

## REVIEW

# Large language models in health care: Development, applications, and challenges

Rui Yang<sup>1</sup> | Ting Fang Tan<sup>2</sup> | Wei Lu<sup>3</sup> | Arun James Thirunavukarasu<sup>4</sup>  | Daniel Shu Wei Ting<sup>2,5</sup> | Nan Liu<sup>5,6</sup> 

<sup>1</sup>Department of Biomedical Informatics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

<sup>2</sup>Singapore National Eye Center, Singapore Eye Research Institute, Singapore Health Service, Singapore, Singapore

<sup>3</sup>StatNLP Research Group, Singapore University of Technology and Design, Singapore

<sup>4</sup>University of Cambridge School of Clinical Medicine, Cambridge, UK

<sup>5</sup>Duke-NUS Medical School, Centre for Quantitative Medicine, Singapore, Singapore

<sup>6</sup>Duke-NUS Medical School, Programme in Health Services and Systems Research, Singapore, Singapore

## Correspondence

Nan Liu, Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore.

Email: [liu.nan@duke-nus.edu.sg](mailto:liu.nan@duke-nus.edu.sg)

## Funding information

None

## Abstract

Recently, the emergence of ChatGPT, an artificial intelligence chatbot developed by OpenAI, has attracted significant attention due to its exceptional language comprehension and content generation capabilities, highlighting the immense potential of large language models (LLMs). LLMs have become a burgeoning hotspot across many fields, including health care. Within health care, LLMs may be classified into LLMs for the biomedical domain and LLMs for the clinical domain based on the corpora used for pre-training. In the last 3 years, these domain-specific LLMs have demonstrated exceptional performance on multiple natural language processing tasks, surpassing the performance of general LLMs as well. This not only emphasizes the significance of developing dedicated LLMs for the specific domains, but also raises expectations for their applications in health care. We believe that LLMs may be used widely in preconsultation, diagnosis, and management, with appropriate development and supervision. Additionally, LLMs hold tremendous promise in assisting with medical education, medical writing and other related applications. Likewise, health care systems must recognize and address the challenges posed by LLMs.

## KEYWORDS

Large language model, AI, Health care

**Abbreviations:** AI, artificial Intelligence; BERT, Bidirectional Encoder Representations from Transformers; BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; CAD, computer-aided diagnosis; EHR, electronic health records; GPT, Generative Pretrained Transformer; LLaMA, large language model meta AI; LLMs, large language models; NLP, nature language processing; PaLM, Pathways Language Model; PMC, PubMed central; USMLE, United States Medical Licensing Examinations.

Rui Yang and Ting Fang Tan are joint-first authors.

Daniel Shu Wei Ting and Nan Liu are Joint-senior authors.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Health Care Science* published by John Wiley & Sons Ltd on behalf of Tsinghua University Press.

## 1 | INTRODUCTION

The field of natural language processing (NLP) has seen significant advances with the development of large language models (LLMs) trained by deep neural networks using massive text datasets, generally with billions of parameters. In 2017, Google first demonstrated “Transformer” architecture for the task of machine translation, which later attained state-of-the-art performance in many NLP tasks [1]. Since then, many LLMs with “Transformer” architecture have been developed, such as Bidirectional Encoder Representations from Transformers (BERT) [2], Generative Pretrained Transformer-3 (GPT-3) [3], Pathways Language Model (PaLM) [4], LLM Meta AI (LLaMA) [5], and GPT-4 [6].

In general, LLMs may be subcategorized based on their pre-training architecture as either encoder-only, decoder-only, or encoder-decoder; the pre-training tasks they undertake or the datasets utilized during their training phase [7]. As the size and computational resources used to train LLMs have increased, “zero-shot” or “few-shot” performance has experienced significant enhancements. When faced with new tasks, models are able to learn and accomplish these novel tasks without prior specialized training, simply by being shown zero or a few examples of these tasks [8]. The rapid development and outstanding performance of LLMs in NLP tasks may result in profound changes to health care work, although barriers to implementation must be overcome [9].

Already, NLP technology has been applied in health care to support preconsultation, diagnosis and management [10–12]. Development and proliferation of LLMs will improve NLP aptitude, and may therefore allow artificial intelligence (AI) to have an even greater impact on clinical care: changing how consultation, diagnosis, and management are conducted, and enhancing accessibility and autonomy by improving the provision of patient education through interactive dialogue with biomedical or clinical language models [13–18]. LLMs also have the potential to assist medical education for clinicians and administrative tasks such as writing letters, clinic notes, and discharge summaries [19–21].

