Review

For reprint orders, please contact: reprints@futuremedicine.com

Future Medicine AI

Unlocking the potential of large language models in healthcare: navigating the opportunities and challenges

Idit Tessler[1,2], Tzahi Yamin[1], Hadar Peeri[2], Eran E Alon[1], Eyal Zimlichman[2], Benjamin S Glicksberg*[,3] & Eyal Klang[2,3]

[1]Department of Otolaryngology - Head & Neck Surgery, Sheba Medical Center, affiliated to Tel Aviv University, Tel Aviv, Israel
[2]The Sagol AI Hub, ARC Innovation Center, Sheba Medical Center, affiliated to Tel Aviv University, Israel
[3]The Division of Data Driven & Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
*Author for correspondence: benjamin.glicksberg@mssm.edu

This paper explores the emerging role of large language models (LLMs) in healthcare, offering an analysis of their applications and limitations. Attention mechanisms and transformer architectures enable LLMs to perform tasks like extracting clinical information and assisting in diagnostics. We highlight research that demonstrates early application of LLMs in various domains and along the care pathway. With their promise, LLMs pose ethical and practical challenges, including data bias and the need for human oversight. This review serves as a guide for clinicians and researchers, outlining potential healthcare applications – ranging from document translation to clinical decision support – while cautioning about inherent limitations and ethical considerations. The aim of this work is to encourage the knowledgeable use of LLMs in healthcare and drive further study in this important emerging field.

Tweetable abstract: This review serves as a guide for clinicians and researchers, outlining potential healthcare applications ranging from document translation to clinical decision support while cautioning about inherent limitations and ethical considerations.

First draft submitted: 4 January 2024; Accepted for publication: 2 April 2024; Published online: 24 April 2024

Keywords: AI • artificial intelligence • attention mechanism • GPT • healthcare • large language model • LLM • natural language processing • NLP • transformer

Artificial intelligence (AI) is in the process of transforming healthcare [1,2], and large language models (LLMs) are gaining fast recognition in this space. These models are trained on large data and can understand patterns of language. They use advanced techniques like attention mechanisms and transformer architectures [3,4]. LLMs are good at creating human-like text and natural language processing (NLP) tasks. They can help in clinical tasks like extracting information from health records, personalizing treatment plans and helping with making diagnoses [1,2]. However, using LLMs in healthcare raises ethical and practical concerns like bias in the training data, the need for human oversight and the impact on healthcare professionals [1]. It is important for clinicians to understand the potential and limitations of LLMs [5].

The rising interest in LLMs for healthcare is already reflected in a marked increase in related publications over the past year (Figure 1). In this review, we will give an overview of LLMs and their potential uses in medicine. We will also discuss the limitations and practical considerations of using LLMs. Our aim with this work is to encourage the knowledgeable use of LLMs in healthcare and drive further study in this important emerging field.

Natural language processing

NLP is a branch of AI that helps computers understand human language [6,7]. It is used in different applications such as information retrieval, question answering, sentiment analysis and speech recognition [6,7]. In healthcare, NLP can extract useful information from unstructured data sources. This can help with tasks such as creating summary

10.2217/fmai-2024-0001   eISSN 2753-8087

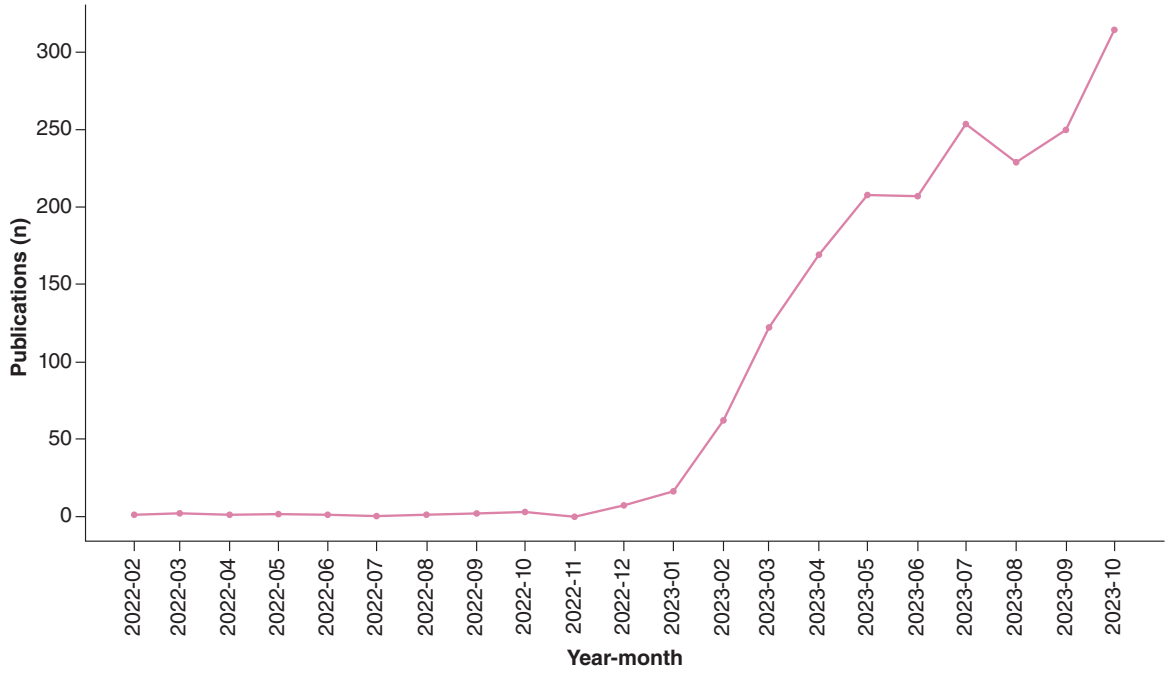

Figure 1. PubMed trend of publications on large language models and related technologies in healthcare. The figure represents a time-series plot of the number of publications per month, as identified in PubMed. Search terms include 'large language models', 'ChatGPT', 'OpenAI', 'Google Gemini', 'Microsoft Bing' or 'BERT' in the title or abstract.

reports, making decisions and identifying patterns in patient data [8–10]. NLP can also be used to personalize treatments, assist with diagnoses and provide best practice recommendations [1,10].

