating both the transformative power and the boundaries of current technology in medicine.

gestion des dossiers médicaux électroniques et offrent des informations précieuses pour la prise de décision clinique en cardiologie. Ils améliorent l'éducation des patients en expliquant des conditions médicales complexes avec des termes simples, disposent d'une vaste bibliothèque de connaissances pour aider les cliniciens à accélérer les tâches administratives, fournissent un retour d'information utile sur le contenu des écrits scientifiques et participent au processus de révision par les pairs. Malgré leurs capacités impressionnantes, les LLM ne sont pas sans limitations. Ils sont susceptibles de générer des contenus incorrects ou plagiés, sont confrontés à des difficultés pour effectuer des tâches sans instructions détaillées et manquent d'originalité. Ces limitations soulignent l'importance de la surveillance humaine dans l'utilisation des LLM en science médicale et dans la pratique clinique. À mesure que les LLM continuent d'évoluer, il sera essentiel de relever ces défis pour maximiser leurs avantages potentiels tout en atténuant les risques. Cette étude explore les fonctions, les opportunités et les contraintes des LLM, en mettant l'accent sur leur impact en cardiologie, illustrant à la fois le pouvoir de transformation et les limites de la technologie actuelle en médecine.

in this review. The detailed search criteria are presented in Table 1. A full list of included literature is further reviewed in Table 2.

How Do LLMs Work?

Although LLMs may appear as enigmatic entities capable of providing meaningful output to the user, the algorithm behind popular LLMs such as ChatGPT can be broken down into simple probabilities. Neural networks traditionally sought to replicate human brain structures with several interconnected "neurons" that were individually tuned by specific parameters to provide a joint probability of a particular outcome. LLMs follow this traditional approach with a few additional nuances.

LLMs uniquely handle sequential data and use the interconnectedness of inputs to gauge a response. As a result, they excel in context and relationships between words in sentences, no matter how far apart they are, thereby generating coherent and relevant text; this process happens via predicting sequences of words and where a particular word in the dictionary will exist concerning other related words via presenting numerical representations for each word provided as an input. Such LLMs are dubbed transformers that can apply knowledge of numerical relationships of inputs and present an output. For example, suppose an input had the phrase "Where are we now? Where are we headed?"; an LLM may interpret these phrases as related with the first 3 words being the same and a 50% probability of the final word in each phrase being "now" or "headed."[4] Furthermore, it may use this probability difference to explain the different interpretations of both phrases. We have included a visual representation of how LLMs work in Figure 1.

The training information provided to current tools for LLMs to learn these relationships comes from vast amounts of text data inputted by the developers, some of which may include the entirety of the internet. In addition, these LLMs can be fed specific training data long after deployment to help attune to specific applications; ChatGPT-4 currently allows users to create custom chatbots for this purpose. LLMs can generate text that is often indistinguishable from that written by humans. They can compose poetry, write articles, generate code, and even simulate dialogue. In the medical field, this could translate to drafting patient information leaflets, creating medical reports, or generating patient correspondence based on brief inputs. With these applications, we can explore the world of possibilities through commercialized LLMs.

Doors Opened By LLMs

Simpler logical tasks

To begin exploring, we can focus on the most discrete, logical tasks that LLMs excel in performing (Table 3).

LLMs are great editors of writing, from formatting and grammar to revising ideas. Early ChatGPT was used to generate paper titles or make text more readable using specific keywords that one would come up with during brainstorming sessions. ChatGPT revised this paper's title. Summarizing large volumes of text, especially for practical tasks such as making abstracts of papers, is easily attainable by ChatGPT.[5] A recent editorial discussing LLMs was written entirely by ChatGPT and was free of large grammatical and logical deficits.[6] Aydin and colleagues[7] went even further by using ChatGPT to write their review article. Their results showed successful compilation and knowledge expression by feeding the algorithm with several published items. Authors have used LLMs to alleviate the painstaking process of formatting references in a specific journal's format, although they did submit corrections because of errors.[5,8]

LLMs now serve authors as the initial step to editing and reviewing their scientific work. Einarsson and colleagues[9] used ChatGPT to review their experimental protocols to improve the statistical power of their papers. Recently, ChatGPT was asked by an author to review a paper and provide feedback directly; the model excelled in providing general feedback, such as the degree of technical detail needed to be simplified or the writing style presented in the article.[10] If human reviewers for journals were compared with ChatGPT, the average overlap in feedback provided was 30.85%, as opposed

