Opportunities and challenges for ChatGPT and large language models in biomedicine and health

Shubo Tian [ID], Qiao Jin [ID], Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim,

Donald C. Comeau, Rezarta Islamaj [ID], Aadit Kapoor, Xin Gao and Zhiyong Lu

Corresponding author: Zhiyong Lu, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD 20894, USA. Tel.: +1 301-594-7089; Fax: +1 301-480-2288; E-mail: zhiyong.lu@nih.gov

Abstract

ChatGPT has drawn considerable attention from both the general public and domain experts with its remarkable text generation capabilities. This has subsequently led to the emergence of diverse applications in the field of biomedicine and health. In this work, we examine the diverse applications of large language models (LLMs), such as ChatGPT, in biomedicine and health. Specifically, we explore the areas of biomedical information retrieval, question answering, medical text summarization, information extraction and medical education and investigate whether LLMs possess the transformative power to revolutionize these tasks or whether the distinct complexities of biomedical domain presents unique challenges. Following an extensive literature survey, we find that significant advances have been made in the field of text generation tasks, surpassing the previous state-of-the-art methods. For other applications, the advances have been modest. Overall, LLMs have not yet revolutionized biomedicine, but recent rapid progress indicates that such methods hold great potential to provide valuable means for accelerating discovery and improving health. We also find that the use of LLMs, like ChatGPT, in the fields of biomedicine and health entails various risks and challenges, including fabricated information in its generated responses, as well as legal and privacy concerns associated with sensitive patient data. We believe this survey can provide a comprehensive and timely overview to biomedical researchers and healthcare practitioners on the opportunities and challenges associated with using ChatGPT and other LLMs for transforming biomedicine and health.

Keywords: ChatGPT; large language model; generative AI; biomedicine and health; opportunities and challenges

INTRODUCTION

The recent release of ChatGPT [1] and the subsequent launch of GPT-4 [2] have captured massive attention among both the general public and domain professionals and has triggered a new wave of development of large language models (LLMs). LLMs such as ChatGPT and GPT-4 are language models (LMs) that have billions of parameters in model size and are trained with data sets containing tens or hundreds of billions of tokens. They are considered as foundation models [3] that are pre-trained on a large-scale data and can be adapted to different downstream tasks. LLMs have achieved impressive performance in a wide range of applications in various fields including the biomedical and health domains. A keyword search of 'large language models' OR 'ChatGPT' in PubMed returned 582 articles by the end of May

Shubo Tian is a research scientist in the BioNLP group led by Dr Zhiyong Lu at the National Center for Biotechnology Information, centering his research on information extraction and knowledge discovery.

Qiao Jin, MD, is a postdoctoral fellow at the National Center for Biotechnology Information, working on biomedical text mining and information retrieval.

Dr Lana Yeganova is a scientist at NLM/NIH. Her interests are in mathematics, machine learning, data science and creating a narrative based on data.

Po-Ting Lai, PhD, is a postdoctoral fellow at the National Center for Biotechnology Information, working on biomedical text mining.

Qingqing Zhu received her BS in software engineering from the University of Electronic Science and Technology of China and her PhD from Peking University. She now works at the National Institutes of Health (NIH) as a postdoc. Her research focuses on text generation and image captioning.

Xiuying Chen is now a PhD student at the King Abdullah University of Science and Technology (KAUST), Saudi Arabia. Her research interests include artificial intelligence and text generation.

Yifan Yang, PhD, is a postdoctoral fellow at the National Center for Biotechnology Information, working on research of fairness and bias of artificial intelligence and biomedical text mining.

Qingyu Chen, PhD, is a postdoctoral fellow at the National Center for Biotechnology Information. His research interests include biomedical text mining and information retrieval, medical image analytics, healthcare applications and biocuration.

Dr Won Kim is a staff scientist at the National Library of Medicine (NLM), focuses on enhancing search functionality on PubMed while conducting research in the fields of machine learning and deep learning.

Dr Donald C. Comeau is a staff scientist at the National Library of Medicine (NLM). His research interests include biomedical information retrieval and database resources.

Dr Rezarta Islamaj is a staff scientist at the National Library of Medicine (NLM). His research interests include biomedical information retrieval and database resources.

Aadit Kapoor is a postbaccalaureate fellow at the NIH IPR. His research focuses on applications of artificial intelligence in medicine.

Xin Gao is a professor at the King Abdullah University of Science and Technology (KAUST), Saudi Arabia. His research interests include bioinformatics, computational biology, artificial intelligence and machine learning.

Zhiyong Lu, PhD, FACMI, is a senior investigator and Deputy Director for Literature Search at NCBI/NLM where he leads text mining and machine learning research and directs overall R&D efforts to improve information access in literature databases such as PubMed and LitCovid.

2023. Moreover, the number of publications on the topic has grown exponentially from late 2022 and doubled every month in the last 6 months, covering the technology and its implications for various biomedical and health applications.

Furthermore, several biomedical-specific LLMs have been developed by either training from scratch or fine-tuning existing pre-trained LLMs with biomedical data [4–9]. To provide a comprehensive overview to biomedical researchers and healthcare practitioners on the possible and effective utilization of ChatGPT and other LLMs in our domain, we performed a literature survey, exploring their potentials in a wide variety of different applications such as biomedical information retrieval, question answering, medical text summarization, information extraction and medical education. Additionally, we delve into the limitations and risks associated with these LMs.

Specifically, due to the remarkable language generation capabilities, our focus centers on ChatGPT and other LLMs within the domain of generative artificial intelligence (AI). We searched articles containing keywords related to LLMs, such as 'GPT', 'ChatGPT' or 'large language model', along with keywords of biomedical applications on PubMed (https://pubmed.ncbi.nlm.nih.gov/), medRxiv (https://www.medrxiv.org/), arXiv (https://arxiv.org/) and Google Scholar (https://scholar.google.com/), and included the articles identified as relevant for our review. To the best of our knowledge, this is the first comprehensive survey of opportunities and challenges on ChatGPT and other LLMs for fundamental applications in seeking information and knowledge discovery in biomedicine and health, although there are several previous survey papers on general LLMs [10, 11] and use of ChatGPT for different specific health applications [12–14]. By discussing the capabilities and limitations of ChatGPT and LLMs, we strive to unlock their immense potential in addressing the current challenges within the fields of biomedicine and health. Furthermore, we aim to highlight the role of these models in driving innovation and ultimately improving healthcare outcomes.

OVERVIEW OF CHATGPT AND DOMAIN-SPECIFIC LLMS
Overview of general LLMs

An LM is a statistical model that computes the (joint) probability of a sequence of words (or tokens). Research on LMs has been going on for a long period of time [15]. In 2017, the transformer model introduced by Vaswani *et al.* [16] became the foundational architecture for most modern LMs including ChatGPT. The transformer architecture includes an encoder of bidirectional attention blocks and a decoder of unidirectional attention blocks. Based on the modules used for model development, most recent LMs can be grouped into three categories: encoder-only LMs such as BERT (Bidirectional Encoder Representations from Transformers) [17] and its variants, decoder-only LMs such as the GPT (Generative Pre-trained Transformer) family [18–20] and encoder–decoder LMs such as T5 (Text-to-Text Transfer Transformer) [21] and BART (Bidirectional and AutoRegressive Transformers) [22]. Encoder-only and encoder–decoder LMs are usually trained with an infilling ('masked LM' or 'span corruption') objective along with an optional downstream task, while decoder-only LMs are trained with autoregressive LMs that predict the next token given the previous tokens.

Although the encoder-only and encoder–decoder models have achieved state-of-the-art performance across a variety of natural language processing (NLP) tasks, they have the downside that

requires significant amount of task-specific data for fine-tuning the model to adapt to the specific tasks. This process needs to update the model parameters and adds complexity to model development and deployment.

Unlike those models, when GPT-3 [19] was released, it demonstrated that large decoder-only LMs trained on large text corpus gained significantly increased capability [23] for natural language generation. After training, the models can be directly applied to various unseen downstream tasks through in-context learning such as zero-shot, one-shot or few-shot prompting [19]. This led to a recent trend toward development of decoder-only LLMs in the following years. Following GPT-3, a number of powerful LLMs such as PaLM [24], Galactica [25] and the most recent GPT-4 [2] have been developed. For more information on these general-domain models, readers are invited to consult [10, 11].

While LLMs are powerful, they are still likely to produce content that is toxic, biased or harmful for humans since the large corpus used for model training could contain both high-quality and low-quality data. Thus, it is extremely important to align LLMs to generate outputs that are helpful, honest and harmless for their human users. To achieve this, Ouyang *et al.* [26] designed an effective approach of fine-tuning with human feedback to fine-tune GPT-3 into the InstructGPT model. They first fine-tuned GPT-3 on a dataset of human-written demonstrations of the desired output to prompts using supervised learning and then further fine-tuned the supervised model through reinforcement learning from human feedback (RLHF). This process was referred to as alignment tuning. It was also applied in the development process of ChatGPT and became an effective practice for development of faithful LLMs.