Attention mechanism

The attention mechanism is a key component used in many LLMs. It allows the model to focus on specific parts of the input when processing language [3,11]. For example, in the sentence 'Review of the patient's history shows diabetes management complexities, primarily due to progressive insulin resistance and adherence to prescribed medication regimen' in a patient's electronic health record (EHR), the attention mechanism might closely associate 'diabetes' with 'insulin resistance' [3]. This is somewhat similar to how the human brain uses the prefrontal cortex to focus attention on stimuli [12].

The function applies weights to input vectors using techniques such as dot-product and multiheaded attention, then sums these products to produce the output of the attention mechanism [11]. The choice of attention function can affect the performance of the LLM [13]. The attention mechanism also has limitations, such as the potential for attention biases based on the training data.

Transformer

The transformer is a type of artificial neural network model that was introduced in 2017 [3]. The model has gained widespread interest in NLP tasks such as language translation and language generation [3,14]. The model is composed of stacked layers of attention mechanism. The mechanism allows the model to selectively attend to different parts of the input when processing language [3,15].

The transformer has been used in a variety of applications. Examples include machine translation, language generation and language modeling [16]. The model is demanding in terms of size and data requirements. It typically requires a large amount of data and computational resources to train [14]. For example, the original transformer model used in the paper by Vaswani *et al.* was trained on a dataset of approximately 8 million sentences and required 4 days of training on eight graphics processing units to reach convergence [3]. The model had six layers in the encoder and six layers in the decoder [3]. Subsequent versions of the transformer, such as Transformer-XL and GPT-3, have been trained on much larger datasets and have achieved better performance on various NLP tasks [17,18].

LLMs are trained using unsupervised learning. This means that they are fed large amounts of unlabeled data and learn to identify patterns in the data without explicit supervision. Therefore, the quality of the training data can significantly impact the model's performance [19,20]. Other technical details to consider when using the transformer include the architecture of the model, which determines the number of layers and the type of attention mechanism [2].

It is also important to consider the training and fine-tuning process when using the transformer. LLMs are typically pretrained on large datasets and then fine-tuned on specific tasks or datasets [16]. The fine-tuning process involves retraining to improve performance on a specific dataset [16]. Careful consideration should be made for the appropriate training and fine-tuning process. Each specific application should be carefully evaluated for the performance of the model on the relevant data [19].

Generative pretraining transformer

Generative pretraining transformer (GPT) is an LLM developed by the company OpenAI. It performs extremely well on various NLP tasks like language translation and generation [18]. The model is composed of a multi-headed attention mechanism and was reportedly trained on a dataset of 410 billion tokens (words) [18].

One of the key features of GPT is its ability to perform few-shot and zero-shot learning. This means it can learn a new task with only a small amount of input data [21]. This is a challenging task in machine learning, but GPT's large size and transformer architecture allow it to transfer knowledge across a wide range of tasks [18]. Additionally, GPT's diverse training dataset of tasks and domains further improves its ability to perform few-shot learning [18,21].

GPT has been used in various applications, including language translation, language generation and question-answering [18]. It has also been used in healthcare tasks such as generating summary reports from EHRs and identifying potential adverse drug events (ADEs) [10].

GPT-3 is an example of a very large transformer model, reportedly with up to 175 billion weights [18]. It requires significant computational resources to train and fine-tune. Thus, it is typically accessed through a cloud-based application programming interface rather than being deployed on a local machine [18,21].

Following GPT-3, OpenAI introduced GPT-3.5 as an interim model, which improved upon its predecessor with refined training techniques and more data. GPT-4, the subsequent iteration, marked a significant leap, incorporating multimodal abilities that enable it to process both text and images, thereby expanding its utility in diverse applications, including more complex medical scenarios. GPT-4 has also been architected to be more reliable and less prone to generating harmful outputs. Moreover, OpenAI released GPT-4-Turbo, a more efficient and cost-effective version tailored for scale, offering faster responses suitable for commercial applications such as AI chatbots in patient care scenarios, where quick and accurate information retrieval is crucial. These advancements signify continuous progress in the field, promising more sophisticated tools for healthcare professionals [22].

Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT), introduced by Google in 2018, represented a significant advancement in the field of NLP. Unlike traditional LLMs, BERT is designed to understand the context of a word in a sentence by looking at the words that come before and after it [23]. This approach allows BERT to capture a more nuanced understanding of language. This is particularly beneficial in complex and specialized fields such as medical research.

In the medical context, BERT's ability to understand context can be particularly useful in interpreting patient records, medical literature and clinical trial data. For instance, BERT can differentiate between different meanings of the same term in medical contexts, such as 'cold' referring to a common viral infection versus a lower temperature [24]. This level of understanding is important for accurate data interpretation in medical research and patient care. MedBERT and BioBERT are some successful implementations and adaptations of BERT into the life and health sciences space [25,26].