Table 1. Search criteria

Step	Details
1. Initial database search	**Databases used:** Google Scholar, PubMed **Key words:** Queries combined terms related to ChatGPT, LLMs, and medicine with specific focus areas (biostatistics, medical writing, USMLE, hospital course discharge): "ChatGPT," "Large Language Models," "LLMs," "medicine," "medical," "biostatistics," "writing," "USMLE," "hospital discharge course," "clinical trials."
2. Screening and selection	**Inclusion criteria:** Articles directly related to statistical analysis capabilities, grammar, automated template generation, medical knowledge, peer review, ethical and legal concerns, and originality of LLMs. **Exclusion criteria:** Editorials, letters to editors, and similar communications if their main idea was not significantly different from included original articles or review articles.
3. Focused search on cardiology science writing	**Database used:** PubMed **Key words:** "ChatGPT," "Large Language Models," "LLMs," "medicine," "medical," "cardiovascular," "cardiology," "academic writing."
4. Supplementary searches and tools	**Additional databases:** Google Scholar **Research tools:** ConnectedPapers.com for related articles by DOI or PubMed ID.
5. Article selection and review	**Initial selection:** The decisions regarding which articles to include in the final version was first made by an experienced author with 5+ years in cardiology and machine learning research. Because this is a literature review, we tried to pick the best sources to highlight aspects of our work without being redundant by including several papers on the same topic. **Final review:** Two clinical cardiologists with more than10 years of clinical experience (1 with machine learning research experience).

LLM, large learning models; USMLE, United States Medical Licensing Examination.

to a reference comparison between 2 human reviewers, which was 28.58%.[11]

Continuous developments of the mathematical capacities of commercially available LLMs are now allowing them to perform statistical analysis appropriately. Although initial experiments showed these chatbots failing at simple tasks such as "2 + 2," with users successfully "brainwashing" these algorithms to believe a different answer, developers have improved the algorithms to integrate numerical input directly. The current version of ChatGPT-4.0 can read comma-separated value files and directly perform regressions, comparison tests, and survival analysis with Python. Ignajatovic and colleagues[12] tested this ability by asking ChatGPT to solve biostatistical problems frequently used in their medical education. Unsurprisingly, GPT-4 succeeded in providing all correct answers within 3 attempts, although initial accuracy was wrong. However, the conclusion of this study needs to be interpreted within some context: algorithms such as ChatGPT are being updated daily. Hence, the output it may have provided yesterday is not the same as its output tomorrow (with—it is hoped—more intelligent answers as time progresses). This direct effect is seen in the work from Calonge et al.,[13] in which they showed that ChatGPT-4 outperforms not only its predecessor GPT-3.5 but also Bard and LLaMA: 2 other well-developed LLMs from the tech industry.

These examples show the power of LLMs in approaching simple logical tasks that would otherwise be tedious in the science writing process. Within the realm of cardiology, specifically, few studies explore the power of LLMs for science writing. Dala and colleagues[14] have shown some test cases with ChatGPT for streamlining administrative tasks; they have demonstrated the ability to generate discharge summaries following a specific format with some clinical input provided by the provider. As the authors astutely state, using LLMs in this sense saves physicians a lot of time by reducing administrative burden and freeing up time for much-needed face-to-face interaction in the cardiology clinics. Other models are also exploring reducing administrative burden in cardiology using niche LLMs.[15–17] When clinicians are seeking to confirm their findings with clinical guidelines, Boonstra et al.[17] have shown that LLMs may provide initial information regarding the legitimacy of their conclusions. Alahdab et al.[18] have further discussed the utility of LLMs in the academic writing process, the most important of which is the added benefit of allowing authors to focus on overarching themes and intellectual aspects of their work rather than minutiae in writing.

The current space of commercially available LLMs is competitive, although ChatGPT is currently the most used tool for medical applications. There are 2 versions of ChatGPT that users have access to: 3.5 is the free version that has less training data and has more limited functionality, whereas 4.0 is a subscription-based service that has more recent data and has complex abilities such as statistical analysis. Other tools commonly being used are Bard, which is made by Google (Googleplex, Mountain View, CA), and LlaMA, which is by Meta (Menlo Park, CA).

Beyond Scholarly Articles: Writings to Aid Clinical Decisions and Patient Education

So far, we have discussed how LLMs can tackle more straightforward tasks such as formatting, template-based writing/rewriting, and statistical analysis. However, the emerging literature on LLM use focuses on interpretive clinical decision making and the breakdown of complex topics into digestible patient understanding (Fig. 2).