In this review, we provide an overview of the development of LLMs designed for biomedical or clinical use. We subsequently explore the potential and trialed applications of LLMs in clinical contexts. Finally, we outline the challenges and limitations which must be addressed to ensure that LLM technology realizes its clinical potential.

## 2 | DEVELOPMENT OF LLMs IN HEALTH CARE

Although LLMs have shown impressive performance across a range of NLP tasks, their efficacy in specialized tasks is limited [22]. A lack of domain-specific knowledge in general LLMs hinders their ability to interpret technical terms and produce accurate, reasoned answers. Moreover, there are significant differences between general corpora and professional corpora, which further hinder the ability of LLMs to perform well in biomedical or clinical tasks [23]. To improve domain-specific performance by addressing these weaknesses, specialized LLMs have been developed.

BERT for Biomedical Text Mining (BioBERT) was built using a large biomedical corpus of PubMed abstracts and PubMed Central full-text articles for fine-tuning [13]. SCIBERT was trained from scratch on the Semantic Scholar corpora (18% computer science papers and 82% biomedical papers), rather than fine-tuning the generalist BERT model [15]. PubMedBERT was developed through a similar schema to SCIBERT, but using corpora sourced entirely from PubMed [16]. Compared to BERT, BioBERT, SCIBERT and PubMedBERT demonstrated superior performance in biomedical NLP tasks. However, clinical use cases of these biomedical language models are limited.

ClinicalBERT was created based upon BERT and BioBERT architectures [14] and trained on the MIMIC-III data set [24]. MIMIC-III comprises demographics, vital signs, laboratory tests, procedures, medications, clinical notes, investigation reports, and mortality data corresponding to over 40,000 critical care patients—a rich source of domain-specific information [24]. ClinicalBERT attained superior performance to BERT and BioBERT across a range of medical NLP tasks, demonstrating the promise of using clinical corpora to fine-tune LLMs to optimize domain-specific performance [14]. In addition, GatorTron, the largest clinical language model available, was trained from scratch using over 90 billion words of text from the deidentified clinical notes of University of Florida Health, PubMed articles and Wikipedia [25]. This model increases the parameter count of LLMs within the clinical domain from 110 million (ClinicalBERT) to 8.9 billion. It has achieved competitive performance across multiple downstream clinical tasks, demonstrating the advantage of using large “Transformer” models.

Moreover, the emerging conversational abilities of LLMs have fostered innovation in health care. The

performance of PaLM in medical questions was optimized through instruction prompt tuning to develop Med-PaLM [26] (ChatGPT-like ChatBot for Health care). Subsequent development at Google has resulted in the production of Med-PaLM 2 [27], which reportedly achieves state-of-the-art performance in United States Medical Licensing Examinations (USMLE) questions, exceeding the performance of ChatGPT [20]. ChatDoctor (an open-source Chatbot for health care) was based on LLaMA and used 100,000 patient-physician conversations to fine-tune [28]; this model showed significant improvements in understanding patients' needs and providing accurate advice. Similarly, Baize-health care [29] is another open-source Chatbot for health care based on LLaMA, which has been fine-tuned using MedQuAD data set [30] (including 46,867 medical dialogues) and performed well in multi-turn conversations. These models will facilitate the further development of conversation models in health care.

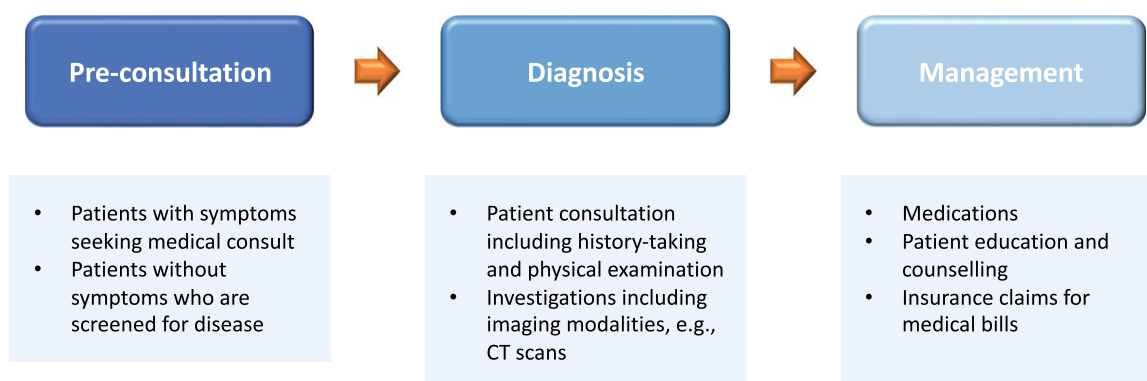
While the performance of general LLMs is impressive, and advancing rapidly, domain-specific models may remain optimal for specialized tasks. Rather than training domain-specific models from the ground up, further research may seek to fine-tune or prompt-tune these general LLMs to optimize performance in particular tasks. Using larger open-source base models and newer interactive LLMs could further improve the capabilities of decentralized researchers around the world, who could then fine-tune LLMs to optimize performance for clinical work. Through bespoke fine-tuning, domain-specific LLMs may be produced to serve narrowly defined, well-specified tasks—minimizing error and maximizing clinical utility. Whether developed from scratch or fine-tuned using existing models, LLM applications will become more sophisticated and begin to impact patients and practitioners at scale.