BERT also differs from models like GPT in its training approach. While GPT is trained using a left-to-right approach, BERT's training involves masking some percentage of the input tokens at random and then predicting those masked tokens, which is known as the masked language model. This training method helps BERT learn a robust representation of linguistic context [23].

In terms of applications, BERT has been successfully used to enhance search algorithms in medical databases, improve the accuracy of clinical decision support systems and aid in the extraction of relevant information from

10.2217/fmai-2024-0001

Table 1. Uses of large language models in healthcare.

Category	Task	Usage
Communication and education	Translation	Translate clinical documents into different languages
	Patient education materials	Generate patient education materials such as medication and discharge instructions
Documentation and administrative tasks	Document generation	Generate clinical document drafts such as progress notes, treatment options and discharge summaries
	Medical coding	Assist with the process of assigning standardized codes to medical diagnoses and procedures
Clinical support and decision-making	Clinical decision support	Provide recommendations based on analysis of patient data, best practices and guidelines
	Predictive analytics	Analyze patient data to predict outcomes such as risk of hospitalization or likelihood of developing certain conditions
Research and trials	Clinical trial enrollment	Identify eligible candidates for clinical trials
	Patient stratification	Identify patients most likely to respond to treatments
Patient interaction and assistance	Virtual assistants	Develop virtual assistants such as chatbots to provide information and education to patients and caregivers
	Adverse drug events	Identify potential adverse drug events using data from different medical sources

medical documents [27]. Its ability to understand the context and details of medical language makes it a valuable tool in the healthcare sector.

Like other LLMs, BERT has its limitations, including the requirement of substantial computational resources for training and the need for large-scale, high-quality training datasets. However, its unique bidirectional approach offers significant advantages in understanding and processing complex language structures, making it a potentially valuable asset in medical research and other fields that require a deep understanding of language.

Comparison with GPT models

While both BERT and GPT models are transformer based and excel in language understanding, their core architectures and applications differ. GPT, with its unidirectional approach, excels in generative tasks like language translation and content creation. In contrast, BERT, with its bidirectional understanding, is more adept at tasks that require understanding of language context, such as sentiment analysis and language understanding. Another important difference is the size of the models. GPTs are at least 10- to 100-times bigger than BERT. While this gives them much more contextual reasoning capability, it also makes them much more expensive to employ. For simpler tasks, BERT may be more cost-effective and efficient.

Gemini

Google Gemini, updated in 2023, is a generative pretrained transformer model with a focus on being informative and comprehensive, integrating features like image capabilities, app integration and enhanced coding features. It now operates on PaLM 2, a more capable LLM, which improves its math and reasoning skills [28].

A review of possible uses of LLMs in healthcare

We have outlined many of the components related to LLMs. In this section, we will highlight some possible clinical applications of LLMs in healthcare (Table 1) [29].

Communication & education

LLMs have shown potential in enhancing communication and education within healthcare settings. LLMs can be employed to generate clear and personalized medication instructions [30]. This is particularly beneficial for patients with complex medication regimens, where understanding dosages, timing and potential side effects is critical. By processing patient-specific information, LLMs can tailor these instructions to individual needs, ensuring they are both comprehensive and easy to comprehend. This personalized approach may significantly improve medication adherence and reduce the likelihood of errors.

In addition to medication instructions, LLMs can be used for instructions for patients post hospitalization, which may include complex care. LLMs may assist in creating detailed, customized discharge instructions that cater to the specific medical conditions and care requirements of each patient. These instructions can include information on follow-up care, lifestyle modifications, symptom monitoring and when to seek medical attention. The use of LLMs in generating these instructions will still require a human physician to validate them for accuracy, but the

instructions could be better tailored to the patient's understanding level, thereby enhancing post-discharge care and reducing readmission rates.

In managing chronic diseases, LLMs could be used for creating comprehensive care plans that address long-term treatment strategies, lifestyle changes and ongoing monitoring [31,32]. They can provide tailored information on disease-specific education, helping patients understand their condition and the importance of consistent management. This includes advice on diet, exercise, medication adherence and recognizing warning signs. Such education materials may improve patient understanding and adherence [10].

Additionally, LLMs can help with other aspects of a patient's understanding, including education. LLMs can translate clinical documents, such as patient consent forms and discharge summaries, into different languages [8,9]. This task can improve communication between healthcare providers and patients with different languages.

While LLMs offer considerable benefits in healthcare communication and education, there are limitations and risks. One significant concern is the accuracy and reliability of the information generated. Errors in LLM outputs, due to biases in training data or algorithmic limitations, could lead to misinformation and potentially harm patient care. Another risk involves privacy and security: handling sensitive patient data requires strict adherence to data protection regulations and ethical standards.

Additionally, over-reliance on LLMs could potentially reduce the human element in patient care, which is vital for empathy and understanding the nuances of patient experiences. It is essential for healthcare providers to critically evaluate and verify LLM-generated content and use these tools as supplements rather than replacements for professional medical advice and patient interactions.

Documentation & administrative tasks

LLMs can be particularly useful for aspects of care that are not directly relevant to treatment. The process of documenting patient interactions by clinical practitioners, such as writing progress notes, is time-consuming and labor-intensive. One study found that physicians spend on average over 16 min of a patient visit on interacting with EHRs, with 24% of that time spent on documentation-related tasks [33]. If medical documentation – particularly in the generation of clinical documents such as progress notes, treatment options and discharge summaries – can be facilitated using LLMs, more physician time can be freed to spend interacting with patients.