Ultimately, it is difficult for machine learning to replace the years of knowledge and experience developed by medical professionals for diagnosis, treatment, and management; however, this does not discount the performance of LLMs in this realm. Kung and colleagues[19] demonstrated that ChatGPT passed the USMLE on all 3 levels; these are very advanced, complicated examinations for which students ordinarily study for years. Yet, with only the knowledge provided by developers without special training in medicine, GPT could match human performance. This discovery led to several authors exploring LLM performance on other

Table 2. Overview of studies included in this review

Author	Type of article	Main finding
Liang et al.[11]	Original article	If human reviewers for journals were compared with ChatGPT, the average overlap in feedback provided was 30.85% as opposed to a reference comparison between 2 human reviewers being 28.58%.
Ignjatovic et al.[12]	Original article	ChatGPT can solve biostatistical problems with reasonable explanations, although they may take multiple attempts for accurate results.
Calonge et al.[13]	Original article	GPT-4 outperforms GPT-3.5, Bard, and LLaMA in enhancing student learning of calculus and statistics.
Luo et al.[15]	Original article	BioGPT, a custom-made LLM, can generate fluent descriptions for biomedical terms as read from reputable sources.
Searle et al.[16]	Original article	LLMs can generate brief hospital course summaries with integration of EMR data.
Kung et al.[19]	Original article	ChatGPT passed the USMLE on all 3 levels.
Takagi et al.[20]	Original article	GPT-4 achieved the passing criteria for the Japanese Medical Licensing Examination.
Guerra et al.[21]	Original article	GPT-4 significantly outperformed medical students, neurosurgery residents, and the national average of Congress of Neurological Surgeons Self-Assessment Neurosurgery Examination users.
Jung et al.[22]	Original article	ChatGPT Passes German State Examination and performed clinically accurate explanations.
Al-Ashwal et al.[23]	Original article	BingAI performed better than GPT-4, GPT-3.5, and Bard in assessing drug—drug interactions.
Datta et al.[25]	Original article	A custom made LLM, AutoCriteria, can extract detailed inclusion criteria from clinical trial documents.
Jeblick et al.[26]	Original article	ChatGPT can generate factually accurate, complete, and nonharmful summaries of radiology reports but has potential for missed relevant information.
Walker et al.[28]	Original article	ChatGPT has potential in explaining details of hepato-pancreatico-biliary conditions to patients.
Gritti et al.[31]	Original article	ChatGPT was able to understand the diagnoses of various congenital heart diseases and explain this in detail to users.
Poola[36]	Original article	ChatGPT has flaws in interpretation of prompts, but these inputs can be specifically engineered for better results.
Ekin[37]	Original article	There are some specific techniques, tips, and best practices when prompting LLMs for everyday tasks.
Salvagno et al.[5]	Review	ChatGPT may play an important role in critical care medicine but there are ethical issues that first need to be addressed.
Aydin et al.[7]	Review	This review was written by ChatGPT with assistance from the coauthors.
Einarsson et al.[9]	Review article	ChatGPT has potential to automate several tasks in academic writing.
Gala et al.[14]	Review article	ChatGPT has the ability to generate discharge summaries following a specific format with some clinical input provided by the provider.
Boonstra et al.[17]	Review article	NLP methods are increasingly relevant in clinical care, and LLMs can automate tasks to ease physician administrative burden.
Sim et al.[24]	Systematic review article	LLMs can accurately parse through EMRs to process free-text related to patient-reported outcomes.
Inam et al.[39]	Review article	Journals are now instituting specific requirements on how LLMs can and cannot be used in the scientific process.
Frosolini et al.[8]	Letter to editor	ChatGPT has potential to fabricate information: namely, references as they pertain to medical science writing.
Inojosa et al.[27]	Letter to editor	ChatGPT has potential in explaining disease details of multiple sclerosis to patients.
Skalidis et al.[33]	Letter to editor	ChatGPT was limited in gathering appropriate and clinically relevant articles during literature search.
Shah et al.[4]	Special communication	Clinical usage of LLMs needs to be fine tuned by relevant training data and evaluating the benefits in real situations.
King[6]	Editorial	This editorial was written by ChatGPT with assistance from Dr King.
Donker[10]	Correspondence	ChatGPT and similar tools may be used for initial peer review and provide good summaries but are limited in their quality of feedback.
Alahdab[18]	Opinion	LLMs help in the academic writing process with the added benefit of allowing authors to focus on overarching themes and intellectual aspects of their work.
Rogasch et al.[32]	Brief communication	ChatGPT can explain PET/CT reports to patients and help them with preparation for imaging.
Salihu et al.[34]	Viewpoint	LLMs have potential to help address clinical explanations and judgement, but have several limitations to consider.
Henrickson et al.[35]	Open forum	These authors designed specific prompts with the goal of optimizing hermeneuticity rather than factual accuracy.

