Computer Methods and Programs in Biomedicine 245 (2024) 108013

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

ChatGPT in healthcare: A taxonomy and systematic review

Jianning Li [a], Amin Dada [a], Behrus Puladi [b,c], Jens Kleesiek [a,e], Jan Egger [a,d,*]

[a] Institute for Artificial Intelligence in Medicine, University Hospital Essen (AöR), Girardetstraße 2, 45131 Essen, Germany
[b] Institute of Medical Informatics, University Hospital RWTH Aachen, Pauwelsstraße 30, 52074 Aachen, Germany
[c] Department of Oral and Maxillofacial Surgery, University Hospital RWTH Aachen, Pauwelsstraße 30, 52074 Aachen, Germany
[d] Center for Virtual and Extended Reality in Medicine (ZvRM), University Hospital Essen, University Medicine Essen, Hufelandstraße 55, 45147 Essen, Germany
[e] TU Dortmund University, Department of Physics, Otto-Hahn-Straße 4, 44227 Dortmund, Germany

ARTICLE INFO

Keywords:
ChatGPT
Healthcare
NLP
Transformer
LLM
OpenAI
Taxonomy
Bard
BERT
LLaMA

ABSTRACT

The recent release of ChatGPT, a chat bot research project/product of natural language processing (NLP) by OpenAI, stirs up a sensation among both the general public and medical professionals, amassing a phenomenally large user base in a short time. This is a typical example of the 'productization' of cutting-edge technologies, which allows the general public without a technical background to gain firsthand experience in artificial intelligence (AI), similar to the AI hype created by AlphaGo (DeepMind Technologies, UK) and self-driving cars (Google, Tesla, etc.). However, it is crucial, especially for healthcare researchers, to remain prudent amidst the hype. This work provides a systematic review of existing publications on the use of ChatGPT in healthcare, elucidating the 'status quo' of ChatGPT in medical applications, for general readers, healthcare professionals as well as NLP scientists. The large biomedical literature database *PubMed* is used to retrieve published works on this topic using the keyword 'ChatGPT'. An inclusion criterion and a taxonomy are further proposed to filter the search results and categorize the selected publications, respectively. It is found through the review that the current release of ChatGPT has achieved only moderate or 'passing' performance in a variety of tests, and is unreliable for actual clinical deployment, since it is not intended for clinical applications by design. We conclude that specialized NLP models trained on (bio)medical datasets still represent the right direction to pursue for critical clinical applications.

1. Introduction

In November 2022 a chat bot called ChatGPT was released. According to itself it is 'a conversational AI language model developed by OpenAI. It uses deep learning techniques to generate human-like responses to natural language inputs. The model has been trained on a large dataset of text and has the ability to understand and generate text for a wide range of topics. ChatGPT can be used for various applications such as customer service, content creation, and language translation'. Since its release, ChatGPT has taken humans by storm and its user base is growing even faster than the current record holder TikTok, reaching 100 million users in just two months after its launch. ChatGPT is already used to generate textual context, presentations and even source code for all kinds of topics. But what does that mean specifically for the healthcare sector? What if the general public or medical professionals turn to ChatGPT for treatment decisions? To answer these questions, we will look at published works that already reported the usage of ChatGPT in the medical field. In doing so, we will explore and discuss ethical concerns when using ChatGPT, specifically within the healthcare sector (e.g., in clinical routines). We also identify specific action items that we believe have to be undertaken by creators and providers of chat bots to avoid catastrophic consequences that go far beyond letting a chat bot do someone's homework. This review makes William B. Schwartz description from 1970 about conversational agents that will serve as consultants by enhancing the intellectual functions of physicians through interactions [120] as up-to-date as ever.

Even though the application of natural language processing (NLP) in healthcare is not new [42,128,140,99], the recent release of ChatGPT, a direct product of NLP, still generated a hype in artificial intelligence (AI) and sparked a heated discussion about ChatGPT's potential capability and pitfalls in healthcare, and attracted the attention of researchers from different medical specialties. The sensation could largely be attributed to ChatGPT's barrier-free (browser-based) and user-friendly interface, allowing medical professionals and the general public with-

* Corresponding author at: Institute for Artificial Intelligence in Medicine, University Hospital Essen (AöR), Girardetstraße, 45131 Essen, Germany.
E-mail address: jan.egger@uk-essen.de (J. Egger).