LLMs have shown promise in synthesizing and contextualizing medical information, and natural language generation can be used to reduce burden [34]. GPT-4 has seen early success in generating ophthalmology operative notes, but requires working in tandem with human clinical expertise and evaluation [35]. Interestingly, one study had clinicians evaluate discharge summaries written by both doctors and ChatGPT. The clinicians had trouble differentiating between the two [36]. Furthermore, AI-written summaries were deemed to be of sufficient quality by general practitioners [36]. Transformer-based models were able to generate effective hospital-course discharge summaries [37].

In addition to documentation-related activities, another important, yet time-consuming part of healthcare operations is the administrative process of assigning billing codes to patient encounters. These types of codes, often in the form of International Classification of Diseases (ICD) codes, need to reflect accurate diagnosis and are important for reporting and insurance reimbursement purposes. However, this process is also time-consuming and not always accurate. There have been mixed successes early on with using LLMs to generate billing codes from patient information. One study even found that GPT3.5 and GPT4 were not perfect in even identifying code identifiers from descriptor text [38].

One paper, however, was able to obtain over 84% accuracy at assigning ICD-10 codes from clinical notes at a health institute in Jordan using a series of BERT models [39]. ChatGPT was used to generate retina-based ICD codes from ophthalmology encounters and achieved a true positive result 70% of the time, with at least one code being correct [39]. These summaries and code automation can save time and reduce the workload for healthcare providers [9,10,23].

Clinical support & decision-making

The integration of LLMs like ChatGPT into clinical care represents a transformative shift in healthcare delivery. Across specialties, LLMs demonstrate a promising capacity to enhance medical decision-making and improve patient outcomes. For instance, knowledge-enhanced auto diagnosis leverages medical domain knowledge to guide the pretraining of a transformer-convolutional neural network LLM foundational multimodal model using chest x-rays and reports. This method, indicating zero-shot prompt performance, a hallmark of LLMs in comparison

10.2217/fmai-2024-0001

with previous methods like convolutional neural networks, rivals the diagnostic accuracy of expert radiologists in certain pathologies [40].

In gastroenterology, a study evaluated ChatGPT's role in assessing acute ulcerative colitis presentations in emergency departments [41]. ChatGPT was tested for its ability to determine disease severity using Truelove and Witts criteria, and the necessity of hospitalization. ChatGPT's assessments showed an 80% consistency with expert gastroenterologists' opinions. This underscores its reliability as a clinical decision support tool in acute ulcerative colitis cases.

The utility of ChatGPT in clinical settings has been exemplified by its ability to provide accurate responses to physician-generated medical queries [42]. In a demonstration of its potential, 33 physicians from 17 specialties reported median accuracy scores that approached complete correctness. This suggests that ChatGPT's advice is largely reliable [42]. This reliability extends into obstetrics, where the obstetric comorbidity index correlates with cesarean delivery rates, revealing significant disparities based on race and ethnicity [43].

In oncology, unmet supportive care needs correlate with increased emergency department visits and hospitalizations. This is particularly true for minority groups [44]. Thus, there is a potential use for LLMs in identifying at-risk populations and guiding resource allocation. LLMs assist in precision oncology by proposing treatment options – some unique and useful – although they have yet to match the credibility of human experts [45]. Similarly, LLMs like ChatGPT have been tested as support tools for tumor board decisions in breast cancer. The results showed alignment with board decisions in a majority of cases [46].

The performance of LLMs in ophthalmology advice suggests they generate responses comparable to those of ophthalmologists, without significant deviation in terms of accuracy or safety [47]. However, the efficacy of these models varies, with some demonstrating higher accuracy in specific medical domains such as myopia [48].

The regulatory oversight of LLMs like GPT-4 becomes crucial given their potential to support medical tasks and the inherent risks of mishandling sensitive data [49]. While foundation models show promise in analyzing EHRs [50], challenges remain, including the risk of generating factually inconsistent summaries [51].

The exploration of LLMs in clinical contexts extends to the development of a German clinical corpus for cardiovascular diseases [52]. Advancements in neurology are evident through studies like the identification of the National Institutes of Health Stroke Scale dimensional structure via machine learning, linking neurological deficits to brain anatomy and function [53].

In neuropsychiatry, NLP technologies help analyze mental health interventions, although challenges in clinical applicability and fairness remain [54].

LLMs can be used for personalized identification of ADEs. These AI systems can process data from multiple medical sources, such as EHRs and knowledge databases. LLMs offer detailed insights into drug interactions by considering individual patient factors, aiding clinicians in informed prescribing and monitoring for ADEs. For patients, they translate complex medical information into understandable language, enhancing awareness about potential ADEs and medication adherence. One study used a BERT-based model to extract and identify key adverse drug reactions from patient discharge summaries and was able to obtain an area under the receiver operating characteristic curve of 0.96 [66]. Separately, Microsoft Bing AI was able to achieve an accuracy of 79% in identifying known drug–drug interactions via querying, which could be used for rapid communication [55].

Overall, in identifying and communicating potential ADEs, LLMs mark a leap forward in personalized medicine. By enhancing patient safety through early ADE detection and tailored communication, these technologies promise to improve drug safety monitoring and patient care. Lastly, NLP approaches have automated the extraction of neurological outcomes from clinical notes, enhancing the scalability of research in neurological outcomes with EHR data [56]. Collectively, these insights affirm the potential of LLMs in augmenting clinical care while emphasizing the need for continuous evaluation and improvement to ensure accuracy, safety and ethical deployment in healthcare settings.