With model size growing bigger, fine-tuning LLMs for downstream tasks becomes inefficient and costly. Alternatively, prompt engineering serves as the key to unlock the power of LLMs given their strong in-context learning ability. As demonstrated by GPT-3, LLMs were able to achieve promising performance on a wide range of natural language tasks through in-context learning by prompting that used a natural language instruction with or without demonstration examples as prompt for the model to generate expected outputs. Wei *et al.* [27] showed that chain-of-thought prompting through a series of intermediate reasoning steps was able to significantly improve LLMs' performance on complex arithmetic, common sense and symbolic reasoning tasks. As a useful approach, designing prompts suitable for specific tasks through prompt engineering became an effective strategy to elicit the in-context learning ability of LLMs. The process of training, fine-tuning with human feedback and unlocking power of LLMs through prompt engineering becomes the paradigm of LLMs as shown in Figure 1.

LLMs for biomedical and health applications

Development of LLMs has been steadily setting new state-of-the-art performance on a variety of tasks in general NLP as well as in biomedical NLP specifically [8, 26, 28–30]. An example is the performance of LLMs on the MedQA dataset, a widely used biomedical question answering dataset that comprises questions in the style of the US Medical Licensing Exam (USMLE) and is used for evaluation of LLMs' reasoning capabilities. In less than half a year, LLM performance has approached a level close to human expert by Med-PaLM 2 [30] from the level of human passing by GPT-3.5 [31], as depicted in Figure 2. These achievements have been accomplished by adapting the LLMs for biomedical QA through different strategies.
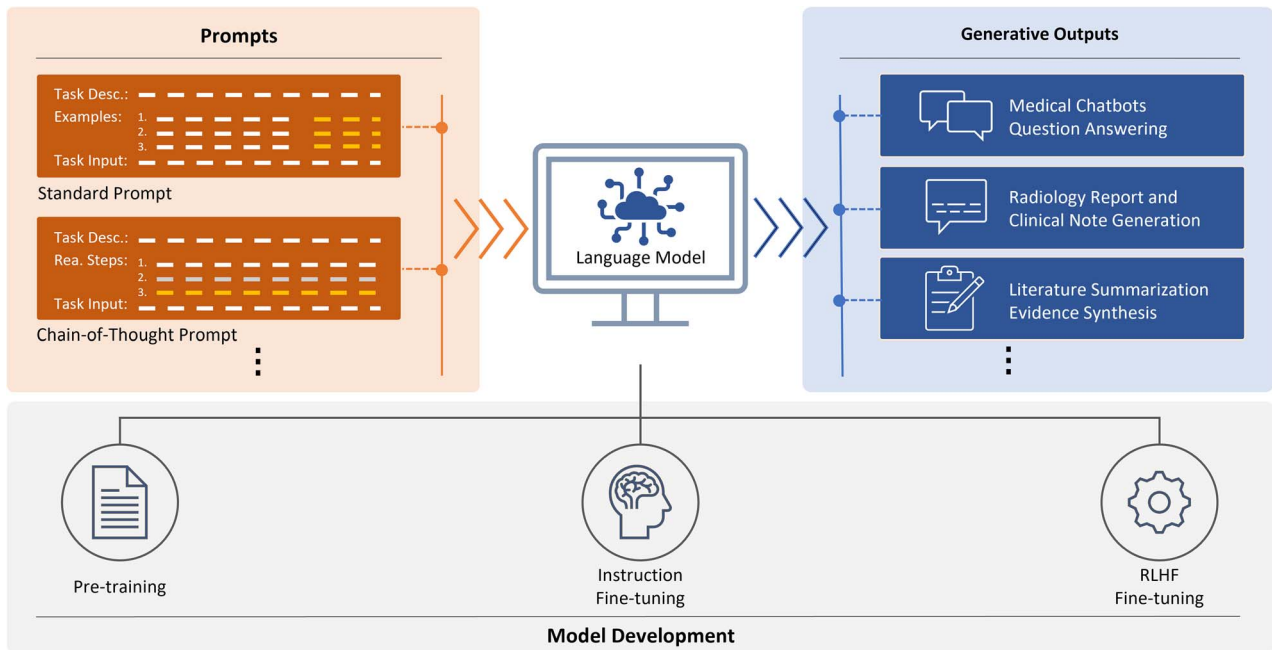
Figure 1. The paradigm of LLMs. Pre-training: LLMs are trained on large scale corpus using an autoregressive LM; Instruction Fine-tuning: pre-trained LLMs are fine-tuned on a dataset of human-written demonstrations of the desired output behavior on prompts using supervised learning; RLHF Fine-tuning: a reward model is trained using collected comparison data, then the supervised model is further fine-tuned against the reward model using reinforcement learning algorithm. Prompts: the instruction and/or example text added to guide LLMs to generate expected outputs. Generative outputs: the outputs produced by the LLMs in response to the users' prompts and inputs.
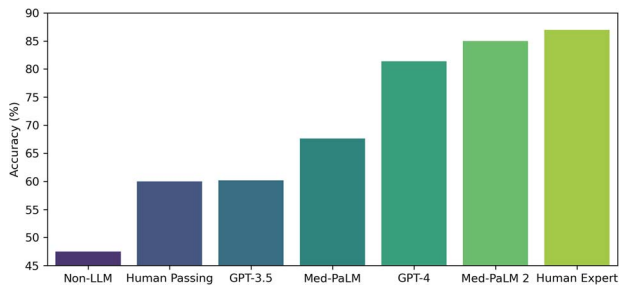

Figure 2. Performance of LLMs versus human on the MedQA (USMLE) dataset in terms of accuracy. Accuracy of LLM performance on the MedQA (USMLE) dataset has increased from the level of human passing by GPT-3.5 to the level close to human expert by Med-PaLM 2 in less than half a year.

There are several strategies that can be applied to adapt Chat-GPT and other LLMs for specific applications in biomedicine and health. When a large amount of data as well as more computing resources and expertise are available, domain-specific LMs can be developed by pre-training from scratch or from checkpoints of existing general LMs. Alternatively, strategies such as fine-tuning with task specific data, instruction fine-tuning and/or RLHF fine-tuning, soft prompt tuning and prompt engineering can be employed to adapt existing pre-trained LMs to specific domain applications. Explanation of these strategies and corresponding examples are listed as follows:

- Pre-training from scratch is to create a specialized LM by pre-training the LM with randomly initialized parameters on a large biomedical corpus using the training objective of either infilling ('masked LM' or 'span corruption') or an autoregressive LM. Both BioMedLM [6] and BioGPT [7] are specialized biomedical LMs developed by pre-training on a corpus of PubMed articles from scratch.

- Pre-training from checkpoints of existing general LMs is to develop a specialized LM by initializing its parameters from the checkpoint of an existing general LM and further pre-training the model on a biomedical corpus with the infilling or autoregressive LM training objectives. The PMC-LLaMA [9] is a model developed by further pre-training the LLaMA-7B [32] model on PubMed Central articles.

- Fine-tuning with task specific data has been frequently used to adapt relatively smaller LMs for specific downstream tasks. This strategy is to fine-tune the existing LMs on the training data of a downstream task with the same training objective of the task. The developers of BioGPT [7] also fine-tuned BioGPT on task-specific data after it was pre-trained from scratch.

- Instruction fine-tuning and/or RLHF fine-tuning is the strategy for aligning LLMs with better instruction responses by fine-tuning the model on data of instruction–response pairs through supervised learning and/or reinforcement learning. Several LLMs including Med-PaLM 2 [30], Clinical Camel [33], ChatDoctor [34] and MedAlpaca [35] have been developed through instruction fine-tuning.

- Soft prompt tuning is the learning of soft prompt vectors that can be used as prompts to LLMs for specific downstream tasks. It is a strategy to take advantage of the benefit from gradient-based learning through a handful of training examples while keeping parameters of the LLMs frozen. The model of Med-PaLM [8] is the result of adapting Flan-PaLM [36] to the biomedical domain through soft prompt tuning.

- Prompt engineering is the process of designing appropriate prompts to adapt LLMs for specific downstream tasks by leveraging the powerful in-context learning capabilities of LLMs without the need of gradient-based learning. Various prompt engineering techniques have been developed and

Table 1: Specialized LLMs in biomedical and health fields

LLM	Size	Description
BioMedLM	2.7B	Developed based on HuggingFace's implementation of the GPT-2 model with small changes through pre-training from scratch on data of 16 million PubMed Abstracts and 5 million PubMed Central full-text articles contained in the Pile dataset [40]
BioGPT	347M and 1.5B	Developed based on the GPT-2 architecture through pre-training from scratch on a corpus of 15 million PubMed articles that have both title and abstract
PMC-LLaMA	7B	Developed by further pre-training from the LLaMA 7B model on 4.9 million PubMed Central articles filtered from the S2ORC Datasets [41] for only five epochs
Med-PaLM 2	N/a	Developed by instruction fine-tuning of PaLM 2 [42] based on a data mixture of medical question answering datasets including MedQA [43], MedMCQA [44], HealthSearchQA [8], LiveQA [45] and MedicationQA [46]
Clinical Camel	13B	Developed by instruction fine-tuning of the LLaMA 13B model on general multi-step conversations in the ShareGPT[a] data and synthetic dialogs transformed from the MedQA data and clinical review articles
ChatDoctor	7B	Developed by instruction fine-tuning of the LLaMA 7B model on more than 100 000 real-world patient-physician conversations collected from two online medical consultation sites
MedAlpaca	7B and 13B	Developed by instruction fine-tuning of the LLaMA 7B and 13B models on the Medical Meadow data, a collection of reformatted instruction–response pairs including datasets for medical NLP tasks and data crawled from various internet resources
Med-PaLM	540B	Developed by adapting Flan-PaLM to medical domain through instruction prompt tuning on 40 examples

[a]https://sharegpt.com/.

applied for adapting LLMs to biomedical and health-related tasks.