CT/PET, computed tomography/positron emission tomography; EMR, electronic medical record; LLMs, large learning models; USMLE, United States Medical Licensing Examination.

board examinations.[20–22] With these strong performances, it is clear LLMs are incredibly valuable in interpreting medical questions provided by cardiologists. Furthermore, these studies show that LLMs have the potential for clinical reasoning, although they are susceptible to the biases of their training data. Al-Ashwal et al.[23] performed a head-to-head comparison among ChatGPT, Bing AI, and Bard in predicting drug—drug interactions in the clinic; early prevention of these interactions prevents dangerous complications and improves targeted decision making for in-hospital management. They discovered that although Bing excelled the most, all LLMs used were proficient, proving valuable in clinical management. Within EMRs, LLMs have also been used to assess subjective unstructured information: namely, patient-reported outcomes.[24]

Within academia, specifically, LLMs have been used as screening tools to assess the quality of work. Datta et al.[25] have created their LLM to interpret clinical trial documents

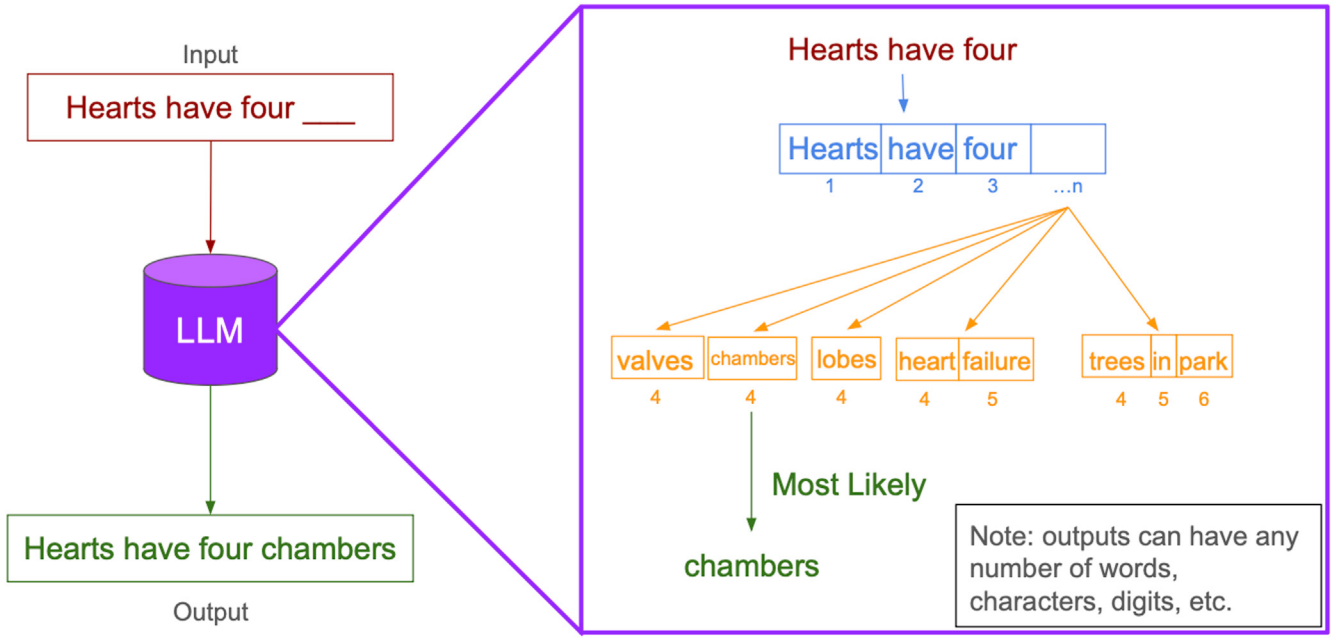

Figure 1. Visualization of algorithmic processing of large language models (LLMs). Note that although "chambers" was diagramed as most likely over "valves," the decision among various options by the algorithm would be based on the full context of the prompt.

and extract eligibility criteria across various diseases. A tool such as this vastly improves the predicted performance of studies by reducing bias while streamlining the process.

LLMs are also being expanded to patient-sided applications to explain complex medical diagnoses in laypersons' terms. This application is precious in historically disenfranchised populations and the pediatric population. One study demonstrated ChatGPT's ability to present findings in radiology reports to children; radiologists agreed that the reports were factually correct, complete, and not harmful for patients to read.[26] Inojosa and colleagues[27] have argued for the utility of ChatGPT in multiple sclerosis to help patients who are unsure of their diagnosis or are left wanting a second opinion because of a lack of knowledge of their conditions.[27] This application was also tested in hepato-pancreatico-biliary conditions with successful results.[28]

Within cardiology, ChatGPT can be particularly valuable in helping patients with dietary preparation for positron emission tomography (PET) imaging; conditions such as cardiac sarcoidosis require a strict ketogenic dietary preparation to suppress baseline myocardial glucose activity, but lack of patient understanding often hinders this process.[29,30]