Research & trials

In the realm of academic writing, LLMs like ChatGPT have shown potential in drafting commentaries on specialized topics such as hematological diseases. However, studies reveal significant limitations in these models, particularly in incorporating the latest research findings and providing detailed, specific content.

For instance, a commentary generated on adeno-associated virus missed recent findings and lacked depth in explaining differences between serotypes [57]. This underscores the necessity for human oversight, especially when

aiming for publication in top-tier journals. LLMs can enhance efficiency in academic writing. However, they require careful inspection to ensure current and relevant research.

The ability of LLMs to generate research questions has also been explored. This is particularly evident in fields like gastroenterology. ChatGPT was used to identify research priorities in areas such as inflammatory bowel disease and the microbiome. While the model produced relevant questions, their lack of originality was noted, averaging a modest 3.6 out of 5 in ratings by a panel of gastroenterologists [58]. This finding suggests that while current LLMs can aid in brainstorming, human creativity remains important for generating innovative research question ideas.

The development of specialized LLMs tailored for healthcare contexts offers promising advancements. GatorTronGPT, a generative clinical LLM, was trained using extensive clinical texts and general English text, showing improved performance in biomedical NLP [59]. This specialized model blurred the lines between AI-generated and human-written content in terms of readability and clinical relevance. This suggests that tailor-made LLMs can significantly enhance medical research and healthcare applications.

Furthermore, LLMs have been instrumental in generating datasets for psychological research, as exemplified by the creation of a corpus of 10,000 artificially generated situations corresponding to the Riverside Situational Q tool [60]. This demonstrates their utility in expanding research tools and aiding in the measurement of situational dimensions.

The transformative potential of LLMs extends to patient education and task automation in fields like urology [61] and their integration into digital mental health solutions [62]. They also show promise in improving the accessibility and efficiency of data transformation tasks, as seen in the development of NL2Rigel for intuitive table generation [63].

Additionally, their application in automatic speech recognition technology for clinical transcription demonstrates a narrowing gap between manual and automated services [64]. This progress points toward LLMs becoming increasingly viable in healthcare settings.

However, the effective use of LLMs in academic research depends on the quality of the provided inputs [65]. In scientific writing, the integrity and accuracy of content generated by LLMs are of concern. When tasked with creating research abstracts, most outputs from ChatGPT were detectable by AI output detectors, raising questions about the ethical and acceptable use of LLMs in this domain [66].

LLMs' output is heavily dependent on the specificity and accuracy of the input, emphasizing the need for well-defined prompts and expert oversight. LLMs hold significant promise in healthcare research and education, being capable of enhancing efficiency, expanding research tools and automating tasks. However, their limitations – particularly in originality, specificity and ethical considerations – necessitate a balanced approach. Their integration into medical research must be guided by careful evaluation and human expertise.

Patient interaction & assistance

In healthcare, the integration of LLMs into virtual assistants marks a significant advancement in patient and caregiver engagement. AI systems like chatbots can create interactive and engaging learning experiences, allowing users to ask specific questions and receive immediate responses. Furthermore, they are accessible 24/7, providing reliable information outside regular healthcare provider hours; and support multiple languages, increasing accessibility and reducing health disparities. In one such example, researchers used a sleep bot to help with tracking sleep logs and perception, the results of which were found to be concordant with FitBit measures [67].

Of course, there is risk with using chatbots and virtual assistants, even those trained by state-of-the-art LLMs. They are certainly not ready for off-the-shelf clinical use. They may hallucinate and provide incorrect information, which is certainly possible without the risk of ramifications for them. In fact, researchers have provided examples of chatbots producing erroneous information [68]. As these technologies evolve, they will undoubtedly become integral to certain aspects of patient care with verification and active management.

Limitations & cautions of using LLMs in healthcare

While LLMs show promising applications in enhancing clinical decision support, providing patient education and streamlining operations, there are notable limitations and cautions that must be addressed to ensure their responsible use.

10.2217/fmai-2024-0001

Consistency & reliability

ChatGPT's assessments on ulcerative colitis matched expert opinions 80% of the time [41]. However, its performance in answering patient questions about gastrointestinal health was inconsistent [69]. LLMs should support, not replace, professional judgment.

Cybersecurity risks

Adversarial attacks demonstrate the vulnerability of AI systems, affecting the accuracy of diagnosis. Ensuring strong cybersecurity measures is crucial.

Misinformation & accountability

LLMs can spread misinformation due to their human-like outputs and issues arising from hallucinations. This raises concerns about the accuracy of medical information and the challenge of accountability. It is critical to verify LLM-provided information.

Ethical & legal considerations

The use of LLMs in healthcare, for training and for inference, prompts ethical and legal questions, particularly around patient privacy and consent. The evolving legal framework around AI use in medicine adds complexity to these issues.

LLMs have potential in healthcare but require cautious implementation. Addressing reliability, security, misinformation and regulatory challenges is crucial for their beneficial integration.