Although previous research has shown that LMs pre-trained with biomedical domain-specific data can benefit various in-domain downstream tasks [37–39], pre-training an LM from scratch or existing checkpoint can be very costly, especially when the sizes of LMs are growing larger. Adapting LLMs through instruction fine-tuning, soft prompt tuning and prompt engineering can be more cost-effective and accessible. In addition, while the strategies listed above can be employed independently, they can also be applied in combination when applicable.

The advances of LLMs in recent years have led to development of a number of specialized biomedical LLMs such as BioMedLM [6], BioGPT [7], PMC-LLaMA [9], Med-PaLM [8], Med-PaLM 2 [30], Clinical Camel [33], ChatDoctor [34] and MedAlpaca [35]. Uses of general LLMs including GPT-3 [19], GPT-3.5 [31], ChatGPT [1], GPT-4 [29], Flan-PaLM [8] and Galactica [25] for biomedical applications are being extensively evaluated. Table 1 provides a list of domain-specific LLMs [6–9, 30, 33–35, 40–46]. Performances of various LLMs on different biomedical application tasks are described in section Applications of ChatGPT and LLMs in Biomedicine and Health.

APPLICATIONS OF CHATGPT AND LLMS IN BIOMEDICINE AND HEALTH

ChatGPT and other LLMs can be used in a wide range of biomedical and health applications. In this survey, we cover applications that are fundamental in satisfying information needs of clinical decision-making and knowledge acquisition, including biomedical information retrieval, question answering, medical text summarization, information extraction and medical education.

Information retrieval

Information retrieval (IR) is an integral part in clinical decision-making [47] and biomedical knowledge acquisition [48], as it covers various information-seeking behaviors such as literature search [49], question answering [50] and article recommendation [51]. LLMs like ChatGPT hold significant potentials in changing the way people interact with medical information online [52].

First and foremost, current LLMs may not be directly used as a search engine because their output can contain fabricated information, commonly known as the hallucination issue. For example, when prompted 'Could you tell me what's the relation between p53 and depression? Please also provide the references by PMIDs', ChatGPT makes up the content of PMID 25772646 (perspectives on thyroid hormone action in adult neurogenesis) to support its incorrect answers. This behavior makes retrieving out-of-context knowledge from ChatGPT potentially dangerous by leading the users to draw incorrect conclusions.

However, LLMs might facilitate the interpretation of traditional IR systems by text summarization. It is also been shown by several pilot studies in biomedicine that when LLMs are provided with enough contexts and background information, they can be very effective at reading comprehension [8, 31] and could generate fluent summaries with high fidelity [53]. These results suggest that ChatGPT might be able to summarize the information returned by a traditional IR system and provide a high-level overview or a direct answer to users' queries. Many search engines have integrated LLMs into their result page. For example, 'You.com' and the 'New Bing' provide ChatGPT-like interactive agents that are contextualized on the web search results to help users navigate them; 'scite.ai' presents LLM-generated summaries with references linked to the retrieved articles for literature search results. While aforementioned features are potentially beneficial for all IR systems, researchers have cautioned that the generated outputs must be carefully verified. Although LLMs can summarize in-context information with high fidelity, there is no guarantee that such summaries are error-free [54].

LLMs like ChatGPT can also be used for query enrichment and improving search results with generating more specific queries, expanding a user's search query to include additional relevant terms, concepts or synonyms that may improve the accuracy and relevance of the search results. For instance, Wang *et al.* [55] used ChatGPT to formulate and refine Boolean queries for systematic reviews. They created an extensive set of prompts to investigate tasks on over 100 systematic review topics. Their experiments were conducted on two benchmarking collections: the CLEF technological assisted review (TAR) datasets [56–58] and the Systematic Review Collection with Seed studies [59]. The ChatGPT-generated queries were compared to the original

queries, the Baseline Conceptual and Objective. Evaluation was performed using precision, recall and F1 and F3 score metrics. Their results show that the ChatGPT-generated queries have higher precision but lower recall compared to the queries generated by the current state-of-the-art method [55].

Question answering

Question answering (QA) denotes the task of automatically answering a given question. In biomedicine, QA systems can be used to assist clinical decision support, create medical chatbots and facilitate consumer health education [50]. Based on whether the supporting materials are available, QA tasks can be broadly classified into open(-domain) QA and machine reading comprehension. In open QA, only the question is provided (e.g. a consumer health search query), and a model needs to use external or internal knowledge to answer the question. In machine reading comprehension, both the question and the material for answering the question are available, e.g. in the case where doctors ask questions about specific clinical notes.

A wide variety of biomedical QA datasets have been introduced over the past decade, including BioASQ [60, 61], MedMCQA [44], MedQA (USMLE) [43], PubMedQA [62], GeneTuring [63]. MedMCQA and MedQA are general medical knowledge tests in the US medical licensing exam (USMLE) and Indian medical entrance exams, respectively. Both datasets are open-domain tasks where only the question and four to five answer options are available. In contrast, the questions and answers in GeneTuring are in the genomics domain such as gene name conversion and nucleotide sequence alignment. On the other hand, BioASQ and PubMedQA provide relevant PubMed articles as supporting materials to answer the given question. The biomedical QA tasks are evaluated using the classification accuracy of the possible answers [four to five provided options for MedMCQA and MedQA (USMLE), entities for GeneTuring, yes/no for BioASQ and yes/no/maybe for PubMedQA].

Table 2 shows the performance of LLMs on three commonly used biomedical QA tasks. Overall, the best results are achieved by either Med-PaLM 2 (on MedQA and PubMedQA) or GPT-4 (on MedMCQA), which are currently the largest LLMs containing hundreds of billions of parameters. Notably, they achieve comparable performance on the MedQA dataset and higher performance on the PubMedQA dataset in comparison to the human expert. FLAN-PaLM and GPT-3.5 also achieve high scores on PubMedQA, but are much worse than Med-PaLM 2 and GPT-4 on the MedQA and MedMCQA datasets. This is probably because PubMedQA mainly requires the reading comprehension capability (reasoning), while the other open QA datasets require both reasoning and knowledge. However, smaller LLMs (<10B), such as BioMedLM and PMC-LLaMA, perform similarly to DRAGON [64], a BERT-sized SOTA model enhanced by domain knowledge. This suggests that auto-regressive LLMs could scale to large-enough model sizes to outperform smaller models augmented by structured domain knowledge.

Answering biomedical questions requires up-to-date and accurate knowledge. To address the hallucination issue [65] in medical QA systems, one of the current solutions is retrieval augmentation, which refers to the approach of combining LLMs with a search system, such as the New Bing for the general domain and Almanac [66] in the clinical domain. For a given question, the system will first retrieve relevant documents as supporting materials and then prompt LLMs to answer the question based on the retrieved documents. In this case, LLMs might generate less hallucinations since they are good at summarizing content. However, such systems are still not free from errors [54] and there

is a need for more systematic evaluations [52]. Another promising direction for tackling the hallucination issue is to augment LLMs with additional tools [67–70]. For example, the GeneTuring dataset contains information-seeking questions for specific SNPs such as rs745940901. However, auto-regressive LLMs possess no knowledge about that SNP and most commercial search engines return no results to this query, so retrieval augmentation might not work either. In this case, the information source is only accessible via the NCBI dbSNP database, and augmenting LLMs with NCBI Web database utility APIs can potentially solve the hallucination issue with regard to specific entities in biomedical databases [67].

Consumers have been relying on web search engines like Google for their medical information needs [71]. It is conceivable that they might turn to LLM chatbots because the dialog interface can directly answer their questions and follow-ups. In fact, there have already been several studies, such as Clinical Camel [33], DoctorGLM [72], ChatDoctor [34], HuaTuo [73] and MedAlpaca [35], that attempt to create clinical chatbots by instruction fine-tuning open source LLMs (e.g. LLaMA) on biomedical corpora. However, most of such studies only use small private datasets for evaluation, and the accuracy, generalizability and actual utility of such dialog systems remains unclear.

Biomedical text summarization

Text summarization in the biomedical and health fields is an important application of natural language processing and machine learning. This process involves condensing lengthy medical texts into shorter, easy-to-understand summaries without losing critical information. Summarization in the medical field can be particularly challenging due to the complexity of the language, terminology and concepts. In this section, we will introduce three application scenarios for text summarization in biomedicine: literature summarization, radiology report summarization and clinical note summarization.

The first important application is medical literature summarization [74]. A well-summarized literature review can help in condensing a large volume of information into a concise, readable format, making it easier for readers to grasp the key findings and conclusions. Toward this goal, Cohan et al. [75] introduced a scholar paper summarization task, where they proposed a large-scale dataset of long and structured scientific papers obtained from PubMed, where the abstracts are regarded as the summary of the paper. Pang et al. [76] achieved state-of-the-art performance on this dataset with top-down and bottom-up inference techniques. Taking a step from paper summarization to literature summarization, Chen et al. [77] proposed a related work generation task, where the related work section is considered as the literature review for the specific field. With the development of LLMs, it is expected that more related documents can be considered [78], and better evaluation metrics can be proposed to evaluate the quality of summaries [79].