Table 3. Strengths and weaknesses of LLMs in medical writing and education

Category	Strengths	Weaknesses
Uses in clinical practice	LLMs have a sufficient degree of clinical reasoning to interpret potentially complex cases. LLMs can help detect clinical administrative errors and thereby avoid unneeded medical mistakes. LLMs can help fill in information for discharge summaries and other template-based documents.	LLMs cannot be fed protected health information unless on a local network with additional privileges. LLMs cannot be liable for poor clinical decision making and require supervision by a doctor.
Uses in teaching	LLMs contain vast bodies of medical knowledge because of their intrinsic training protocols and can therefore breakdown complex medical concepts into digestible pieces for students. LLMs can simulate case scenarios for educators to improve teaching material.	LLMs cannot provide specific and detailed feedback unless properly prompted by users.
Uses in research	LLMs can easily edit text and assist with grammatical changes to improve content of papers. LLMs are excellent at summarizing large volumes of text into digestible pieces of information for interpretation. LLMs can act as interim peer reviewers to critique early drafts of authors' manuscripts. LLMs can perform statistical analysis either via direct input of files or by providing drafts of code for users to use.	LLMs are prone to plagiarism, and they struggle in generating creative ideas. LLMs are prone to bias because of variability of user/developer inputted data. LLMs have potential to fabricate references and are generally poor at literature search.

LLMs, large learning models.

Rogasch et al.[31] show that ChatGPT successfully prepared patients for their PET scans and explained the results, thereby improving diagnostic performance.[32] Within pediatric cardiology, ChatGPT could understand the diagnoses of various congenital heart diseases and explain this in detail to users.[31]

Limitations

Undoubtedly, LLMs such as ChatGPT have generated sensationalist energy with their clinical performance and utility in medical science writing. However, these results need to be considered with some significant limitations.

First, LLMs are prone to fabrication and plagiarism. This is logical considering that training data for LLMs comes from other already published works on the Internet. In our earlier discussion of the review article written by ChatGPT and published by Aydin et al.,[7] they mentioned that the model had significant challenges with developing summaries of the articles without copying the text it was given. Within cardiovascular science writing, Skalidis et al.[33] showed ChatGPT was limited in gathering appropriate and clinically relevant articles during literature search. As an experiment, when we prompted ChatGPT to summarize some articles when writing this review, we repeatedly saw it copy phrases and sentences from previously published work. Recently, authors had to submit corrections to their paper because an editorial pointed out that ChatGPT had falsified its reference for an article.[8] When we consider ChatGPT for tasks such as formatting references, we must bear in mind it has no logical basis for what the reference should look like but instead tries to use its knowledge to interpret an answer. As such, volume numbers, years, author names, and titles can easily be forged by ChatGPT and need meticulous review by authors after using it. We have noticed that ChatGPT fails with literature search, providing references from years far beyond what the training data contained. Considering these shortcomings, it is imperative for all authors to meticulously review their work before using work generated by LLMs to ensure accuracy. Ultimately, human authors are liable for the work they produce and publish, even if they use LLMs for supplementation.

Second, LLMs fail when they are not prompted with sufficient detail. Because LLMs are trained with copious amounts of information, they struggle to focus on minute tasks because they weigh other information more heavily. In addition, there is, at present, no mechanism by which ChatGPT can say it simply does not know the answer, a standard answer provided by humans. As such, ChatGPT provides either broad answers without explaining the nuances or gives inaccurate information. Minute changes within the prompts have shown drastic changes in responses by

Table 4. Guidelines of the top 25 cardiology and cardiovascular medicine journals as per SCimago rankings regarding the use of generative artificial intelligence in scientific writing

Journal	Publishing company	Inclusion of AI tools in authorship	Use of AI in content generation	Use of AI in image generation	Use of AI in peer review
Journal of the American College of Cardiology	Elsevier	X	✓	X	X
Circulation	AHA; Lippincott Williams & Wilkins	X	✓	X	X
JAMA Cardiology	AMA	X	✓	X	X
Nature Reviews Cardiology	Nature	X	✓	X	N/A
European Journal of Heart Failure	Wiley	X	✓	X	N/A
JACC: Heart Failure	Elsevier	X	✓	X	X
Circulation Research	AHA; Lippincott Williams & Wilkins	X	✓	X	X
JACC: Cardiovascular Imaging	Elsevier	X	✓	X	X
European Heart Journal	Oxford University Press	X	✓	X	X
Circulation: Heart Failure	AHA; Lippincott Williams & Wilkins	X	✓	X	X
Circulation: Arrhythmia and Electrophysiology	AHA; Lippincott Williams & Wilkins	X	✓	X	X
Circulation. Genomic and precision medicine	AHA; Lippincott Williams & Wilkins	X	✓	X	X
JACC: Cardiovascular Interventions	Elsevier	X	✓	X	X
Arteriosclerosis, Thrombosis, and Vascular Biology	AHA; Lippincott Williams & Wilkins	X	✓	X	X
Stroke	AHA; Lippincott Williams & Wilkins	X	✓	X	X
European Stroke Journal	Sage	X	✓	X	X
Journal of Heart and Lung Transplantation	Elsevier	X	✓	X	X
JACC: Basic to Translational Science	Elsevier	X	✓	X	X
JACC: Clinical Electrophysiology	Elsevier	X	✓	X	X
Circulation: Cardiovascular Quality and Outcomes	AHA; Lippincott Williams & Wilkins	X	✓	X	X
Cardiovascular Research	Oxford University Press	X	✓	X	X
JACC: CardioOncology	Elsevier	X	✓	X	X
Cardiovascular Diabetology	Springer	N/A	N/A	N/A	N/A
Chest	Elsevier	X	✓	X	X
Circulation: Cardiovascular Interventions	AHA; Lippincott Williams & Wilkins	X	✓	X	X

✓, permitted; X, not permitted; N/A, guidelines not available.
AHA, American Heart Association; AI, artificial intelligence.
Reprinted from Inam et al.[39] with permission from *Current Problems in Cardiology*.