Conclusion

In this review, we explored diverse applications of LLMs in healthcare, as well as highlighted several key studies within several utility areas. The versatility of LLMs is evident in their role in enhancing patient communication and education. These models facilitate clearer, more efficient communication among healthcare professionals. They can also translate between languages to alleviate language barriers. In the realm of documentation and administrative tasks, LLMs offer significant time-saving benefits. They streamline paperwork, reduce manual errors and help in organizing and analyzing large volumes of data. Thus these models allow healthcare providers to focus more on patient care. LLMs can also play an important role in clinical support and decision-making. By processing large amounts of medical literature and patient data, they can assist in formulating diagnoses and communicating potential treatment plans to overseeing healthcare professionals. This not only speeds up the decision-making process but also helps in identifying the most effective treatments, potentially improving patient outcomes. In clinical trials, LLMs can expedite the process of patient recruitment by summarizing complex inclusion/exclusion criteria and recommending patients for trials. Finally, in terms of patient interaction and assistance, LLMs in the form of chatbots and virtual assistants can allow for seamless interaction and personalized health information and aid in monitoring patient health, thereby enhancing patient engagement and satisfaction. Put together, LLMs have the potential to transform clinical medicine across a variety of use cases. However, LLM technology raises important ethical and practical considerations [20]. Examples of such considerations include biases, data privacy, the need for human oversight, legal liability and the impact on healthcare professionals. While we still have a way to go before successful implementation, further progress will undoubtedly improve patient outcomes.

Future perspective

Over the next 5–10 years, LLMs will continue to permeate the biomedical field, personalizing medicine, accelerating drug discovery and enhancing healthcare. More data types and modalities will be represented. A large focus of the next decade will be in fostering greater trust of LLMs overall and undertaking more nuanced analysis into potential biases that exist and methods to overcome them. Safety and regulatory frameworks will evolve to provide safer implementation of LLMs into health systems and hospitals, greatly expanding their reach. A growing practice will be to foster implementation sciences, or operationalizing LLMs into health workflows without disrupting them. As models and hardware advance, larger models will be built on bigger and more diverse training datasets, providing more advanced foundational models in various biomedical disciplines. As application programming interface costs continue to go down, it will be easier to fine-tune models for specific healthcare use purposes. Retrieval-augmented generation, and iterations beyond, will also allow for more precise utilizations of LLMs in practice. Additionally, LLMs will help to ease operational burdens to free up the valuable time of healthcare practitioners. While there

Executive summary

Background
- Artificial intelligence (AI), especially large language models (LLMs), shows potential to enhance healthcare through understanding and generating human-like text.
- Ethical and practical issues like data bias and the impact on professionals need careful consideration.

Natural language processing
- Natural language processing helps computers understand human language and is applicable in healthcare for extracting patient information and aiding diagnosis.
- Attention mechanisms and transformers are key LLM technologies, enabling focused processing of language inputs.

Attention mechanism
- An attention mechanism is a type of artificial neural network algorithm which is the basis of most state-of-the-art LLMs today.
- It enables models to focus on relevant parts of text, improving language understanding by applying context to words.

Transformer
- A transformer is a type of neural network model that is composed of repeating attention mechanism layers.
- Transformers excel in natural language processing tasks.
- They require substantial data for training and computational resources for training and inference.
- LLMs, including transformers, are trained using unsupervised learning, affecting model performance based on data quality.

Generative pretraining transformer
- Generative pretraining transformer (GPT) models perform well in language generation tasks.
- Applications in healthcare are varied and include, for instance, generating medical reports and identifying adverse drug events.

Bidirectional Encoder Representations from Transformers
- Bidirectional Encoder Representations from Transformers (BERT) is one of the first developed LLMs. Although composed of hundreds of millions of weights, it is much smaller than GPTs.
- BERT understands language context by analyzing words around the focus word, beneficial in medical applications.
- Med-BERT and BioBERT are adaptations for healthcare, enhancing medical document analysis.

Gemini
- Gemini is an updated generative model produced by Google, focusing on informative and comprehensive responses.

Possible uses of LLMs in healthcare
- LLMs enhance communication, education, documentation and clinical decision-making.
- They generate personalized medical instructions and assist in diagnostics, though accuracy and reliability concerns remain.

Communication & education
- LLMs can create tailored medication instructions and patient education materials, improving understanding and adherence.

Documentation & administrative tasks
- Automating documentation and billing codes with LLMs saves healthcare professionals' time, allowing more patient interaction.

Clinical support & decision-making
- LLMs like ChatGPT can assist in diagnostics and treatment planning, showing potential to match expert radiologists in accuracy.

Research & trials
- LLMs can contribute to academic writing and research question generation but require human oversight for accuracy and relevance.

Patient interaction & assistance
- Virtual assistants and chatbots powered by LLMs enhance patient engagement but need careful management to avoid misinformation.

Conclusion
- LLMs hold transformative potential across various healthcare applications but face challenges like data bias and privacy concerns.
- Ongoing research and ethical considerations are crucial for their successful implementation.

10.2217/fmai-2024-0001

will still be limitations and risks, the increased focus on safe implementation of LLMs in the biomedical field is encouraging and will be an interesting journey to take part in.

Financial disclosure

The authors have no financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, stock ownership or options and expert testimony.