https://doi.org/10.1016/j.cmpb.2024.108013
Received 7 June 2023; Received in revised form 29 December 2023; Accepted 8 January 2024
Available online 15 January 2024

out a technical background to easily communicate with the *Transformer*- and reinforcement learning-based language model. Currently, the interface is designed for question answering (QA), i.e., ChatGPT responds in texts to the questions/prompts from users. All established or potential applications of ChatGPT in different medical specialties and/or clinical scenarios hinge on the QA feature, distinguished only by how the prompts are formulated (Format-wise: open-ended, multiple choice, etc. Content-wise: radiology, parasitology, toxicology, diagnosis, medical education and consultation, etc.). Numerous publications featuring these applications have also been generated and indexed in *PubMed* since the release. This systematic review dives into these publications, aiming to elucidate the current state of employment, as well as the limitations and pitfalls of ChatGPT in healthcare, amidst the ChatGPT AI hype.

Based on the findings derived from existing publications on ChatGPT in healthcare, this systematic review addresses the following research questions:

- RQ1: What are the different medical applications where ChatGPT has already been tested?
- RQ2: What are the strengths, limitations and main concerns of ChatGPT for healthcare, especially with respect to the field they are applied to?
- RQ3: What are the key research gaps that are being investigated or should be investigated according to the existing works?
- RQ4: How can existing publications on ChatGPT in healthcare be categorized according to a taxonomy?

The rest of the manuscript is organized as follows: Section 2 briefly introduces NLP, transformers and large language Models (LLMs), on which ChatGPT is built. Section 3 introduces the inclusion criteria and taxonomy used in the systematic review, and discusses in detail the selected publications. Section 4 presents the answers to the above research questions (RQ1 - RQ4). Section 5 and Section 6 summarize and conclude the review.

2. Background

For completeness of the review, this section briefly introduces natural language processing (NLP), the current state-of-the-art language architecture - transformers, large language models (LLMs) and their applications in the medical domain. The key methods and the common medical corpus involved in training large language model for medical applications are also discussed.

2.1. Natural language processing (NLP)

NLP [27] is an interdisciplinary research field that aims to develop algorithms for the computational understanding of written and spoken languages. Some of the most prominent applications include text classification, question answering, speech recognition, language translation, chat bots, and the generation or summarization of texts. Over the past decade, the progress of NLP has been accelerated by deep learning techniques, in conjunction with increasing hardware capabilities and the availability of massive text corpora. Given the fast growth of digital data and the growing need for automated language processing, NLP has become an indispensable technology in various industries, such as healthcare, finance, education, and marketing.

2.2. Transformer

In 2017, Vaswani et al. [138] introduced the Transformer model architecture, replacing previously widespread recurrent neural networks (RNN) [98], Long short-term memory networks (LSTM) [55] and Word2Vec [28]. Transformers are feedforward networks combined with specialized attention blocks that enable the model to attend to distinct segments of its input selectively. Attention blocks overcome two important limitations of RNNs. First, they enable Transformers to process input in parallel, whereas in RNNs each computation step depends on the previous one. Second, they allow Transformers to learn long-term dependencies. Since their introduction, Transformers consecutively achieved state-of-the-art results on various NLP benchmarks. Further developments include novel training tasks [29,69,143], adaptions of the network architecture [52,81], and reduction of computational complexity [73,81,51]. However, the limited training data and the model complexities remained one of the primary factors of model performance. Transformers have also been used for tasks beyond NLP, such as image and video processing [121], and they are an active area of research in the deep learning community.

A basic Transformer network comprises of an encoder and a decoder stack, each consisting of several identical blocks of feed-forward neural networks and multi-head attention [138]. Both the input and output of a Transformer are text sequences, where the words are tokenized and represented as elements in a high-dimensional vector. An embedding layer then projects the vector into a lower dimension space. The order of the sequences are also modeled into embeddings by a positional encoding. In other words, the embeddings are learned [17], and contain the semantic and positional information of the text sequences. Multi-head attention allows the network to cover different parts and information of long sequences, and it is one of the key components of a Transformer network.