### 3 | APPLICATIONS OF LLMs IN HEALTH CARE

A typical patient journey in health care (as outlined in Figure 1) includes: (1) Preconsultation: where patients register for medical consult or undergo health screening; (2) Diagnosis: which includes patient consultation and examination as well as adjunctive investigations; and (3) Management: which includes medications, patient counseling and education, and reimbursements for medical bills. LLMs show promise to enhance the patient experience at each of these touch points in the patient journey.

#### 3.1 | Preconsultation

To cope with the exponential increase in patient load, LLMs can facilitate triaging patients and directing efficient use of resources. By combining with cloud services, a cloud-based intelligent self-diagnosis and department recommendation service built on LLM was able to predict the possible disease categories based on patients' history of presenting complaints, and then recommended the specific medical subspecialty for patients to book a doctor's appointment [17]. This provided a recommendation framework for patients to seek appropriate medical review, and minimized wastage of medical resources. Furthermore, LLMs can also be integrated with remote diagnosis systems or telehealth services to enhance access to care for patients facing geographical barriers or mobility limitations. LLMs can serve as real-time self-assessment tools via SMS or other platforms on mobile devices for individuals in remote settings like sub-Saharan Africa [12]. This tool provided timely access to health care by assessing symptoms of tropical diseases, suggesting a likely diagnosis, and providing medical advice. In addition to assessing



**FIGURE 1** Potential touch points along a patient's care journey for the application of large language models.

patients' symptoms, LLMs can improve predictions in the context of patients' medical background by harnessing additional information like comorbidities, risks factors and medication lists.

### 3.2 | Diagnosis

Patient consultation, especially the first encounter, is important not only in the diagnostic process but essential in laying the foundation of the patient-physician relationship. However, the reality of clinical practice is that appointments are often overbooked with minimal consultation time for each patient. Furthermore, each consultation consists of reviewing electronic health records (EHR), history-taking, physical examination, and patient counseling. Therefore LLMs can be leveraged to generate concise summaries of each patient's medical background including comorbidities, previous consultations or admissions, medication lists, past treatment progress and response [11, 18, 25]. This synthesizes and brings to the physician's attention relevant patient information that may be essential in disease diagnosis. This also facilitates a more efficient and comprehensive consultation, allowing the physician to dedicate more time for patient interaction and lead to greater patient satisfaction.

In the diagnostic process, adjunctive investigations are often required to support clinical suspicion and confirm the diagnosis. LLMs can serve as clinical decision support systems to guide physicians to select the most appropriate radiological investigations in specific clinical scenarios [18]. This can minimize unnecessary burden on existing resource capacity, especially in public hospitals or low-resource settings, where imaging modalities and technical support are limited. Moreover, patients who may not require contrast computed tomography scans, for example, can avoid having unnecessary radiation or contrast exposure.

There have been increasing applications of deep learning models utilized to build computer-aided diagnosis (CAD) systems to automate efficient and highly accurate interpretation of these medical imaging modalities for disease detection and classification. However, most clinicians remain skeptical and hesitant to adopt these into real-world clinical practice, attributed to the opaque model decision-making processes. By incorporating LLMs into these CAD systems, clinicians can ask open-ended questions about specific input images or patients to understand the rationale of the CAD's decision output [18]. This human-computer interaction could potentially enhance model interpretability and encourage uptake into existing diagnostic workflows.

Furthermore, clinicians may even uncover new insights or imaging biomarkers of disease in the process.