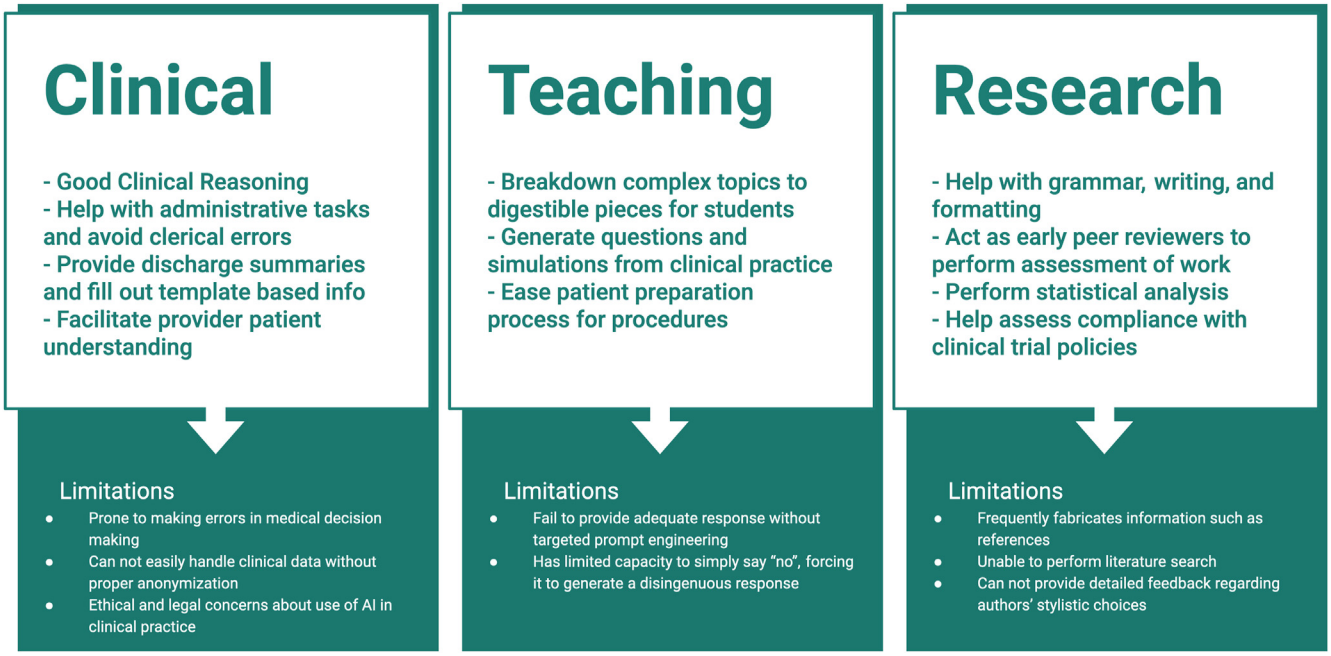

Figure 2. Graphical representation of core conclusions.

ChatGPT, some of which may even contradict.[34] Authors publish "prompting engineering" guides to address these shortcomings to help users get specific answers.[35–37]

Importantly, there remains a dilemma with using patient data within publicly available LLMs. At present, any data input into models such as ChatGPT need to be completely anonymized, as any prompts provided have the potential to be used as future training data for these models. This process of anonymization is susceptible to human error, especially with large volumes of data such as generating discharge summaries. Clinicians who seek to use LLMs in their practice need to be diligent with the protection of patient data to ensure confidentiality. To help alleviate some of these differences, some companies are working with specific health networks to create encrypted local LLMs that are Health Insurance Portability and Accountability Act (HIPAA) compliant.[38]

Finally, LLMs' lack of originality and creativity makes work that relies on their language boring. With repeated paraphrasing, haughty writing, and redundancies, ChatGPT and similar tools are simply poor writers. Within cardiology science writing, we must present results with appropriate clinical judgement to shape the lives of our future patients. Copying responses from LLMs will not allow us to achieve this goal successfully. Appropriately, many journals are now restricting how authors may use ChatGPT to write their papers. Table 3 is reprinted from a recent paper provided by Inam and colleagues,[39] which shows current guidelines for LLM use by major cardiovascular science journals. Nevertheless, with the continuous development of these LLMs, the future looks bright to address some of these shortcomings.