We next examine how summarization techniques can help medical applications such as radiology report summarization [80]. This is the process of condensing lengthy and detailed radiology reports into concise, informative and easily understandable summaries. Radiology reports contain critical information about a patient's medical imaging results, such as X-rays, CT scans, MRI scans and ultrasound examinations. Representative datasets include MIMIC-CXR [81], which is a large-scale radiography dataset comprising 473 057 chest X-ray images and 206 563 reports. Hu et al. [80] utilized an anatomy-enhanced multimodal model to achieve state-of-the-art results in terms of the ROUGE and CheXbert [82] metrics. In the era of LLMs, Ma et al. [83]

Table 2: Performance of LLMs on biomedical QA tasks

Model	Learning	MedQA (USMLE)	PubMedQA (r.r./r.f.)	MedMCQA (dev/test)
Human expert [31]	–	87.0	78.0/90.4	90.0
Human passing [31]	–	60.0	–	50.0
Med-PaLM 2 [30]	Mixed	86.5[a]	81.8/–	72.3[a]/–
GPT-4 [29]	Few-shot	86.1	80.4/–	73.7/–
FLAN-PaLM [8]	Few-shot	67.6	79.0/–	57.6/–
GPT-3.5 [31]	Few-shot	60.2	78.2/–	59.7/62.7
Galactica [25]	Mixed	44.4	77.6[a]/–	52.9[a]/–
BioMedLM [6]	Fine-tune	50.3	74.4/–	–
BioGPT [7]	Fine-tune	–	–/81.0	–
PMC-LLaMA [9]	Fine-tune	44.7	69.5/–	–/50.5
Non-LLM SOTA [64]	Fine-tune	47.5	73.4/–	–

Note: All numbers are accuracy in percentages. Underline values denote the best performance by language models. r.r.: reasoning-required; r.f.: reasoning-free.
[a]Fine-tuning.

proposed ImpressionGPT, which leverages the in-context learning capability of LLMs for radiology report summarization. Wang and colleagues proposed ChatCAD [84], a framework that summarizes and reorganizes information from a radiology report to support query-aware summarization.

Finally, clinical notes summarization [85] aims to summarize other non-radiology clinical notes, which helps doctors and other healthcare professionals quickly grasp the essential information about a patient's condition, treatments and progress. While radiology report summarizations are more focused, delivering insights based on imaging studies, clinical note summarization involves summarizing the overall status, progress and plan for a patient based on various clinical observations, examinations and patient interactions [86]. McInerney *et al.* [87] proposed and evaluated models that extract relevant text snippets from patient records to provide a rough case summary. Recently, Peng *et al.* [88] demonstrated that while ChatGPT can condense pre-existing systematic reviews, it frequently overlooks crucial elements in the summary, particularly failing to mention short-term or long-term outcomes that are often associated with different levels of risk. Patel and Lam [89] discussed the possibility of using an LLM to generate discharge summaries, and Tang *et al.* [90] tested performance of ChatGPT on their in-house medical evidence dataset. As a concluding work, Ramprasad *et al.* [91] discussed the current challenges in summarizing evidence from clinical notes.

Information extraction

Information extraction involves extracting specific information from unstructured biomedical text data and organizing the extracted information into a structured format. The two most studied IE tasks are (a) named entity recognition (NER): recognizing biological and clinical entities (e.g. diseases) asserted in the free text and (b) relation extraction (RE): extracting relations between entities in the free text.

Pre-trained LMs have been widely used in NER and RE methods. The encoder-only LMs such as BERT are typically fine-tuned with annotated data via supervised learning before being applied for NER and RE tasks. Instead, using decoder-only LMs for NER and RE will usually model them as text generation tasks to directly generate the entities and the relation pairs. Current state-of-the-art (SOTA) NER and RE performance were mostly achieved by models based on encoder-only LMs that were pre-trained on biomedical and clinical text corpus [92, 93] or machine learning method [94].

Recently, several studies have been conducted to explore the use of GPT-3 and ChatGPT for biomedical NER and RE tasks.

For example, Agrawal *et al.* [95] used GPT-3 for NER task on the CASI dataset and showed that GPT-3 was able to outperform the baseline model by observing a single input–output pair. Caufield *et al.* [96] developed SPIRES by recursively querying GPT-3 to obtain responses and achieved an F1-score of 40.65% for RE on the BC5CDR dataset [94] using zero-shot learning without fine-tuning on the training data. Gutiérrez *et al.* [97] used 100 training examples to explore GPT-3's in-context learning for biomedical information extraction and discovered that GPT-3 outperformed PubMedBERT, BioBERT-large and ROBERTa-large in few-shot settings on several biomedical NER and RE datasets. A benchmark study conducted by Chen *et al.* [98] employed prompt engineering method to evaluate ChatGPT's performance on biomedical NER and RE in the BLURB benchmark datasets including BC5CDR-chemical [94], BC5CDR-disease [94], NCBI-disease [99], BC2GM [100], JNLPBA [101], ChemProt [102], DDI [103] and GAD [104] in a zero-shot or few-shot manner. Chen *et al.* [105] performed a pilot study to establish the baselines of using GPT-3.5 and GPT-4 for biomedical NER and RE at zero-shot and one-shot settings. They selected 180 examples with entities or relations and 20 examples without entities or relations from each of the BC5CDR-chemical, NCBI-disease, ChemProt and DDI datasets and designed consistent prompts to evaluate the performance of GPT-3.5 and GPT-4. Tables 3 and 4 summarize performance of different LMs on some commonly used NER and RE benchmark datasets.

The drawback of models achieving SOTA NER and RE performance is their need of labeled data. The remarkable in-context learning abilities of LLMs such as ChatGPT exhibited great potential and provided significant advantages for biomedical NER and RE in circumstances where labeled data are not available. However, they are still not able to surpass performance of LMs that are fine-tuned on task-specific datasets. In addition, several challenges still exist in the use of ChatGPT and other LLMs for information extraction. The generative outputs of ChatGPT and other LLMs sometimes will re-phrase the identified entities or predicted relations that make them difficult to verify. ChatGPT and other LLMs can also produce entities and relations that sound plausible but not factually true. Searching for prompts that are appropriate for NER and RE can be challenging as well. Given all these challenges, extensive research is needed to explore effective approaches to leverage ChatGPT and other LLMs for biomedical information extraction.

Medical education

The use of LLMs in medical education is an exciting and rapidly growing area of research and development. In particular, LLMs

Table 3: Performance of LLMs for NER compared to SOTA on selected datasets (F1-score in %)

LM	Method	BC2GM	BC5CDR-chemical	BC5CDR-disease	JNLPBA	NCBI-disease
SOTA	Task fine-tuning	84.52	93.33	85.62	79.10	87.82
GPT-3	Few-shot	41.40	73.00	43.60		51.40
GPT-3.5	Zero-shot		29.25			24.05
	One-shot		18.03			12.73
ChatGPT	Zero-shot or few-shot	37.54	60.30	51.77	41.25	50.49
GPT-4	Zero-shot		74.43			56.73
	One-shot		82.07			48.37

Table 4: Performance of LLMs for RE compared to SOTA on selected datasets (F1-score in %)

LM	Method	BC5CDR	CHEMPROT	DDI	GAD
SOTA	Task fine-tuning	57.03	77.24	82.36	83.96
BioGPT	Task fine-tuning and few-shot	46.17		40.76	
GPT-3	Few-shot		25.90	16.10	66.00
SPIRES	Zero-shot	40.65			
GPT-3.5	Zero-shot		57.43	33.49	
	One-shot		61.91	34.40	
ChatGPT	Zero-shot or few-shot		34.16	51.62	52.43
GPT-4	Zero-shot		66.18	63.25	
	One-shot		65.43	65.58	

have a potential to mature into education applications and provide alternative learning avenues for students to help them acquire and retain knowledge more efficiently.

One of the attractive features of ChatGPT is its ability to interact in a conversational way [106]. The dialog format makes it possible for ChatGPT to answer follow-up questions and communicate in a conversational format. An early application of ChatGPT in education is a pilot study conducted by Khan Academy [107]. Although the application is not in healthcare education, in general education for students in grades K-12, it is an illustration of the model's integration in an educational environment. Khanmigo, a real-time chat bot, analyzes the answers and guides the student toward the solution by asking questions and providing encouragement.

In addition, ChatGPT is equipped with the ability to provide insights and explanations, suggesting that LLMs may have the potential to become interactive medical education tools to support learning. One of the features making them suitable for education is their ability to answer questions and provide learning experiences for individual students, helping them learn more efficiently and effectively.

ChatGPT can also be used for generating case scenarios [108, 109] or quizzes to help medical students practice and improve their diagnostic and treatment planning abilities [110]. For example, the author in [109] engages in a dialog with a chatbot, asking it to simulate a patient with undiagnosed diabetes and the common labs that may need to be run.