2.3. Large language models (LLMs)

LLMs [20] refer to massive Transformer models trained on extensive datasets. Substantial research has been conducted on scaling the size of Transformer models [71]. The popular Bidirectional encoder representations from transformers (BERT) model [33], which in 2019 achieved record-breaking performance on seven tasks in the Glue Benchmark [139], possesses 110 million parameters. GPT-3 [22] had also already reached 175 billion parameters by 2021. At the same time, the size of the training datasets has continued to grow. BERT, for example, was trained on a dataset comprising of 3.3 billion words, while the recently published LLaMA [135] was trained on 1.4 trillion tokens. Despite their success, LLMs face several challenges, including the need for massive computational resources and the potential of adopting bias and misinformation from training data. Additionally, overconfidence when expressing wrong statements and a general lack of uncertainty remains to be a significant concern in NLP applications. As LLMs continue to improve and become more widespread, addressing these challenges and ensuring they are used ethically and responsibly is essential. ChatGPT is another representative LLM released by OpenAI, and other tech giants have also released their LLMs, such as the previously mentioned LLaMA from Meta, as a response. Fig. 1 illustrates the evolution of LLMs.

The training of LLMs unfolds in two distinct phases. Initially, in the pre-training phase, models are exposed to an extensive corpus of unlabeled text data, learning by predicting subsequent tokens in a given text through autoregressive training. This unsupervised approach allows for training on massive datasets without the need for manual labeling. Subsequently, the model undergoes instruction fine-tuning and alignment, a critical step to refine its understanding and application of various facts and concepts acquired during pre-training. This phase is crucial for mitigating undesired learning, such as certain human biases, insensitive language, and inaccuracies, ensuring the model's reliability, especially in sensitive domains like medicine where precision and sensitivity are crucial.

2.4. Medical LLMs and corpora

In the medical domain, specialized LLMs and medical corpora are developed and curated for different medical tasks and specialties. For

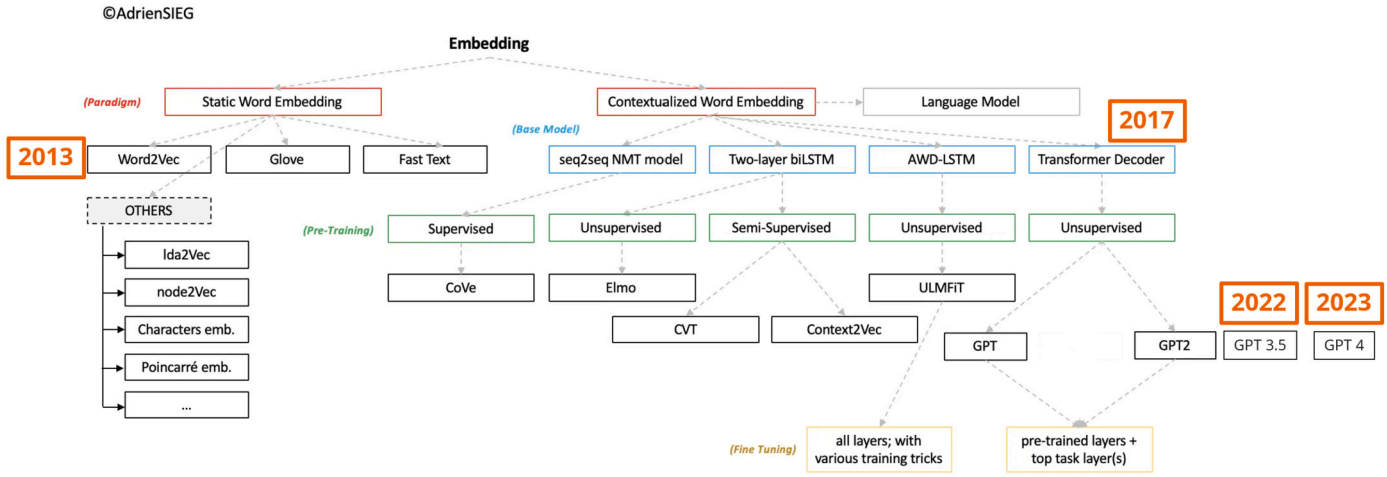

Fig. 1. Evolution of large language models (LLMs) (adapted from [123]).