Example Prompts and Responses

In this section, we have used ChatGPT-4 to write some text and provide the unedited responses.[3] This is intended to show readers the potential for LLMs to help them with their writing.

Calculations

[User input] *"I have a patient with peak aortic velocity of 3.5 m/s, mean transaortic gradient (MG) of 40 mm Hg, and aortic valve area (AVA) of 1.0 cm^2 on cardiac magnetic resonance imaging (CMRI). What disease do they have?"*

[ChatGPT] *"These parameters suggest the patient may have moderate to severe aortic stenosis (AS). The classification between moderate and severe AS can sometimes depend on the combination of these parameters and may require clinical correlation with the patient's symptoms and other findings. Your patient's parameters do not fit neatly into one category due to the mean gradient and AVA suggesting severe AS, but the peak velocity is more indicative of moderate AS. Given the borderline values, this patient's case might require further evaluation, possibly including a comprehensive clinical assessment and potentially additional cardiac imaging or tests, to fully understand the severity of the aortic stenosis and to guide treatment decisions."*

Formatting references

[User input] *"Cite this paper in the Index Medicus format.*

2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk:

A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines

David C. GofffJr, Donald M. Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B. D'Agostino, Raymond Gibbons, Philip Greenland, Daniel T. Lackland, Daniel Levy, Christopher J. O'Donnell, Jennifer G. Robinson, J. Sanford Schwartz, Susan T. Shero, Sidney C. SmithJr, Paul Sorlie, Neil J. Stone and Peter W. F. Wilson

Originally published 12 Nov 2013 https://doi.org/10.1161/01.cir.0000437741.48606.98 Circulation 2014;129:S49–S73."

[ChatGPT] *"Goff DC Jr, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB Sr, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC Jr, Sorlie P, Stone NJ, Wilson PWF. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129:S49–S73. https://doi.org/10.1161/01.cir.0000437741.48606.98"*

Explaining diseases

[User input] *"I am 5 years old and I have tetralogy of Fallot. What does this mean?"*

[ChatGPT] *"Having tetralogy of Fallot means your heart is a bit different inside. Imagine your heart is like a house with 4 rooms, but one of the rooms has a wall that's not built right, and some of the doors are too small or in the wrong place. This makes it hard for your heart to send all the clean, fresh air (oxygen) your body needs to go everywhere it should.*

Doctors and nurses, who are like heart fixers, can help a lot. They know exactly what to do to make your heart work better. Sometimes they give you medicine, and sometimes they might do a special kind of fixing called surgery. This surgery is like fixing the walls and doors in the house, so everything works just right.

After the doctors help your heart, you'll be able to run, play, and do lots of fun things just like your friends. You might need to see the doctor more often than other kids, just to make sure your heart stays strong and happy.

Remember, having tetralogy of Fallot makes you unique, and with the right care, you can do almost anything you set your mind to!"

Conclusions

LLMs are powerful medical science writing tools from academic- and clinical-ease perspectives. They perform logical tasks such as formatting, editing, statistical analysis, and more complex functions such as summarizing articles, providing explanations, and guiding clinical assessment. They are limited in their novelty of writing, accuracy, and explorative analysis, but continuous development of these algorithms can help LLMs be a pocket assistant for physicians.

Ethics Statement

This research paper followed the appropriate ethical guidelines for science and publication.

Patient Consent

The authors confirm that patient consent is not applicable to this article.

Funding Sources

This work was supported by funds from the National Science Foundation (NSF number 2125872).

Disclosures

Dr Yanamala serves as an advisor for Turnkey Techstart. Dr Sengupta serves as an advisor for RCE Technologies. During the preparation of this work, the authors used ChatGPT to demonstrate examples of its use and edit our title. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. The section entitled "Example prompts and responses" has unedited ChatGPT responses to show authentic responses.