LLMs can also be used to help medical students improve their communication skills. By analyzing natural language inputs and generating human-like responses, LLMs can help students practice their communication skills in a safe and controlled environment. For example, an LLM might be used to simulate patient interactions, allowing students to practice delivering difficult news or explaining complex medical concepts in a clear and concise manner.

Other applications

Besides the previously elaborated fundamental applications, ChatGPT and other LLMs can be used for other applications that are also important in biomedicine and health, such as coreference resolution, text classification and knowledge synthesis. Use of ChatGPT and other LLMs for these applications has been comparably less explored. We briefly summarize these applications as follows.

Coreference resolution is the process of finding all mentions that refer to the same entities in a text. It is an essential task of identifying coreference links to support the discovery of complex information in biomedical texts. Some datasets used for research of coreference resolution include MEDSTRACT [111], FlySlip [112], GENIA-MedCo [113], DrugNerAR [114], BioNLP-ST'11 COREF [115], HANAPIN [116] and CRAFT-CR [117]. While pre-trained LMs such as BioBERT and SpanBERT haven been used in research of coreference reference, LLMs including ChatGPT have not been explored [118].

Text classification aims to assign one or more pre-defined labels to a given text such as a sentence, a paragraph or a document. It plays an important role in biomedical sentiment analysis and document classification. HoC [119] has been the dataset used widely for research of biomedical text classification. It contains 1580 PubMed abstracts that were manually annotated for multi-label document classification of hallmarks of cancer. State-of-the-art performances on HoC were achieved by fine-tuned pre-trained LMs [7]. Experiments had shown that performances of LLMs such as ChatGPT and GPT-4 were suboptimal compared to fine-tuned pre-trained LMs [98, 105].

Knowledge synthesis attempts to extract and summarize useful information from large amount of data to generate comprehensive knowledge and new insights. It is a critical step in biomedical knowledge discovery and translation. Pre-trained with large number of diverse data, LLMs like ChatGPT are believed to have encoded rich biomedical and clinical knowledge [8]. Some

experiments had been conducted to evaluate use of LLMs for summarizing, simplifying and synthesizing medical evidence [88, 90, 120]. LLMs will make it possible to automate knowledge synthesis on a large scale to accelerate biomedical discoveries and improve medical education and health practices.

Given the powerful capabilities of LLMs, we can expect that more creative LLMs-powered applications will be developed in the domains of biomedicine and health.

LIMITATIONS AND RISKS OF LLMS

While LLMs like ChatGPT demonstrate powerful capabilities, these models are not without limitations. In fact, the deployment of LLMs in high-stakes applications, particularly within the biomedical and health domain, presents challenges and potential risks. The limitations, challenges and risks associated with LLMs have been extensively discussed in previous research [19, 121, 122], and in this survey, we will specifically focus on those relevant to the context of biomedical and health domains.

Hallucination

All LMs have the tendency to hallucinate—producing content that may seem plausible but is not correct. When such content is used for providing medical advice or in clinical decision-making, the consequence can be particularly harmful and even disastrous. The potential danger associated with hallucinations can become more serious as the capabilities of LLMs continue to advance, resulting in more convincing, persuasive and believable hallucinations. These systems are known to lack transparency—inability to relate to the source, which creates a barrier for using the provided information. For healthcare professionals to use LLMs in support of their decision-making, great caution should be exercised to verify the generated information.

Another concern is that LLMs may not be able to capture the full complexity of medical knowledge and clinical decision-making or produce erroneous results. While LLMs can analyze vast amounts of data and identify patterns, they may not be able to replicate the nuanced judgment and experience of a human clinician. The usage of non-standard terminologies presents an additional complication.

Fairness and bias

In recent years, fairness has gained the attention of ML research communities as a crucial consideration for both stable performance and unbiased downstream prediction. Many studies have shown that LMs can amplify and perpetuate biases [2, 123] because they learned from historical data. This may inadvertently perpetuate biases and inequalities in healthcare. In a recent study, researchers show that text generated with GPT-3 can capture social bias [124]. Although there exists a lot of research in the general domain regarding fairness in ML and NLP including gender and racial bias, little work has been done in biomedical domain. Many current datasets do not have demographic information, as this relates to privacy concerns in medical practices. An unfair and biased model in the biomedical and health domain can lead to detrimental outcomes and affect the quality of treatment a patient receives [125–127].

Privacy

The corpora used for LLMs training usually contain a variety of data from various sources, which may include private personal information. Huang *et al.* [128] found that LMs can leak personal information. It was also reported that GPT-4 has the potential to be used for attempt to identify private individuals and associate personal information such as geographic location and phone number [122]. Biomedical and clinical text data used for training LLMs may contain patient information and pose serious risks to patient privacy. LLMs deployed for biomedical and health applications can also present risks to patient privacy, as they may have access to patient characteristics, such as clinical measurements, molecular signatures and sensory tracking data.

Legal and ethical concerns

Debates on legal and ethical concerns of using AI for medicine and healthcare have been carried out continuously in recent years [129]. The widespread interests in ChatGPT recently also raised many concerns on legal and ethical issues regarding the use of LLMs like ChatGPT in medical research and practices [130, 131]. It was advocated to establish a robust legal framework encompassing transparency, equity, privacy and accountability. Such framework can ensure safe development, validation, deployment and continuous monitoring of LLMs, while taking into account limitations and risks [132].

The acknowledgement of ChatGPT as an author in biomedical research has been particularly identified as an ethical concern. Biomedical researchers may have their opinions on whether ChatGPT or other LLMs should be welcomed to their ranks. In fact, several papers already list ChatGPT as an author [133–136]. However, after several ethical concerns were raised [137], several of these papers ended up removing ChatGPT from the author list [133, 138]. One of the problems of allowing machine written articles is whether they could be reliably recognized. In one report, humans not only detected 68% of generated abstracts but also flagged 14% of human abstracts as machine-generated [139].

The most valid criticism of LLM-assisted generation of scientific papers is accountability. There are no consequences to the LLM if the output is wrong, misleading or otherwise harmful. Thus, they cannot take responsibility for writing the article [140–144]. Another issue is copyright—in many jurisdictions, machine-generated material may not receive a copyright [130, 140], which poses an obvious problem for journals.

Questions also arise regarding the disclosure of LLM usage during a project or in preparing a paper [142, 144]. There is a long-standing tradition of reporting tools that were used for a project. On one side, if the LLM had a material impact on the study, it should be reported. On the other side, we do not report the spell checker that was used to prepare a paper. Should we report the LLM? Distinguishing these extremes in the context of an LLM will require time and experience.

Lack of comprehensive evaluations

LLMs must be comprehensively evaluated with regard to their performance, safety and potential bias before any implementations in biomedicine. However, evaluating these biomedical LLMs is not trivial. While some traditional NLP tasks such as NER and RE have reliable automatic evaluation metrics like F1 scores, users mostly use LLMs to get free-text response for their biomedical information needs, such as question answering and text summarization. Generally, expert evaluations of such free-text LLM outputs are considered as the gold standard, but getting such evaluations is labor-intensive and not scalable. For example, a panel of clinicians were employed to evaluate the Med-PaLM answers to medical questions among several axes such as scientific consensus, content appropriateness and extent of possible harm [8]. However, only 140 questions have been evaluated in the study, probably due to the high cost of expert annotations.

Another issue for manual evaluation is that there is no consensus on what axes should be evaluated or the scoring guidelines, so the manual evaluation results from different studies are not directly comparable. Therefore, it is imperative to arrive at a reporting consensus, such as the PRISMA statement [145] for systematic reviews, for evaluating biomedical LLMs.

Alternatively, there are two main approaches to evaluate LLM answers without involving expert evaluators. The most common practice is to convert the task into USMLE-style multi-choice questions (such as MedQA, PubMedQA and MedMCQA) and evaluate the accuracy of LLM-generated answer choices. These tasks serve as a good proxy for evaluating the knowledge reasoning capabilities of LLMs. However, they are not realistic since the answer choices will not be provided in real-life user questions in biomedicine. The other solution is to evaluate the LLM-generated response against a reference answer or summary with automatic metrics. These automatic scoring can be based on lexical overlap such as BLEU [146], ROUGE [147] and METEOR [148], as well as semantic similarity like BERTScore [149], BARTScore [150] and GPTScore [151]. Although these automatic metrics can evaluate free-text LLM outputs in a large scale, they often do not strongly correlate with human judgments [79, 90]. As such, it is also vital to design new evaluation metrics, potentially with LLMs, that can be both scalable and accurate.

Open-source versus closed-source LLMs

When it comes to implementing LLMs for applications, one important decision the users need to make is whether to choose open-source or closed-source LLMs. Both categories have their pros and cons and pose distinct challenges and risks to users. Open-source LLMs play an important role in facilitating innovation and adaptation of new technologies in the community. They offer users with more transparency and more control over the models, but less support. Using open-source LLMs, users could adapt and customize the LLMs to their specific needs with higher flexibility. However, open-source LLMs users usually depend on the community for support. When they don't have strong internal technical resources, they might face challenges in adaptation, customization and implementation and run into the risk of unsuccessful deployment of LLMs. On the other hand, closed-source LLMs can provide users with more dedicated support, but less transparency and less control over the models. Closed-source LLMs are often backed by large corporations with substantial resources for support in development, improvement and deployment. However, closed-source LLMs users could experience difficulties in customizing the LLMs to their needs due to the lack of transparency and control of the LLMs and may encounter the risks of vendor dependency and leaking sensitive information when they are submitted to the LLMs. Furthermore, costs of using open-source versus closed-source LLMs can also be very different. Implementing open-source LLMs might require more initial investment but fewer long-term costs that are predictable, while using closed-source LLMs might need less investment at the beginning but higher subscription costs that may increase because of vendor pricing changes. The decision to use open-source or closed-source LLMs depends on the specific needs, priorities and resources of the users. It is important for the users to carefully evaluate the benefits and pitfalls of both options before making the decision.