example, MIMIC-III [65] and MIMIC-IV [66] are among the most influential datasets for electronic health record (EHR); The MedDialog dataset [146] is comprised of conversation-style texts between doctors and patients, which can be used in medical dialogue systems; The PubMed 200k RCT [32] and RCMR 280k [88] datasets consist of abstracts derived from *PubMed*, and they are commonly used for the classification of sentences in medical literature abstracts; SumPubMed [48] is a dataset derived from *PubMed*, and it is specifically designed for medical literature summarization; Multimodal datasets consist of both texts and images. For example, MIMIC-CXR [64] provides paired radiology reports and chest radiographs, and it can therefore be used for automated generation of radiology reports given radiographs. The ROCO dataset [109] provides paired images and captions derived from *PubMed Central*. Besides these English corpora, large-scale datasets have also been developed for other languages, such as Chinese (MedDialog [146], Huatuo-26M and Huatuo_Encyclopedia_QA [87]), Japanese (MedTxt-RR-JA [102]), French (The QUAERO French Medical Corpus) and German [41]. However, these non-English medical language datasets are relative scarce.

Bidirectional language representation (i.e., BERT [33]) is one of the most popular methods used to improve the performance of LLMs in medical tasks, and it has been adopted in several influential medical LLMs, such as PubMedBERT [46], ClinicalBERT [61] and BioBERT [84]; Incorporating existing medical knowledge base, such as the Unified Medical Language System (UMLS), into language models has also shown to improve results (KeBioLM [145]); Furthermore, study has shown that pre-training on diverse datasets, albeit not healthcare-related, still improves performance in a medical NLP task, compared to training on the target domain medical dataset alone [30].

2.5. ChatGPT

GPT-3 is a 175 billion parameter encoder-only model developed by OpenAI, trained on a diverse dataset of about 500 billion tokens. The data for this model is sourced from a wide array of texts, including various unspecified websites, collections of books, and Wikipedia. The model is particularly noted for its few-shot learning ability, allowing it to perform tasks with minimal examples or guidance effectively. Building upon GPT-3, ChatGPT is an instruction-tuned extension designed to interact and respond in a conversational manner.

3. Methodology

The search strategy used in this systematic review is illustrated in Fig. 2, according to the PRISMA guidelines. We use *PubMed* as the only source to search candidate publications. Since the majority of the papers are very short (without abstracts), eligibility is determined at first screening based on the inclusion criteria below.

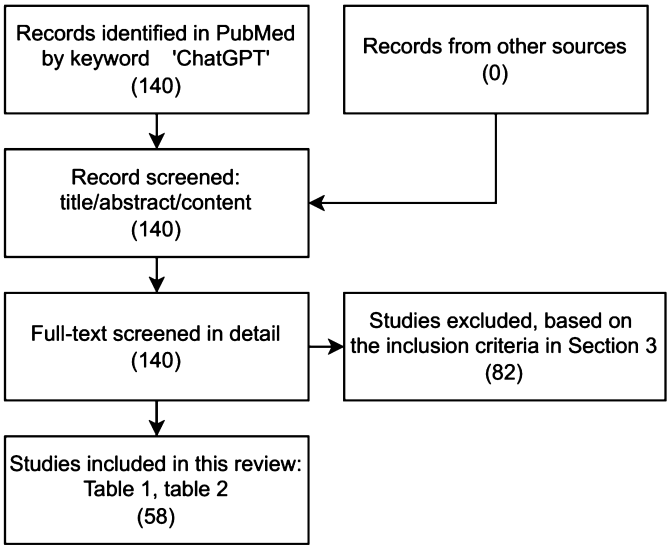

Fig. 2. Search strategy used in this systematic review.

3.1. Inclusion criteria

The review is expressly dedicated to the ChatGPT released in November 2022 by OpenAI, excluding its predecessors (*GPT-3.5*, *CPT-4*), other large language models (LLMs) such as *InstructGPT* and general NLP medical applications [91]. By March 20, 2023, a total of 140 publications are retrieved in *PubMed* (https://pubmed.ncbi.nlm.nih.gov/) using the keyword ChatGPT. Among them, article written in languages other than English (e.g., French [111]), without full text access (e.g., [79]), or whose main content has little to do with (or is not specific to) either ChatGPT (e.g., [56,131,40,45]) or healthcare (e.g., [124,130,34,8,49,15,114,26,85,144,129,53]) are excluded. Other representative exclusions include [54,70], which deal with CPT-3, and [72,37,116,2], where the authors claimed that ChatGPT assisted with the writing of the papers or case reports, but did not provide any discussion of the appropriateness of the generated texts and how the texts were incorporated into the main content. Generic comments that are not specific to healthcare, such as [133,144,19,60], where the authors comment on the authorship of ChatGPT and using ChatGPT in scientific writing, are also excluded. Several repetitive articles were found from the *PubMed* search results. Table 1 and Table 2 show the full

Table 1
Summary of *Level 1* and *Level 2* papers.