### 3.3 | Management

Optimal management requires a multidisciplinary approach with an increasing emphasis on people-centered care to empower patients to participate and take ownership of their own health [31]. Medication compliance is a significant challenge affecting treatment effectiveness, often attributed to forgetfulness and poor patient insight. Engaging patients through patient education initiatives may boost compliance and encourage patients to be responsible for their own health. LLMs may enable tailored patient education to improve understandability and engagement. Recently, Macy, an AI pharmacist comprising a photorealistic animated avatar incorporating ChatGPT, was capable of delivering medication counseling via video. This was structured in plain-language and included essential information like medication dosage and frequency, precautions and red flag symptoms to watch out for. While solely experimental, it was developed in under 30-min and at no significant cost, demonstrating the potential of LLMs to revolutionize patient education and beyond [32].

LLMs can also generate educational content at an appropriate level for patients, such as postprocedure counseling, medication counseling, and lifestyle modifications [33]. This allows complex medical terminology to be communicated effectively and appropriately to patients in simple terms to facilitate understanding. For example, LLMs can be used for autocomplete text simplification tasks, to translate jargon-heavy medical reports or explanations into simplified sentences by prompting simple words to follow what has been typed by the physician [34]. This expedites the process of text simplification while preserving control of the information translated, allowing the physician to ensure quality and accuracy of the information communicated. LLMs can also efficiently automate multilingual translations to cater to a wider diversity of patients. Another example would be AnsibleHealth, a virtual clinic for chronic lung diseases, which explored the use of ChatGPT to simplify radiology reports and jargon-heavy medical records to facilitate patient comprehension [19].

Managing patients with mental health conditions is challenging, requiring multimodal and multidisciplinary approaches. LLMs can potentially be effective in addressing the clinical need for access to psychiatric care services and treatment, supplementing the shortage of health care professionals and enhancing patient compliance [35]. For example, Woebot is a fully automated text-



based conversational assistant that delivers cognitive behavior therapy services for adolescents with depression. Woebot significantly reduced symptoms of depression, compared to the control group who was given information-only e-book materials [35, 36]. SERMO is another conversational tool that guides patients with mental health conditions in regulating their emotions to better handle negative thoughts [37]. It automatically detects the type of emotion based on user text inputs, and recommends mindfulness activities or exercises tailored to the specific emotions.

Furthermore, LLMs have the potential to streamline administrative processes to increase efficiency while reducing the administrative burden on physicians and enhancing patient experience. This can encompass drafting discharge summaries and operation reports, extracting succinct clinical information from EHR to complete medical reports and translating them into billable codes for reimbursement claims, as well as automating responses to general patient queries (e.g., requests for medication top-up, appointment booking and rescheduling) [38–40].

### 3.4 | Medical education and medical writing

In addition to health care applications from the patient perspective, LLMs hold immense potential in reshaping medical education and research. Existing LLMs have been able to pass undergraduate and postgraduate medical examinations [19, 41, 42]. Moreover, answers generated by ChatGPT to USMLE were accompanied by justifications that have a high level of concordance and offered new insights [17, 31]. The logical flow of explanations and deductive reasoning with additional supplementary information provided allows students to easily follow and comprehend. For example, this can be targeted at an undergraduate medical student who may have answered the question incorrectly, to uncover new perspectives or remedial knowledge from the ChatGPT-generated explanations. ChatGPT can also suggest innovative and unique mnemonics to aid memorizing. The interactive interface of LLMs can complement existing student-directed learning, where Socratic style of teaching has been surveyed as preferable by students over didactic lectures [20, 43, 44].

LLMs can also add value to medical research. LLMs can improve the efficiency of research article writing by automating tasks such as literature review, generating text and guiding manuscript writing style and formatting [44]. Biswas recently published a perspective piece that was written by ChatGPT, though still requiring editing by

a human author [45, 46]. LLMs can also match patients to potential clinical trial opportunities relevant to patients' conditions and within inclusion and exclusion criteria. This can facilitate research patient recruitment, while enabling access to potentially breakthrough treatments that may not be otherwise available or affordable for patients [45, 46].

## 4 | CHALLENGES OF LLMs IN HEALTH CARE

### 4.1 | Data privacy

One of the challenges in validation and implementation of LLMs with real-world clinical patient data would be the risk of leaking confidential and sensitive patient information. For example, adversarial attacks on a LLM GPT-2 were successful in extracting the model's training data [47, 48]. By querying GPT-2 structured questions, training data including personal identifiable information and internet relay chat conversations were extracted verbatim. Moreover, despite anonymizing sensitive patient health information, some algorithms demonstrated the capability to reidentify these patients [49–51]. To mitigate these challenges, possible strategies include pseudonymization or filtering patient identifiers, differential privacy, and auditing of LLMs using data extraction attacks [47, 48, 52].