DISCUSSION AND CONCLUSION

In this survey, we reviewed the recent progress of LLMs with a focus on generative models like ChatGPT and their applications in the biomedical and health domains. We find that biomedical and health applications of ChatGPT and other LLMs are being extensively explored in the literature and that some domain-specialized LLMs have been developed. Performance of specialized and non-specialized LLMs for biomedical applications have been evaluated on a variety of tasks. Our findings also revealed that performance of LLMs varies on different biomedical downstream tasks. LLMs were able to achieve new state-of-the-art performance on text generation tasks such as medical QA. However, they still underperformed the existing fine-tuning approach of smaller LMs for information extraction.

Looking forward, the opportunities for LLMs present promising prospects for deployment of LLM-powered systems for biomedical and health applications in real-life scenarios. In the era of LLMs, the future direction of medical summarization holds significant promise. We can anticipate that LLMs will be increasingly utilized to automatically summarize extensive medical literature, radiology reports and clinical notes. This would facilitate quicker access to vital information and support decision-making processes for healthcare professionals. Additionally, they are expected to better handle complex medical terminology and context, thus improving the quality of summaries. Another potential area of growth is in patient communication. LLMs could be used to transform complex medical jargon into layman's terms, aiding patients in understanding their health conditions and treatment options. Furthermore, a medical classroom furnished with LLMs can bring students more personalized learning experiences and more focus on study of critical thinking and problem-solving skills. A clinical system integrated with LLMs can beneficially provide patients and physicians with efficient and quality healthcare services through accurate diagnosis, precision medicine, appropriate decision-making and proper clinical documentation in preparing succinct clinical reports, concise clinical notes and warm-hearted patient letters.

As a matter of fact, a number of articles have been published on perspectives of using ChatGPT for biomedical and clinical applications in practice [130, 131]. Many experiments have been conducted to evaluate use of ChatGPT in various scenarios in biomedical and clinical workflows. However, until now, no actual deployment of ChatGPT or any other LLMs has been reported. Because of the high-stakes nature of biomedical and health settings, deployment of LLMs like ChatGPT into practice requires more prudence given their limitations and risks. In particular, the transparency challenge that the training data of ChatGPT and other LLMs remain closed source increases difficulties in the evaluation of LLMs.

While the potential benefits are immense, we must also be mindful of the risks and challenges as discussed previously. Strategies and techniques need to be developed and deployed for overcoming the limitations of LLMs. To alleviate generation of nonsensical or harmful content, retrieval augmentation techniques can be used, effective prompts need to be crafted and rigorous evaluation methods shall be applied [67, 152]. To mitigate bias and improve fairness, training data need to be diversified, bias and fairness of LLMs shall be analyzed and bias detection shall be implemented. To protect privacy of individuals, sensitive personal information shall be limited and deidentified when they are used in LLMs. Regulations shall be created and issued to secure legal and proper use of LLMs. The research community is working hard on development of such strategies and techniques. Ensuring the ethical use of AI in healthcare, maintaining patient privacy, mitigating biases in AI models and increasing transparency of AI models are some of the significant considerations for future

development in this area. Therefore, a multidisciplinary approach, including input from healthcare professionals, data scientists, ethicists and policymakers, will be crucial to guide the future direction of future research and development in the era of LLMs.

Key Points

- We examined diverse applications of LLMs including ChatGPT in biomedicine and health.
- LLMs have achieved significant advances in the field of text generation tasks, but modest advances in other biomedical and health applications.
- The recent rapid progress of LLMs indicates their great potential of providing valuable means for accelerating discovery and improving health.
- The use of LLMs like ChatGPT in biomedicine and health entails various risks and challenges including fabricated information as well as legal and privacy concerns.
- A multidisciplinary approach is crucial to guide the future direction of research and development of LLMs for biomedical and health applications.

FUNDING

REFERENCES

1. OpenAI. *Introducing ChatGPT*. OpenAI Blog Post 2022. https://openai.com/blog/chatgpt (4 May 2023, date last accessed).
2. OpenAI. *GPT-4 Technical Report*. arXiv Preprint 2023; arXiv:2303.08774.
3. Bommasani R, Hudson DA, Adeli E, *et al*. *On the Opportunities and Risks of Foundation Models*. arXiv Preprint 2022; arXiv:2108.07258.
4. Shin H-C, Zhang Y, Bakhturina E, *et al*, BioMegatron: Larger Biomedical Domain Language Model. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2020;4700–6.
5. Yang X, Chen A, PourNejatian N, *et al*. *GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records*. arXiv Preprint 2022; arXiv:2203.03540.
6. Bolton E, Hall D, Yasunaga M, *et al*. BioMedLM: a Domain-Specific Large Language Model for Biomedical Text. *Stanford CRFM Blog* 2022. https://crfm.stanford.edu/2022/12/15/biomedlm.html (22 February 2023, date last accessed).
7. Luo R, Sun L, Xia Y, *et al*. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 2022;**23**(6):bbac409.
8. Singhal K, Azizi S, Tu T, *et al*. *Large Language Models Encode Clinical Knowledge*. arXiv Preprint 2022; arXiv:2212.13138.
9. Wu C, Zhang X, Zhang Y, *et al*. *PMC-LLaMA: Further Finetuning LLaMA on Medical Papers*. arXiv Preprint 2023; arXiv:2304.14454.
10. Zhao WX, Zhou K, Li J, *et al*. *A Survey of Large Language Models*. arXiv Preprint 2023; arXiv:2303.18223.
11. Yang J, Jin H, Tang R, *et al*. *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*. arXiv Preprint 2023; arXiv:2304.13712.
12. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023;**6**(online). https://www.frontiersin.org/articles/10.3389/frai.2023.1169595.
13. Thapa S, Adhikari S. ChatGPT, bard, and large language models for biomedical research: opportunities and pitfalls. *Ann Biomed Eng* 2023;**51**:2647–51.
14. Sharma S, Pajai S, Prasad R, *et al*. A critical review of Chat-GPT as a potential substitute for diabetes educators. *Cureus* 2023;**15**(5):e38380.
15. Bengio Y, Ducharme R, Vincent P, *et al*. A Neural Probabilistic Language Model. *J Mach Learn Res* 2003;**3**:1137–55.
16. Vaswani A, Shazeer N, Parmar N, *et al*. *Attention Is All You Need*. arXiv Preprint 2017; arXiv:1706.03762.
17. Devlin J, Chang M-W, Lee K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume **1** (Long and Short Papers) 2019; 4171–86.
18. Radford A, Narasimhan K, Salimans T, *et al*. Improving Language Understanding by Generative Pre-Training. *OpenAI Research Papers* 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (3 May 2023, date last accessed).
19. Brown TB, Mann B, Ryder N, *et al*. Language Models are Few-Shot Learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems* 2020; 1877–1901.
20. Radford A, Wu J, Child R, *et al*. Language Models are Unsupervised Multitask Learners. *OpenAI Research Papers* 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (28 March 2023, date last accessed).
21. Raffel C, Shazeer N, Roberts A, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2020;**21**(140):1–67.
22. Lewis M, Liu Y, Goyal N, *et al*. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 2020; 7871–80.
23. Wei J, Tay Y, Bommasani R, *et al*. *Emergent Abilities of Large Language Models*. arXiv Preprint 2022; arXiv:2206.07682.
24. Chowdhery A, Narang S, Devlin J, *et al*. *PaLM: Scaling Language Modeling with Pathways*. arXiv Preprint 2022; arXiv:2204.02311.
25. Taylor R, Kardas M, Cucurull G, *et al*. *Galactica: A Large Language Model for Science*. arXiv Preprint 2022; arXiv:2211.09085.
26. Ouyang L, Wu J, Jiang X, *et al*. *Training language models to follow instructions with human feedback*. arXiv Preprint 2022; arXiv:2203.02155.
27. Wei J, Wang X, Schuurmans D, *et al*. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv Preprint 2023; arXiv:2201.11903.
28. OpenAI. OpenAI codex. OpenAI Blog Post 2021. https://openai.com/blog/openai-codex (3 May 2023, date last accessed).
29. Nori H, King N, McKinney SM, *et al*. *Capabilities of GPT-4 on Medical Challenge Problems*. arXiv Preprint 2023; arXiv:2303.13375.
30. Singhal K, Tu T, Gottweis J, *et al*. *Towards Expert-Level Medical Question Answering with Large Language Models*. arXiv Preprint 2023; arXiv:2305.09617.
31. Liévin V, Hother CE, Winther O. *Can Large Language Models Reason About Medical Questions?* arXiv Preprint 2023; arXiv:2207.08143.