Ref.	Scenario	Category	Main Content	*Tag*
[119]	clinical workflow	editorial	discussion of the potential use, limitations and risks of ChatGPT in nursing practice	*Level 1*
[118]	medical research	perspective	comments about ChatGPT in scientific writing; Use ChatGPT to summarize and compare across papers	*Level 1*
[106]	medical research	editorial	generic comments on using ChatGPT in orthopaedic research	*Level 1*
[101]	medical research	letter to the editor	comments on using ChatGPT in scientific publications and generating research ideas	*Level 1*
[75]	miscellaneous	letter to the editor	comments on the potential use and pitfalls of ChatGPT in healthcare	*Level 1*
[134]	miscellaneous	editorial	discuss with ChatGPT about synthetic biology (e.g., applications, ethical regulations, history, research trends, etc.)	*Level 1*
[31]	medical research	editorial	comments on the pros and cons of using ChatGPT in medical research	*Level 1*
[83]	miscellaneous	original article	comments on the potential usage of ChatGPT in radiology (generate radiological reports, education, diagnostic decision-making, communicate with patients, compose radiological research article)	*Level 1*
[10]	medical education & research	letter to the editor	comments on the pros and cons of ChatGPT in medical education and research	*Level 1*
[43]	miscellaneous	primer	short comment on ChatGPT for Urologists	*Level 1*
[59]	consultation	correspondence	ChatGPT for antimicrobial consultation	*Level 1*
[58]	medical research	article (preprint)	comments on ChatGPT in peer-review	*Level 1*
[95]	miscellaneous	editorial	comments on ChatGPT in translational medicine	*Level 1*
[16]	consultation	letter to the editor	comments on the pros and cons of ChatGPT in public/community health (e.g., answer generic public health questions)	*Level 1*
[96]	miscellaneous	article	comments on the ethics of using ChatGPT in Health Professions Education	*Level 1*
[76]	medical research	letter to the editor	brief comments on using ChatGPT in medical writing	*Level 1*
[105]	medical education	editorial	comment on ChatGPT in nursing education	*Level 1*
[13]	miscellaneous	commentary	comment on ChatGPT in translational medicine	*Level 1*
[142]	miscellaneous	editorial	comment on ChatGPT in healthcare	*Level 1*
[74]	medical research	editorial	comment on ChatGPT in medical writing	*Level 1*
[9]	medical research	editorial	comment on using ChatGPT for scientific writing in sports & exercise medicine	*Level 1*
[14]	medical research	perspective	comment on medical writing	*Level 1*
[117]	miscellaneous	review	systematic review on ChatGPT in healthcare	*Level 1*
[7]	medical research	editorial	comment on the hallucination issue of ChatGPT in medical writing	*Level 1*
[94]	medical research	editorial	ChatGPT draft an article on vaccine effectiveness	*Level 2*
[137]	medical research	review	review on ChatGPT in medical research, including use examples	*Level 2*
[11]	medical research	original article	use ChatGPT to compile a review article on Digital Twin in healthcare	*Level 2*
[108]	clinical workflow	comment	use ChatGPT to generate a discharge summary for a patient who had hip replacement surgeries including follow-up care suggestions	*Level 2*
[115]	clinical workflow	letter to the editor	ChatGPT gives diagnosis, prognosis and explanation for a clinical toxicology case of acute organophosphate poisoning	*Level 2*
[24]	medical research	editorial	ChatGPT answers questions about computational systems biology in stem cell research but its answers lack depth	*Level 2*
[50]	medical research	letter to the editor	use ChatGPT to search literature of a given topic, but majority of returned publications are fabricated	*Level 2*
[100]	medical (anatomy) education	letter to the editor	ChatGPT answers anatomy-related questions; Result shows ChatGPT is currently incapable of giving accurate anatomy information	*Level 2*
[1]	consultation	letter to the editor	ChatGPT answers questions on cardiopulmonary resuscitation	*Level 2*
[23]	miscellaneous	Discussions with Leaders (Invitation Only)	comment and use examples of ChatGPT in nuclear medicine	*Level 2*
[3]	medical education	editorial	ChatGPT answers multiple-choice questions on nuclear medicine; Results suggest ChatGPT does not possesses the knowledge of a nuclear medicine physician	*Level 2*
[25]	medical research	brief report	comments on using ChatGPT in healthcare (e.g., compose medical notes) and medical research (e.g., generate abstracts, research topics)	*Level 2*
[57]	consultation	commentary	ChatGPT answers cancer-related questions information	*Level 2*
[18]	consultation	commentary	ChatGPT answers epilepsy-related questions	*Level 2*
[127]	consultation	article	comments on ChatGPT in diabetes self management and education (DSME)	*Level 2*
[38]	medical research	editorial	ChatGPT generates a curriculum about AI for medical students and a list of recommended readings	*Level 2*