### 4.2 | Questionable credibility and accuracy of information

Some have criticized LLMs for the questionable credibility and accuracy of information generated. Open domain nonspecific LLMs may be at risk of perpetuating inaccurate information from open internet sources, or generalize poorly across different contextual settings [47, 48, 52, 53]. The term “hallucination effect” has been used to describe trivial guessing behaviors observed in LLMs [54]. For example, an experiment using GPT-3.5 to answer sample medical questions from USMLE, found that the model often predicted options A and D. In the ChatGPT-generated perspective article, three fabricated citations were identified during editing by the human author [45]. This may potentially be hazardous to users who are unable to discern seemingly credible but inaccurate or misleading answers. Despite its potential as an educational tool and source of information for patients, medical students, and the research community, human oversight and additional quality measures are essential in ensuring accuracy and quality control of the generated content.

### 4.3 | Data bias

LLMs are commonly trained on vast and diverse data which is often biased. Consequently, the content generated by LLMs may perpetuate and even amplify bias, such as ethnicity, gender, and socioeconomic background [55]. These biases are especially problematic in health care, where differential treatment may lead to exacerbation of disparities in mortality and morbidity. For example, a study focusing on skin cancer may predominantly involve participants with fair-skinned individuals, resulting in an LLM that is less adept at identifying skin cancer in those with darker-skinned individuals. This could lead to misdiagnosis and delayed or inappropriate treatments, further widening health disparities. The absence of minority groups in training data may make LLMs exacerbate these biases, leading to inaccurate results. Moving towards fair artificial intelligence and combating bias will be a significant challenge for LLMs [56].

### 4.4 | Interpretability of LLMs

The lack of interpretability of the decision-making process of LLMs remains a barrier to adoption into clinical practice [57]. LLM-generated responses are largely not accompanied by justifications or supporting information sources. This is further exacerbated by the tendency of LLMs to fabricate facts in a seemingly confident manner or rely on trivial guessing, as elaborated above. In the context of safety-critical tasks in health care, this may limit trust and acceptance by physicians and patients, where the consequences of delivering inaccurate medical advice may be detrimental. Proposed methods to improve interpretability include a selection inference multi-step reasoning framework by Creswell et al. to generate a series of casual reasoning steps toward the final generated response [58]. Another method proposed leveraging ChatGPT using chain-of-thought prompting (i.e. step-by-step instructions) [59] for knowledge graph extraction, where extracted entities and relationships from the raw input text are presented in a structured format, which was then used to train an interpretable linear model for text classification [60]. Uncertainty-aware LLM applications may be another useful feature, where differential weights of input data or reporting confidence scores of generated responses can enhance the trust in proposed LLM applications [61].

### 4.5 | Roles of LLMs

Another challenge LLMs may face lies in defining its role and identity in scientific research and clinical practice

[62]. Questions that may arise include: Can AI be a researcher or a physician? Can the AI be responsible for the content it generates? How to distinguish the text generated by AI versus humans? What to do when physicians have different views than AI? It is worth noting that LLMs may fabricate false content, so it is necessary to avoid overusing [55]. From preconsultation to diagnosis to treatment, or in medical education, medical research, LLMs can serve in complementary roles rather than substitutes for physicians. Although LLMs can undergo self-improvement, physician oversight is still required to ensure the generated content is accurate and clinically relevant.

### 4.6 | Deployment of LLMs

LLaMA's open source facilitates the deployment of LLMs on resource-constrained devices, such as laptops, phones, and Raspberry Pi systems [5]. Alpaca's fine-tuning based on LLaMA, enables the rapid (within hours) and cost-effective (costing under US\$600) development of models that exhibit performance comparable to that of GPT-3.5 [63]. This makes it possible to train personalized language models with high performance at a reduced cost, but it is important to recognize that these models also inherit various biases. When applied to general purposes, they may generate harmful or sensitive content, potentially compromising user security. Furthermore, the ease of deployment may increase the likelihood of LLMs being misused or even maliciously trained to disseminate deeply falsified information and detrimental content. Such outcomes could undermine public trust in AI and have deleterious effects on the whole society. To ensure that LLMs are harnessed for their intended purposes and to reduce the risks associated with their misuse, it is crucial to develop and implement various safeguards. These may include technical solutions for filtering out sensitive and harmful content, the establishment of stringent terms of use and deployment specifications. By adopting such measures, the potential dangers of deploying LLMs on small personal devices can be effectively controlled.