Writing disclosure

No writing assistance was utilized in the production of this manuscript.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25(1), 44–56 (2019).
2. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* 2(10), 719–731 (2018).
3. Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. *arXiv* https://doi.org/10.48550/arXiv.1706.03762 (2017).
4. Budzianowski P, Wen T-H, Tseng B-H *et al.* [1810.00278] MultiWOZ – a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. *arXiv* https://doi.org/10.48550/arXiv.1810.00278 (2018).
5. Briganti G. A clinician's guide to large language models. *Future Med. AI* 1(1), (2023).
6. Kamath U, Liu J, Whitaker J. *Deep Learning for NLP and Speech Recognition. Vol. 84*. Springer International Publishing, Cham, Switzerland (2019).
7. Jurafsky, Dan. *Speech & Language Processing*. Pearson Education India (2000). https://www.amazon.com/Processing-Introduction-Computational-Linguistics-Recognition/dp/9332518416
8. Sorin V, Barash Y, Konen E, Klang E. Deep learning for natural language processing in radiology – fundamentals and a systematic review. *J. Am. Coll. Radiol.* 17(5), 639–648 (2020).
9. Sorin V, Barash Y, Konen E, Klang E. Deep-learning natural language processing for oncological applications. *Lancet Oncol.* 21(12), 1553–1556 (2020).
10. Chan L, Beers K, Yau AA *et al.* Natural language processing of electronic health records is superior to billing codes to identify symptom burden in hemodialysis patients. *Kidney Int.* 97(2), 383–392 (2020).
11. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. *Neurocomputing* 452, 48–62 (2021).
12. Posner MI, Petersen SE. The attention system of the human brain. *Annu. Rev. Neurosci.* 13, 25–42 (1990).
13. DeRose JF, Wang J, Berger M. [2009.07053] Attention flows: analyzing and comparing attention mechanisms in language models. *arXiv* https://doi.org/10.48550/arXiv.2009.07053 (2020).
14. Wolf T, Debut L, Sanh V *et al.* HuggingFace's transformers: state-of-the-art natural language processing. *arXiv* 27(2) 1160–1170 (2019).
15. Gillioz A, Casas J, Mugellini E, Khaled OA. Overview of the transformer-based models for NLP tasks. Presented at: *2020 Federated Conference on Computer Science and Information Systems*, Sofia, Bulgaria, 26 September 2020. doi:10.15439/2020F20
16. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv* https://doi.org/10.48550/arXiv.1810.04805 (2018).
17. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. *arXiv* https://doi.org/10.48550/arXiv.1901.02860(2019).
18. Brown TB, Mann B, Ryder N *et al.* Language models are few-shot learners. *Advances in neural information processing systems* 33, 1877–1901 (2020).
19. Kluge E-HW. Artificial intelligence in healthcare: ethical considerations. *Healthc. Manage. Forum* 33(1), 47–49 (2020).
20. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. (Eds.) Adam Bohr, Kaveh Memarzadeh In: *Artificial intelligence in healthcare*. Elsevier, MA, USA 295–336 (2020).
21. Floridi L. GPT-3: Its Nature, Scope, Limits, and Consequences. Minds & Machines *SSRN J.* 30, 681–694 (2020).

22. Tessler I, Wolfovitz A, Livneh N *et al.* Advancing medical practice with artificial intelligence: ChatGPT in healthcare. *Isr. Med. Assoc. J.* 26(2), 80–85 (2024).
23. Soffer S, Glicksberg BS, Zimlichman E, Klang E. BERT for the processing of radiological reports: an attention-based natural language processing algorithm. *Acad. Radiol.* 29(4), 634–635 (2022).
24. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. *J. Am. Med. Inform. Assoc.* 26(11), 1297–1304 (2019).
25. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* 4(1), 86 (2021).
26. Lee J, Yoon W, Kim S *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4), 1234–1240 (2020).
27. Chen T, Wu M, Li H. A general approach for improving deep learning-based medical relation extraction using a pre-trained model and fine-tuning. *Database (Oxford)* baz116(2019).
28. Singhal K, Azizi S, Tu T *et al.* Large language models encode clinical knowledge. *Nature* 620(7972), 172–180 (2023).
29. Moor M, Banerjee O, Abad ZSH *et al.* Foundation models for generalist medical artificial intelligence. *Nature* 616(7956), 259–265 (2023).
30. Mukherjee S, Durkin C, Pebenito AM, Ferrante ND, Umana IC, Kochman ML. Assessing ChatGPT's ability to reply to queries regarding colon cancer screening based on multisociety guidelines. *Gastro Hep. Adv.* 2(8), 1040–1043 (2023).
31. Mira FA, Favier V, Dos Santos Sobreira Nunes H *et al.* Chat GPT for the management of obstructive sleep apnea: do we have a polar star? *Eur. Arch. Otorhinolaryngol*. 1–7 (2023).
32. Ferro Desideri L, Roth J, Zinkernagel M, Anguita R. Application and accuracy of artificial intelligence-derived large language models in patients with age related macular degeneration. *Int. J. Retina Vitreous* 9(1), 71 (2023).
33. Overhage JM, McCallie D. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. *Ann. Intern. Med.* 172(3), 169–174 (2020).
34. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl. Sci. Proc.* 2020, 191–200 (2020).
35. Waisberg E, Ong J, Masalkhi M *et al.* GPT-4 and ophthalmology operative notes. *Ann. Biomed. Eng.* 51(11), 2353–2355 (2023).
36. Clough RA, Sparkes WA, Clough OT, Sykes JT, Steventon AT, King K. Transforming healthcare documentation: harnessing the potential of AI to generate discharge summaries. *Br. J. Gen. Pract. Open* 10.3399/BJGPO.2023.0116 (2023).
37. Hartman V, Campion TR. A day-to-day approach for automating the hospital course section of the discharge summary. *AMIA Jt Summits Transl. Sci. Proc.* 2022, 216–225 (2022).
38. Soroush A, Glicksberg BS, Zimlichman E *et al.* Assessing GPT-3.5 and GPT-4 in generating international classification of diseases billing codes. *medRxiv* https://doi.org/10.1101/2023.07.07.23292391 (2023).
39. Al-Bashabsheh E, Alaiad A, Al-Ayyoub M, Beni-Yonis O, Zitar RA, Abualigah L. Improving clinical documentation: automatic inference of ICD-10 codes from patient notes using BERT model. *J. Supercomput.* 79, (11) 12766–12790(2023).
40. Zhang X, Wu C, Zhang Y, Xie W, Wang Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat. Commun.* 14(1), 4542 (2023).
41. Levartovsky A, Ben-Horin S, Kopylov U, Klang E, Barash Y. Towards AI-augmented clinical decision-making: an examination of ChatGPT's utility in acute ulcerative colitis presentations. *Am. J. Gastroenterol.* 118(12), 2283–2289 (2023).
42. Goodman RS, Patrinely JR, Stone CA *et al.* Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw. Open* 6(10), e2336483 (2023).
43. Wetcher CS, Kirshenbaum RL, Alvarez A *et al.* Association of maternal comorbidity burden with cesarean birth rate among nulliparous, term, singleton, vertex pregnancies. *JAMA Netw. Open* 6(10), e2338604 (2023).
44. Penedo FJ, Natori A, Fleszar-Pavlovic SE *et al.* Factors associated with unmet supportive care needs and emergency department visits and hospitalizations in ambulatory oncology. *JAMA Netw. Open* 6(6), e2319352 (2023).
45. Benary M, Wang XD, Schmidt M *et al.* Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open* 6(11), e2343689 (2023).
46. Sorin V, Klang E, Sklair-Levy M *et al.* Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer* 9(1), 44 (2023).
47. Bernstein IA, Zhang YV, Govil D *et al.* Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Netw. Open* 6(8), e2330320 (2023).
48. Lim ZW, Pushpanathan K, Yew SME *et al.* Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* 95, 104770 (2023).

49. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit. Med.* 6(1), 120 (2023).
50. Wornow M, Xu Y, Thapa R *et al.* The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit. Med.* 6(1), 135 (2023).
51. Tang L, Sun Z, Idnay B *et al.* Evaluating large language models on medical evidence summarization. *NPJ Digit. Med.* 6(1), 158 (2023).
52. Richter-Pechanski P, Wiesenbach P, Schwab DM *et al.* A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters. *Sci. Data* 10(1), 207 (2023).
53. Cheng B, Chen J, Königsberg A *et al.* Mapping the deficit dimension structure of the National Institutes of Health Stroke Scale. *EBioMedicine* 87, 104425 (2023).
54. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. *Transl. Psychiatry* 13(1), 309 (2023).
55. 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.* 15, 137–147 (2023).
56. Fernandes MB, Valizadeh N, Alabsi HS *et al.* Classification of neurologic outcomes from medical notes using natural language processing. *Expert Syst. Appl.* 214,(15) 119171 (2023).
57. Klang E, Levy-Mendelovich S. Evaluation of OpenAI's large language model as a new tool for writing papers in the field of thrombosis and hemostasis. *J. Thromb. Haemost.* 21(4), 1055–1058 (2023).
58. Lahat A, Shachar E, Avidan B, Shatz Z, Glicksberg BS, Klang E. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Sci. Rep.* 13(1), 4164 (2023).
59. Peng C, Yang X, Chen A *et al.* A study of generative large language model for medical research and healthcare. *NPJ Digit. Med.* 6(1), 210 (2023).
60. Neuman Y, Cohen Y. A dataset of 10,000 situations for research in computational social sciences psychology and the humanities. *Sci. Data* 10(1), 505 (2023).
61. Gupta R, Pedraza AM, Gorin MA, Tewari AK. Defining the role of large language models in urologic care and research. *Eur. Urol. Oncol.* 7(1), 1–13 (2024).
62. Torous J, Benson NM, Myrick K, Eysenbach G. Focusing on digital research priorities for advancing the access and quality of mental health. *JMIR Ment. Health* 10, e47898 (2023).
63. Huang Y, Zhou Y, Chen R *et al.* Interactive table synthesis with natural language. *IEEE Trans. Vis. Comput. Graph.* (Epub ahead of print) 10.1109/TVCG.2023.3329120 (2023).
64. Seyedi S, Griner E, Corbin L *et al.* Using HIPAA (Health Insurance Portability and Accountability Act)-compliant transcription services for virtual psychiatric interviews: pilot comparison study. *JMIR Ment. Health* 10, e48517 (2023).
65. Piazza P, Checcucci E, Puliatti S *et al.* The long but necessary journey towards optimization of the cause–effect relationship between input and output for accountable use of ChatGPT for academic purposes. *Eur. Urol. Focus* 9(6), 1065–1067 (2023).
66. Gao CA, Howard FM, Markov NS *et al.* Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit. Med.* 6(1), 75 (2023).
67. Jang H, Lee S, Son Y *et al.* Exploring variations in sleep perception: comparative study of chatbot sleep logs and Fitbit sleep data. *JMIR Mhealth Uhealth* 11, e49144 (2023).
68. Chakraborty C, Bhattacharya M, Lee S-S. Need an AI-enabled, next-generation, advanced ChatGPT or large language models (LLMs) for error-free and accurate medical information. *Ann. Biomed. Eng.* 52(2), 134–135 (2024).
69. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? *Diagnostics (Basel)* 13(11), 1950 (2023).