Table 2
Summary of *Level 3* papers. Note: 'preprint' in the 'Journal' column is the status of the papers at the time of conducting the review.

Ref.	Scenario	Summary	Results/Conclusion	Version	Journal
[113]	clinical workflow	decide an imaging procedure or evaluate whether a procedure is proper for breast cancer/pain patients	specialized ChatGPT is needed	Jan. 9, 2023	preprint
[112]	clinical workflow	ChatGPT supports clinical decision-making, by answering questions from Merck Sharpe & Dohme (MSD) clinical vignettes	ChatGPT achieves an overall accuracy of (71.7%) on 36 clinical vignettes covering the entire clinical workflow	Jan. 9,2023	preprint
[62]	medical education	compare ChatGPT with medical students in (an internal) parasitology exam (79 questions)	ChatGPT is not comparable to medical student (Acc. 89.6%) in parasitology questions	Dec. 15, 2022	JEEHP
[44]	medical education	ChatGPT takes US Medical Licensing Examination (USMLE)	ChatGPT achieved passing score	Dec. 15, 2022	JMIR
[5]	clinical workflow	ChatGPT writes patient letters (e.g., communicates diagnostic results, gives treatment advice) for 38 clinical scenarios	ChatGPT achieved high scores on both the factual correctness and humanness criterion	-	Lancet Digit Health
[126]	medical education	compare ChatGPT with medical students in parasitology exam (288 questions) from the Doctor of Veterinary Medicine (DVM) exam	ChatGPT and students achieve similar scores	-	Cell
[90]	clinical workflow	ChatGPT answers clinical decision support (CSD) alerts from Epic EHR	ChatGPT's answers are biased and redundant, their acceptability in CDS is low	-	preprint
[77]	medical education	ChatGPT takes USMLE (June 2022)	ChatGPT achieved passing score, and its explanations contain novel insights	-	PLOS Digital Health
[39]	medical education	ChatGPT takes life-support exams (AHA BLS / CLS Exams 2016)	ChatGPT did not reach passing score	Jan. 9 and 30, 2023	Resuscita-tion
[68]	consultation	ChatGPT provides cancer-related information and feedback on cancer misconceptions	ChatGPT provides highly accurate cancer information	Dec. 15, 2022	JNCI Cancer Spectrum
[86]	medical research	compared 50 ChatGPT-generated abstracts with real abstracts from scientific publications	Grammarly can detect ChatGPT-generated abstracts with high accuracy	-	AJOG
[110]	consultation	evaluate ChatGPT using 100 questions about retinal diseases	ChatGPT is highly accurate on general questions but less accurate for treatment options	-	Acta Oph-thalmologica
[35]	consultation	compare ChatGPT with humans on 85 genetics/genomics questions	ChatGPT and humans perform similarly	-	preprint
[67]	consultation medical education	ChatGPT answers 284 question from various medical specialties	ChatGPT achieved overall high accuracies	-	preprint
[141]	medical education	ChatGPT takes Chinese National Medical Licensing Examination	ChatGPT's performance on the exam is well below passing level	-	preprint
[80]	medical research	ChatGPT identifies research questions in gastroenterology (e.g., microbiome, endoscopy)	ChatGPT generates highly relevant but non-novel research questions	Dec. 15, 2022	Scientific Reports
[47]	medical research	ChatGPT generates systematic review topics in plastic surgeries	ChatGPT performs moderately in generating novel systematic review ideas	-	Aesthetic Surgery Journal
[125]	consultation medical education	evaluate ChatGPT using 100 OE questions about pathology	ChatGPT scored around 80%	Jan. 30, 2023	Cureus