### 4.7 | Clinical domain-specific LLMs

There is no doubt that LLMs are having a significant impact in the health care field. Regardless of whether it is preconsultation, diagnosis, management, medical education, or medical writing, all these areas will undergo transformative changes due to the development of LLMs. In this regard, it is essential to recognize that when LLMs

are deployed in real clinical settings, different medical specialties will encounter a variety of unique challenges [64]. For example, the type and quality of data may differ significantly between domains. Additionally, the diverse application scenarios and tasks of LLMs will lead to inconsistencies in the standards expected by clinical professionals. In light of this, when deploying LLMs in clinical environments, we should recognize the variations across clinical specialties and make appropriate adjustments according to the specific application scenarios.

## 5 | CONCLUSION

LLMs are poised to bring about significant transformation in health care and will be ubiquitous in this field. To make LLMs more serviceable for health care, training from scratch with medical databases or fine-tuning with generic LLMs would be effective approaches. Besides, LLMs can further perform multi-modal feature fusion with diverse data sources, including image data and tabular data, resulting in better performance, even beyond human level. While the use of LLMs presents numerous benefits, we should recognize that LLMs cannot take full responsibility for generated content. It is essential to ensure that AI-generated content is properly reviewed to avoid any potential harm. As the threshold for the deployment of LLMs diminishes, improving deployment specifications also deserves attention. Simultaneously, efforts should be made to promote the integration of LLMs in clinical practice, improve the interpretability of LLMs in clinical settings and enhance human-machine collaboration to better support clinical decision-making. By leveraging LLMs as a complementary tool, physicians can maximize the benefits of AI while mitigating potential risks and achieve better clinical outcomes for patients. Ultimately, the successful integration of LLMs in health care will require the collaborative efforts of physicians, data scientists, administrators, patients, and regulatory bodies.

### AUTHOR CONTRIBUTIONS

**Rui Yang:** Conceptualization (equal); writing—original draft (equal); writing—review and editing (equal). **Ting Fang Tan:** Conceptualization (equal); writing—original draft (equal); writing—review and editing (equal). **Wei Lu:** Conceptualization (equal); writing—review and editing (equal). **Arun James Thirunavukarasu:** Conceptualization (equal); writing—review and editing (equal). **Daniel Shu Wei Ting:** Conceptualization (equal); supervision (lead); writing—review and editing

(equal). **Nan Liu:** Conceptualization (lead); project administration (lead); supervision (lead); writing—original draft (supporting); writing—review and editing (equal).

### ACKNOWLEDGMENTS

Not applicable.

### CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

### DATA AVAILABILITY STATEMENT

Not applicable.


### ETHICS STATEMENT

Not applicable.

### INFORMED CONSENT

Not applicable.

### ORCID

Arun James Thirunavukarasu  <http://orcid.org/0000-0001-8968-4768>

Nan Liu  <https://orcid.org/0000-0003-3610-4883>

### REFERENCES

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:1–11. [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
2. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. <https://doi.org/10.48550/arXiv.1810.04805>
3. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–901. <https://doi.org/10.48550/arXiv.2005.14165>
4. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. *arXiv:2204.02311*. 2022. <http://arxiv.org/abs/2204.02311>
5. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. LLaMA: open and efficient foundation language models. 2023. <http://arxiv.org/abs/2302.13971>
6. OpenAI. GPT-4 Technical Report. 2023. <http://arxiv.org/abs/2303.08774>
7. Amatriain X. Transformer models: an introduction and catalog. 2023. <http://arxiv.org/abs/2302.07730>
8. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. 2020. <http://arxiv.org/abs/2001.08361>
9. Zhavoronkov A. Caution with AI-generated content in biomedicine. *Nature Med*. 2023;29(3):532. <https://doi.org/10.1038/d41591-023-00014-w>