References

1. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Machine Learning Res 2011;12:2493-537.
2. Google AI. Bard. Available at: bard.google.com. Accessed February 7, 2024.
3. OpenAI: ChatGPT. Available at: https://chat.openai.com. Accessed February 7, 2024.
4. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023;330:866.
5. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Crit Care 2023;27:75.
6. King MR. The future of AI in medicine: a perspective from a chatbot. Ann Biomed Eng 2023;51:291-5.
7. Aydın Ö, Karaarslan E. OpenAI ChatGPT generated literature review: digital twin in healthcare. SSRN Electron J. Available at: https://www.ssrn.com/abstract = 4308687. Accessed February 7, 2024.
8. Frosolini A, Gennaro P, Cascino F, Gabriele G. In reference to "role of ChatGPT in public health," to highlight the AI's incorrect reference generation. Ann Biomed Eng 2023;51:2120-2.
9. Einarsson H, Lund SH, Jónsdóttir AH. Application of ChatGPT for automated problem reframing across academic domains. Comput Educ Artif Intell 2024;6:100194.
10. Donker T. The dangers of using large language models for peer review. Lancet Infect Dis 2023;23:781.
11. Liang W, Zhang Y, Cao H, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. Available at: https://arxiv.org/abs/2310.01783. Accessed February 7, 2024.
12. Ignjatović A, Stevanović L. Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: a descriptive study. J Educ Eval Health Prof 2023;20:28.
13. Calonge DS, Smail L, Kamalov F. Enough of the chit-chat: a comparative analysis of four AI chatbots for calculus and statistics. J Appl Learn Teach. Available at: https://journals.sfu.ca/jalt/index.php/jalt/article/view/1053. Accessed February 7, 2024.
14. Gala D, Makaryus AN. The utility of language models in cardiology: a narrative review of the benefits and concerns of ChatGPT-4. Int J Environ Res Public Health 2023;20:6438.
15. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23:bbac409.
16. Searle T, Ibrahim Z, Teo J, Dobson RJB. Discharge summary hospital course summarisation of in patient electronic health record text with

clinical concept guided deep pre-trained transformer models. J Biomed Inform 2023;141:104358.

17. Boonstra MJ, Weissenbacher D, Moore JH, Gonzalez-Hernandez G, Asselbergs FW. Artificial intelligence: revolutionizing cardiology with large language models. Eur Heart J 2024;45:332-45.

18. Alahdab F. Potential impact of large language models on academic writing. BMJ Evid Based Med 2024;29:201-2.

19. 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:e0000198.

20. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023;9:e48002.

21. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 2023;179:e160-5.

22. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT passes German state examination in medicine with picture questions omitted. Dtsch Ärztebl Int. Available at: https://www.aerzteblatt.de/10.3238/arztebl.m2023.0113. Accessed February 7, 2024.

23. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard against conventional drug-drug interactions clinical tools. Drug Healthc Patient Saf 2023;15:137-47.

24. Sim J, Huang X, Horan MR, et al. Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: a systematic review. Artif Intell Med 2023;146:102701.

25. Datta S, Lee K, Paek H, et al. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. J Am Med Inform Assoc 2024;31:375-85.

26. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024;34:2817-25.

27. Inojosa H, Gilbert S, Kather JN, Proschmann U, Akgün K, Ziemssen T. Can ChatGPT explain it? Use of artificial intelligence in multiple sclerosis communication. Neurol Res Pract 2023;5:48.

28. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023;25:e47479.

29. Rojulpote C, Bhattaru A, Jean C, et al. Effect of immunosuppressive therapy and biopsy status in monitoring therapy response in suspected cardiac sarcoidosis. JACC Cardiovasc Imaging 2022;15:1944-55.

30. Selvaraj S, Seidelmann SB, Soni M, et al. Comprehensive nutrient consumption estimation and metabolic profiling during ketogenic diet and relationship with myocardial glucose uptake on FDG-PET. Eur Heart J Cardiovasc Imaging 2022;23:1690-7.

31. Gritti MN, AlTurki H, Farid P, Morgan CT. Progression of an artificial intelligence chatbot (ChatGPT) for pediatric cardiology educational knowledge assessment. Pediatr Cardiol 2024;45:309-13.

32. Rogasch JMM, Metzger G, Preisler M, et al. ChatGPT: can you prepare my patients for [18 F]FDG PET/CT and explain my reports? J Nucl Med 2023;64:1876-9.

33. Skalidis I, Cagnina A, Fournier S. Use of large language models for evidence-based cardiovascular medicine. Eur Heart J Digit Health 2023;4:368-9.

34. Salihu A, Gadiri MA, Skalidis I, et al. Towards AI-assisted cardiology: a reflection on the performance and limitations of using large language models in clinical decision-making. EuroIntervention 2023;19:e798-801.

35. Henrickson L, Meroño-Peñuela A. Prompting meaning: a hermeneutic approach to optimising prompt engineering with ChatGPT. AI Soc. Available at: https://link.springer.com/10.1007/s00146-023-01752-8. Accessed February 7, 2024.

36. Poola I. Overcoming ChatGPTs inaccuracies with pre-trained AI prompt engineering sequencing process. Int J Technol Emerg Sci 2023;3:16-9.

37. Ekin S. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. Available at: https://www.techrxiv.org/doi/full/10.36227/techrxiv.22683919.v2. Accessed May 2023.

38. Microsoft. Copilot for Microsoft 365. Available at: https://www.microsoft.com/en-us/microsoft-365/business/copilot-for-microsoft-365. Accessed April 20, 2024.

39. Inam M, Sheikh S, Minhas AMK, et al. A review of top cardiology and cardiovascular medicine journal guidelines regarding the use of generative artificial intelligence tools in scientific writing. Curr Probl Cardiol 2024;49:102387.