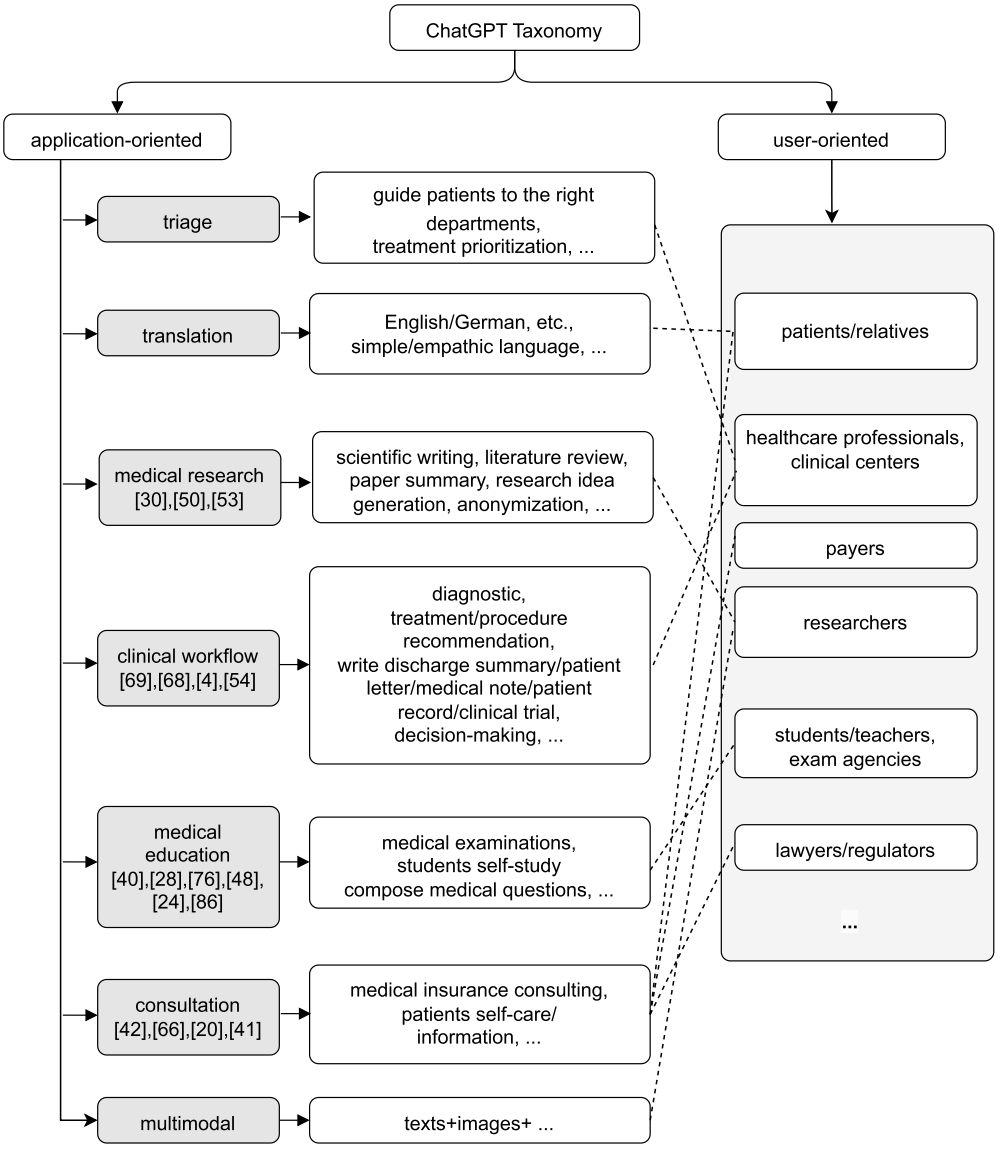

Fig. 3. Application- and user-oriented Taxonomy used in the ChatGPT review. The references shown in the application boxes are the *Level 3* publications.

list of selected publications based on the inclusion (exclusion) criteria. *PubMed* is a 'pure' medical search engine. However, *PubMed* does not index all publications in the medical field, e.g., it could be that a technical venue publishes a ChatGPT paper in the healthcare area (which is not indexed under *PubMed*). Hence, we incorporated IEEE as a 'technical' search engine to check if there are additional works that would fit into our review. The search for 'ChatGPT' resulted in five publications until March 2023. From these publications, two had performed something with ChatGPT in healthcare. However, the publications did not explicitly focus on healthcare, and the healthcare aspect was more a sub-section. Because our review focuses on publications where ChatGPT was solely used in healthcare, we did not include these IEEE contributions in our review.