10. He Y, Zhu Z, Zhang Y, Chen Q, Caverlee J. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. 2020. <https://doi.org/10.48550/arXiv.2010.03746>
11. Li C, Zhang Y, Weng Y, Wang B, Li Z. Natural language processing applications for Computer-Aided diagnosis in oncology. *Diagnostics*. 2023;13(2):286. <https://doi.org/10.3390/diagnostics13020286>
12. Omoregbe NAI, Ndaman IO, Misra S, Abayomi-Alli OO, Damaševičius R. Text Messaging-Based medical diagnosis using natural language processing and fuzzy logic. *J Healthc Eng*. 2020;2020(4):1–14. <https://doi.org/10.1155/2020/8839524>
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–40. <https://doi.org/10.1093/bioinformatics/btz682>
14. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019. <https://doi.org/10.48550/arXiv.1904.03323>
15. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019. <https://doi.org/10.18653/v1/d19-1371>
16. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthcare*. 2022;3(1):1–23. <https://doi.org/10.1145/3458754>
17. Wang J, Zhang G, Wang W, Zhang K, Sheng Y. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. *Journal of Cloud Computing*. 2021;10:1–12. <https://doi.org/10.1186/s13677-020-00218-2>
18. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;307(2):230163. <https://doi.org/10.1148/radiol.230163>
19. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health*. 2023;2(2):e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
20. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. <https://doi.org/10.2196/45312>
21. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology*. 2023;307(2):230171. <https://doi.org/10.1148/radiol.230171>
22. Thirunavukarasu A, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) with Applied Knowledge Test questions: what are the opportunities and limitations of artificial intelligence chatbots in primary care? (Preprint). 2023. <https://doi.org/10.2196/preprints.46599>
23. Lei L, Liu D. A new medical academic word list: a corpus-based study with enhanced methodology. *J English Acad Purp*. 2016;22:42–53. <https://doi.org/10.1016/j.jeap.2016.01.008>
24. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. <https://doi.org/10.1038/sdata.2016.35>
25. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. *npj digital Medicine*. 2022;5(1):1–9. <https://doi.org/10.1038/s41746-022-00742-2>
26. Med-PaLM. Med-PaLM [Internet]. Available from: <https://sites.research.google/med-palm/>
27. Matias Y. Our latest health AI research updates. Google [Internet]. Available from: <https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/>
28. Li Y, Li Z, Zhang K, Dan R, Zhang Y. ChatDoctor: a medical chat model fine-tuned on LLaMA model using medical domain knowledge. 2023. <http://arxiv.org/abs/2303.14070>
29. Xu C, Guo D, Duan N, McAuley J. Baize: an open-source chat model with parameter-efficient tuning on self-chat data. 2023. <http://arxiv.org/abs/2304.01196>
30. Ben Abacha A, Demner-Fushman D. A question-entailment approach to question answering. *BMC Bioinformatics*. 2019;20(1):511. <https://doi.org/10.1186/s12859-019-3119-4>
31. World Health Organization. WHO global strategy on people-centred and integrated health services: interim report. World Health Organization; 2015. <https://apps.who.int/iris/handle/10665/155002>
32. Kenneth Leung on LinkedIn. Available from: [https://www.linkedin.com/posts/kennethleungty\\_generativeai-ai-pharmacist-activity-7031533843429949440-pVZb](https://www.linkedin.com/posts/kennethleungty_generativeai-ai-pharmacist-activity-7031533843429949440-pVZb)
33. Bala S, Keniston A, Burden M. Patient perception of Plain-Language medical notes generated using artificial intelligence software: pilot Mixed-Methods study. *JMIR Formative Research*. 2020;4(6):e16670. <https://doi.org/10.2196/16670>
34. Van H, Kauchak D, Leroy G. AutoMeTS: the autocomplete for medical text simplification. 2020. <https://doi.org/10.48550/arXiv.2010.10573>
35. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *Canadian J Psychi*. 2019;64(7):456–64. <https://doi.org/10.1177/0706743719828977>
36. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment Health*. 2017;4(2):e19. <https://doi.org/10.2196/mental.7785>
37. Denecke K, Vaaheesan S, Arulnathan A. A mental health chatbot for regulating emotions (SERMO)—concept and usability test. *IEEE Transact Emerg Topics Comput*. 2021;9:1170–82. <https://doi.org/10.1109/tetc.2020.2974478>
38. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Semin Ophthalmol*. 2023;38(5):503–7. <https://doi.org/10.1080/08820538.2023.2209166>
39. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digital Health*. 2023;5(3):e107–8. [https://doi.org/10.1016/S2589-7500\(23\)00021-3](https://doi.org/10.1016/S2589-7500(23)00021-3)
40. Insights CB. How artificial intelligence is reshaping medical billing & insurance. CB Insights Research [Internet].