32. Touvron H, Lavril T, Izacard G, *et al. LLaMA: Open and Efficient Foundation Language Models*. arXiv Preprint 2023; arXiv:2302.13971.

33. Toma A, Lawler PR, Ba J, *et al. Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding*. arXiv Preprint 2023; arXiv:2305.12031.

34. Li Y, Li Z, Zhang K, *et al. ChatDoctor: A Medical Chat Model Fine-tuned on LLaMA Model using Medical Domain Knowledge*. arXiv Preprint 2023; arXiv:2303.14070.

35. Han T, Adams LC, Papaioannou J-M, *et al. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data*. arXiv Preprint 2023; arXiv:2304.08247.

36. Chung HW, Hou L, Longpre S, *et al. Scaling Instruction-Finetuned Language Models*. arXiv Preprint 2022; arXiv:2210.11416.

37. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**(4):1234–40.

38. Gu Y, Tinn R, Cheng H, *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans Comput Healthcare* 2021;**3**:2:1–2:23.

39. Miolo G, Mantoan G, Orsenigo C. *ELECTRAMed: A New Pre-Trained Language Representation Model for Biomedical NLP*. arXiv Preprint 2021; arXiv:2104.09585.

40. Gao L, Biderman S, Black S, *et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. arXiv Preprint 2020; arXiv:2101.00027.

41. Lo K, Wang LL, Neumann M, *et al. S2ORC: The Semantic Scholar Open Research Corpus*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020; 4969–83.

42. Anil R, Dai AM, Firat O, *et al. PaLM 2 Technical Report*. arXiv Preprint 2023; arXiv:2305.10403.

43. Jin D, Pan E, Oufattole N, *et al.* What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl Sci* 2021;**11**(14):6421.

44. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. *Proceedings of the Conference on Health, Inference, and Learning* 2022; 248–60.

45. Abacha AB, Agichtein E, Pinter Y. *Overview of the Medical Question Answering Task at TREC 2017 LiveQA*. The Twenty-Sixth Text REtrieval Conference (TREC 2017) Proceedings 2017. https://trec.nist.gov/pubs/trec26/papers/Overview-QA.pdf (30 May 2023, date last accessed).

46. Abacha AB, Mrabet Y, Sharp M, *et al. Bridging the Gap Between Consumers' Medication Questions and Trusted Answers. MEDINFO 2019: Health and Wellbeing e-Networks for All* 2019; 25–29.

47. Ely JW, Osheroff JA, Chambliss ML, *et al.* Answering Physicians' clinical questions: obstacles and potential solutions. *J Am Med Inform Assoc* 2005;**12**(2):217–24.

48. Gopalakrishnan V, Jha K, Jin W, *et al.* A survey on literature based discovery approaches in biomedical domain. *J Biomed Inform* 2019;**93**:103141.

49. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* 2011;**2011**:baq036.

50. Jin Q, Yuan Z, Xiong G, *et al.* Biomedical Question Answering: A Survey of Approaches and Challenges. *ACM Comput Surv* 2022;**55**:35:1–35:36.

51. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics* 2007;**8**(1):423.

52. Jin Q, Leaman R, Lu Z. Retrieve, summarize, and Verify: how will ChatGPT impact information seeking from the medical literature? *J Am Soc Nephrol* 2023;**34**(8):1302–4.

53. Jeblick K, Schachtner B, Dexl J, *et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports*. arXiv Preprint 2022; arXiv:2212.14882.

54. Liu NF, Zhang T, Liang P. *Evaluating Verifiability in Generative Search Engines*. arXiv Preprint 2023; arXiv:2304.09848.

55. Wang S, Scells H, Koopman B, *et al. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?* arXiv Preprint 2023; arXiv:2302.03495.

56. Kanoulas E, Li D, Azzopardi L, *et al.* CLEF 2017 technologically assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings* 2017;**1866**:1–29.

57. Kanoulas E, Li D, Azzopardi L, *et al.* CLEF 2018 Technologically Assisted Reviews in Empirical Medicine Overview. *CEUR Workshop Proceedings* 2018;**2125**:1–34.

58. Kanoulas E, Li D, Azzopardi L, *et al.* CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. *CEUR Workshop Proceedings* 2019;**2380**:1–31.

59. Wang S, Scells H, Clark J, *et al.* From Little Things Big Things Grow: A Collection with Seed Studies for Medical Systematic Review Literature Search. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2022; 3176–86.

60. Krithara A, Nentidis A, Bougiatiotis K, *et al.* BioASQ-QA: a manually curated corpus for biomedical question answering. *Sci Data* 2023;**10**(1):170.

61. Tsatsaronis G, Balikas G, Malakasiotis P, *et al.* An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* 2015;**16**(1):138.

62. Jin Q, Dhingra B, Liu Z, *et al.* PubMedQA: A Dataset for Biomedical Research Question Answering. *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; 2567–77.

63. Hou W, Ji Z. *GeneTuring tests GPT models in genomics*. bioRxiv Preprint 2023; 2023.03.11.532238.

64. Yasunaga M, Bosselut A, Ren H, *et al. Deep Bidirectional Language-Knowledge Graph Pretraining*. arXiv Preprint 2022; arXiv:2210.09338.

65. Ji Z, Lee N, Frieske R, *et al.* Survey of hallucination in natural language generation. *ACM Comput Surv* 2023;**55**(12):1–38.

66. Zakka C, Chaurasia A, Shad R, *et al. Almanac: Retrieval-Augmented Language Models for Clinical Medicine*. arXiv Preprint 2023; arXiv:2303.01229.

67. Jin Q, Yang Y, Chen Q, *et al. GeneGPT: Augmenting Large Language Models with Domain Tools for Improved Access to Biomedical Information*. arXiv Preprint 2023; arXiv:2304.09667.

68. Parisi A, Zhao Y, Fiedel N. *TALM: Tool Augmented Language Models*. arXiv Preprint 2022; arXiv:2205.12255.

69. Qin Y, Hu S, Lin Y, *et al. Tool Learning with Foundation Models*. arXiv Preprint 2023; arXiv:2304.08354.

70. Gao L, Madaan A, Zhou S, *et al. PAL: Program-aided Language Models*. arXiv Preprint 2023; arXiv:2211.10435.

71. Fox S, Duggan M. *Health Online 2013*. Pew Research Center Blog Post 2013. https://www.pewresearch.org/internet/2013/01/15/health-online-2013/ (7 May 2023, date last accessed).

72. Xiong H, Wang S, Zhu Y, *et al. DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task*. arXiv Preprint 2023; arXiv:2304.01097.

73. Wang H, Liu C, Xi N, *et al. HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge*. arXiv Preprint 2023; arXiv:2304.06975.

74. Qazvinian V, Radev DR, Mohammad SM, *et al.* Generating extractive summaries of scientific paradigms. *J Artif Intell Res* 2013;**46**(1):165–201.

75. Cohan A, Dernoncourt F, Kim DS, *et al.* A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume **2** (Short Papers) 2018; 615–21.

76. Pang B, Nijkamp E, Kryściński W, *et al.* *Long Document Summarization with Top-down and Bottom-up Inference.* arXiv Preprint 2022; arXiv:2203.07586.

77. Chen X, Alamro H, Li M, *et al.* Capturing Relations between Scientific Papers: An Abstractive Model for Related Work Section Generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 2021; 6068–77.

78. Laundry N. *Turn GPT-4 Into your Personal Literature Review Bot. The Academic's Field Guide to Writing Code Blog Post on Medium* 2023. https://medium.com/the-academics-field-guide-to-coding/turn-gpt-4-into-your-personal-literature-review-bot-ff955ba6fbb0 (5 June 2023, date last accessed).

79. Wang LL, Otmakhova Y, DeYoung J, *et al.* *Automated Metrics for Medical Multi-Document Summarization Disagree with Human Evaluations.* arXiv Preprint 2023; arXiv:2305.13693.

80. Hu J, Chen Z, Liu Y, *et al.* *Improving Radiology Summarization with Radiograph and Anatomy Prompts.* arXiv Preprint 2022; arXiv:2210.08303.

81. Johnson AEW, Pollard TJ, Greenbaum NR, *et al.* *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs.* arXiv Preprint 2019; arXiv:1901.07042.

82. Smit A, Jain S, Rajpurkar P, *et al.* *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT.* arXiv Preprint 2020; arXiv:2004.09167.

83. Ma C, Wu Z, Wang J, *et al.* *ImpressionGPT: An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT.* arXiv Preprint 2023; arXiv:2304.08448.

84. Wang S, Zhao Z, Ouyang X, *et al.* *ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models.* arXiv Preprint 2023; arXiv:2302.07257.

85. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. *J Am Med Inform Assoc* 2015;**22**(5):938–47.

86. Feblowitz JC, Wright A, Singh H, *et al.* Summarization of clinical information: a conceptual model. *J Biomed Inform* 2011;**44**(4): 688–99.

87. McInerney DJ, Dabiri B, Touret A-S, *et al.* *Query-Focused EHR Summarization to Aid Imaging Diagnosis.* Proceedings of the 5th Machine Learning for Healthcare Conference 2020; 632–59.

88. Peng Y, Rousseau JF, Shortliffe EH, *et al.* AI-generated text may have a role in evidence-based medicine. *Nat Med* 2023;**29**:1593–94.

89. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health* 2023;**5**(3):e107–8.

90. Tang L, Sun Z, Idnay B, *et al.* Evaluating Large Language Models on Medical Evidence Summarization. *npj Digit Med* 2023; 6:1–8.

91. Ramprasad S, Mcinerney J, Marshall I, *et al.* Automatically Summarizing Evidence from Clinical Trials: A Prototype Highlighting Current Challenges. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* 2023; 236–47.

92. Luo L, Wei C-H, Lai P-T, *et al.* AIONER: All-in-One Scheme-Based Biomedical Named Entity Recognition Using Deep Learning. *Bioinformatics* 2023;**39**:btad310.

93. Alrowili S, Shanker V. BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. *Proceedings of the 20th Workshop on Biomedical Language Processing* 2021; 221–27.

94. Wei C-H, Peng Y, Leaman R, *et al.* Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database* 2016;**2016**:baw032.

95. Agrawal M, Hegselmann S, Lang H, *et al.* Large Language Models are Few-Shot Clinical Information Extractors. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* 2022; 1998–2022.

96. Caufield JH, Hegde H, Emonet V, *et al. Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): A Method for Populating Knowledge Bases Using Zero-Shot Learning.* arXiv Preprint 2023; arXiv:2304.02711.

97. Jimenez Gutierrez B, McNeal N, Washington C, *et al. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again.* Findings of the Association for Computational Linguistics: EMNLP 2022 2022; 4497–4512.

98. Chen Q, Sun H, Liu H, *et al.* A comprehensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics* 2023;**39**:btad557.

99. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Inform* 2014;**47**:1–10.

100. Smith L, Tanabe LK, Ando RJ, *et al.* Overview of BioCreative II gene mention recognition. *Genome Biol* 2008;**9**(2):S2.

101. Collier N, Ohta T, Tsuruoka Y, *et al.* Introduction to the Bio-entity Recognition Task at JNLPBA. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)* 2004; 73–78.

102. Krallinger M, Rabal O, Akhondi S, *et al.* Overview of the BioCreative VI chemical-protein interaction Track. *Proceedings of the sixth BioCreative Challenge Evaluation Workshop* 2017; 142–47.

103. Herrero-Zazo M, Segura-Bedmar I, Martínez P, *et al.* The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions. *J Biomed Inform* 2013;**46**(5): 914–20.

104. Bravo À, Piñero J, Queralt-Rosinach N, *et al.* Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics* 2015;**16**(1):55.

105. Chen Q, Du J, Hu Y, *et al. Large Language Models in Biomedical Natural Language Processing: Benchmarks, Baselines, and Recommendations.* arXiv Preprint 2023; arXiv:2305.16326.

106. OpenAI. *Educator considerations for ChatGPT.* OpenAI ChatGPT Documentation 2023. https://platform.openai.com/docs/chatgpt-education/educator-considerations-for-chatgpt (7 May 2023, date last accessed).

107. https://www.khanacademy.org/.

108. Pandey M, Mishra V. Large language models in medical education and quality concerns. *J Qual Health Care Econ* 2023;**6**:1–3.

109. Eysenbach G. *The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers.* JMIR Medical Education 2023;**9**:e46885.

110. Khan RA, Jawaid M, Khan AR, *et al.* ChatGPT - reshaping medical education and clinical management. *Pak J Med Sci* 2023;**39**: 605–7.

111. Pustejovsky J, Castaño J, Saurí R, *et al.* Medstract: Creating Large-Scale Information Servers from Biomedical Texts. *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain* 2002; 85–92.

112. Gasperin C, Karamanis N, Seal RL. Annotation of Anaphoric Relations in Biomedical Full-Text Articles Using a Domain-Relevant Scheme. *Proceedings of DAARC* 2007; 19–24.

113. Su J, Yang X, Hong H, *et al.* Coreference Resolution in Biomedical Texts: a Machine Learning Approach. *Dagstuhl Seminar Proceedings* 2008;**8131**:1. 10.4230/DagSemProc.08131.4.

114. Segura-Bedmar I, Crespo M, de Pablo-Sánchez C, *et al.* Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. *BMC Bioinformatics* 2010;**11**(2):S1.

115. Nguyen N, Kim J-D, Tsujii J. Overview of BioNLP 2011 Protein Coreference Shared Task. *Proceedings of BioNLP Shared Task 2011 Workshop* 2011; 74–82.

116. Batista-Navarro RT, Ananiadou S. Building a Coreference-Annotated Corpus from the Domain of Biochemistry. *Proceedings of BioNLP 2011 Workshop* 2011; 83–91.

117. Cohen KB, Lanfranchi A, Choi MJ, *et al.* Coreference annotation and resolution in the Colorado richly annotated full text (CRAFT) corpus of biomedical journal articles. *BMC Bioinformatics* 2017;**18**(1):372.

118. Lu P, Poesio M. Coreference Resolution for the Biomedical Domain: A Survey. *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference* 2021; 12–23.

119. Baker S, Silins I, Guo Y, *et al.* Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics* 2016;**32**(3):432–40.

120. Shaib C, Li M, Joseph S, *et al.* Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 2023; 1387–1407.

121. Tamkin A, Brundage M, Clark J, *et al. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models*. arXiv Preprint 2021; arXiv:2102.02503.

122. OpenAI. *GPT-4 System Card*. OpenAI Research Papers 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf (7 May 2023, date last accessed).

123. Shah DS, Schwartz HA, Hovy D. *Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 2020; 5248–64.

124. Abid A, Farooqi M, Zou J. Persistent Anti-Muslim Bias in Large Language Models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* 2021; 298–306.

125. Obermeyer Z, Powers B, Vogeli C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;**366**(6464):447–53.

126. Sourlos N, Wang J, Nagaraj Y, *et al.* Possible bias in supervised deep learning algorithms for CT lung nodule detection and classification. *Cancer* 2022;**14**(16):3867.

127. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *N Engl J Med* 2020;**383**(9):874–82.

128. Huang J, Shao H, Chang KC-C. Are Large Pre-Trained Language Models Leaking Your Personal Information? *Findings of the Association for Computational Linguistics: EMNLP 2022* 2022; 2038–47.

129. Naik N, Hameed BMZ, Shetty DK, *et al.* Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg* 2022;**9**:9.

130. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 2023;**11**(6):887.

131. Li J, Dada A, Kleesiek J, *et al. ChatGPT in Healthcare: A Taxonomy and Systematic Review*. medRxiv Preprint 2023; 2023.03.30.23287899.

132. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. *Cureus* 2023;**15**(5):e39305.

133. Kung TH, Cheatham M, Medenilla A, *et al.* Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;**2**(2):e0000198.

134. O'Connor S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? *Nurse Educ Pract* 2023;**66**:103537.

135. Zhavoronkov A. Rapamycin in the context of Pascal's wager: generative pre-trained transformer perspective. *Onco Targets Ther* 2022;**9**:82–4.

136. GPT Generative Pretrained Transformer, Thunström AO, Steingrimsson S. *Can GPT-3 Write an Academic Paper on Itself, With Minimal Human Input?* HAL Open Science 2022. https://hal.science/hal-03701250/ (18 May 2023, date last accessed).

137. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature* 2023;**613**(7945):620–1.

138. O'Connor S. Corrigendum to "open artificial intelligence platforms in nursing education: tools for academic progress or abuse?" [nurse Educ. Pract. 66 (2023) 103537]. *Nurse Educ Pract* **2023**(67):103572.

139. 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* 2023;**6**(1):75.

140. Lee JY. Can an artificial intelligence chatbot be the author of a scholarly article? *J Educ Eval Health Prof* 2023;**20**:6.

141. Siegerink B, Pet LA, Rosendaal FR, *et al.* ChatGPT as an author of academic papers is wrong and highlights the concepts of accountability and contributorship. *Nurse Educ Pract* 2023;**68**:103599.

142. Goto A, Katanoda K. Should we acknowledge ChatGPT as an author? *J Epidemiol* 2023;**33**:333–4.

143. Ide K, Hawke P, Nakayama T. Can ChatGPT be considered an author of a medical article? *J Epidemiol* 2023;**33**:381–2.

144. Norris C. Large language models like ChatGPT in ABME: author guidelines. *Ann Biomed Eng* 2023;**51**(6):1121–2.

145. Page MJ, McKenzie JE, Bossuyt PM, *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;**372**:n71.

146. Papineni K, Roukos S, Ward T, *et al.* Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* 2002; 311–18.

147. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out* 2004; 74–81.

148. Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* 2005; 65–72.

149. Zhang T, Kishore V, Wu F, *et al. BERTScore: Evaluating Text Generation with BERT*. arXiv Preprint 2020; arXiv:1904.09675.

150. Yuan W, Neubig G, Liu P. BARTScore: Evaluating Generated Text as Text Generation. *Advances in Neural Information Processing Systems* 2021;**34**:27263–77.

151. Fu J, Ng S-K, Jiang Z, *et al. GPTScore: Evaluate as You Desire*. arXiv Preprint 2023; arXiv:2302.04166.

152. Jin Q, Wang Z, Floudas CS, *et al. Matching Patients to Clinical Trials with Large Language Models*. arXiv Preprint arXiv.org 2023; arXiv:2307.15051.