- Available from: <https://www.cbinsights.com/research/artificial-intelligence-healthcare-providers-medical-billing-insurance/>
41. Varanasi L. AI models like ChatGPT and GPT-4 are acing everything from the bar exam to AP Biology. Here's a list of difficult exams both AI versions have passed. 2023. Website. <https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1>
  42. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. 2022. <https://doi.org/10.48550/arxiv.2212.13138>
  43. Burk-Rafel J, Santen SA, Purkiss J. Study behaviors and USMLE step 1 performance: implications of a student Self-Directed parallel curriculum. *Acad Med*. 2017;92:S67–74. <https://doi.org/10.1097/ACM.0000000000001916>
  44. Abou-Hanna JJ, Owens ST, Kinnucan JA, Mian SI, Kolars JC. Resuscitating the socratic method: student and faculty perspectives on posing probing questions during clinical teaching. *Acad Med*. 2021;96(1):113–7. <https://doi.org/10.1097/ACM.0000000000003580>
  45. Biswas S. ChatGPT and the future of medical writing. *Radiology*. 2023;307(2):223312. <https://doi.org/10.1148/radiol.223312>
  46. BuildGreatProducts.club. The Potential of Large Language Models(LLMs) in Healthcare: Improving Quality of Care and Patient Outcomes. In: Medium [Internet]. Available from: <https://medium.com/@BuildGP/the-potential-of-large-language-models-in-healthcare-improving-quality-of-care-and-patient-6e8b6262d5ca>
  47. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting training data from large language models. 2020. <https://doi.org/10.48550/arXiv.2012.07805>
  48. Yang X, Lyu T, Li Q, Lee C-Y, Bian J, Hogan WR, et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Med Inform Decis Mak*. 2019;19(Suppl 5):232. <https://doi.org/10.1186/s12911-019-0935-4>
  49. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science*. 2013;339(6117):321–4. <https://doi.org/10.1126/science.1229566>
  50. Na L, Yang C, Lo C-C, Zhao F, Fukuoka Y, Aswani A. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. *JAMA Network Open*. 2018;1(8):e186040. <https://doi.org/10.1001/jamanetworkopen.2018.6040>
  51. Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. *Science*. 2018;362(6415):690–4. <https://doi.org/10.1126/science.aau4832>
  52. Du L, Xia C, Deng Z, Lu G, Xia S, Ma J. A machine learning based approach to identify protected health information in Chinese clinical text. *Int J Med Inform*. 2018;116:24–32. <https://doi.org/10.1016/j.ijmedinf.2018.05.010>
  53. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med*. 2021;13(586):eabb1655. <https://doi.org/10.1126/scitranslmed.abb1655>
  54. OpenAI. ChatGPT: Optimizing Language Models for Dialogue. In: OpenAI [Internet]. Available from: <https://openai.com/blog/chatgpt/>
  55. Volovici V, Syn NL, Ercole A, Zhao JJ, Liu N. Steps to avoid overuse and misuse of machine learning in clinical research. *Nature Med*. 2022;28(10):1996–9. <https://doi.org/10.1038/s41591-022-01961-6>
  56. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. *Patterns*. 2021;2(10):100347. <https://doi.org/10.1016/j.patter.2021.100347>
  57. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE transactions on neural networks and learning systems*. 2021;32(11):4793–813. <https://doi.org/10.1109/TNNLS.2020.3027314>
  58. Creswell A, Shanahan M, Higgins I. Selection-Inference: exploiting large language models for interpretable logical reasoning. 2022. <http://arxiv.org/abs/2205.09712>
  59. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022. <http://arxiv.org/abs/2201.11903>
  60. Shi Y, Ma H, Zhong W, Mai G, Li X, Liu T, et al. ChatGraph: interpretable text classification by converting ChatGPT knowledge to graphs. 2023. <http://arxiv.org/abs/2305.03513>
  61. Youssef A, Abramoff M, Char D. Is the algorithm good in a bad world, or has it learned to be bad? The ethical challenges of “locked” versus “continuously learning” and “autonomous” versus “assistive” AI tools in healthcare. *Am J Bioeth*. 2023;23(5):43–5. <https://doi.org/10.1080/15265161.2023.2191052>
  62. Liebrezn M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digital Health*. 2023;5(3):e105–6. [https://doi.org/10.1016/S2589-7500\(23\)00019-5](https://doi.org/10.1016/S2589-7500(23)00019-5)
  63. Stanford CRFM. Alpaca: a strong, replicable instruction-following model. Available from: <https://crfm.stanford.edu/2023/03/13/alpaca.html>
  64. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47(1):33. <https://doi.org/10.1007/s10916-023-01925-4>

**How to cite this article:** Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. *Health Care Sci*. 2023;2:255–263. <https://doi.org/10.1002/hcs2.61>