3.2. Taxonomy

We propose a taxonomy, as shown in Fig. 3, to categorize the selected publications included in the review. The taxonomy is based on applications, including 'triage', 'translation', 'medical research', 'clinical workflow', 'medical education', 'consultation', 'multimodal', each targeting one or multiple end-user groups, such as patients, healthcare professionals, researchers, medical students and teachers, etc. An application-based taxonomy allows more compact and inclusive grouping of papers, compared to categorizing papers by specific medical specialties. For example, scientific progress and findings generated through clinical practices are documented in the form of publications and/or reports, and literature reviews and novel ideas are usually required for medical researchers of all disciplines to publish their works. Thus, papers on 'scientific writing', 'literature reviews', 'research ideas generation', etc., can be grouped into the 'medical research' category. Similarly, the 'consultation' category comprises papers where ChatGPT is used in medical consulting settings for both corporations (e.g., insurance companies, medical consulting agencies, etc.) and individuals (e.g., patients) seeking medical information and advice. The 'clinical workflow' category includes ChatGPT's applications in a variety of clinical scenarios, such as diagnostic decision-making, treatment and imaging procedure recommendation, and writing of discharge summary, patient letter and medical note. Furthermore, clinical departments, regardless of medical specialties, may benefit from a translation system for patients/visitors who are non-native language speakers ('translation'). A triage system [12] guiding patients to the right departments would reduce the burden of clinical facilities and centers in general. Note that different categories are not necessarily completely independent, since all applications are reliant upon the QA-based interface of

ChatGPT. By formulating the same questions differently according to different scenarios, ChatGPT's role can change. For instance, reformulating multiple choice questions about a medical specialty in medical exams to open-ended questions, ChatGPT's role changes from a medical student ('medical education') to a medical consultant ('consultation') or a clinician providing diagnosis or giving prescriptions ('clinical workflow'). To avoid such ambiguity, categorization of a paper is solely based on the scenario explicitly reported in the paper. The connections between the applications and end-users in Fig. 3 are also not unique. In this review, only the most obvious connections are established, such as 'medical education' - 'students/teachers/exam agencies', 'medical research' - 'researchers'. The following of the review will show that existing publications on ChatGPT in healthcare can all find a proper categorization based on the proposed taxonomy. Besides the taxonomy, we further assign a *tag* (*Level 1 - Level 3*) to the selected papers to indicate the depth and particularity of the papers on the '*ChatGPT in Healthcare*' topic:

- *Level 1*: Generic comments about the potential applications of ChatGPT in healthcare or in a specific medical specialty and/or scenario;
- *Level 2*: Comments with one or more example use cases of ChatGPT in a specific medical specialty and/or scenario and moderate discussion about the correctness of ChatGPT's answers;
- *Level 3*: Qualitative and quantitative evaluation of ChatGPT's answers to a decent amount of specialty- and/or scenario-specific questions, with insightful discussion about the correctness and appropriateness of the ChatGPT's answers.

Shortly prior to our review, a systematic review of ChatGPT in healthcare was published by Sallam, M. [117]. An inclusive taxonomy and a proper differentiation among the selected publications (*tag*: *Level 1*, *Level 2*, *Level 3*) is, however, lacking. We believe that the tag helps readers quickly filter and locate papers of interest. This review put more emphasis on *Level 3* papers, since they provide a clearer picture of the real capability of ChatGPT in different healthcare applications.

3.3. General profile of Level 1 and Level 2 papers

A list of *Level 1* and *Level 2* papers are summarized in Table 1. It is not unexpected that the majority of shortlisted papers fall into the *Level 1* and *Level 2* category. As seen from Table 1, most of *Level 1* and *Level 2* papers are short editorial comments or letters to the editor from multidisciplinary journals like *Nature* (https://www.nature.com/) and *Science* (https://www.science.org/), or specialty journals like nuclear medicine [3,75], plastic surgery [101,47], synthetic biology [134] and orthopaedic [106]. These publications usually deliver high-level comments about the potential impact and pitfalls of ChatGPT in healthcare [142], with a focus on medical publishing. Scientific journals are among the immediate stakeholders of the publishing industry on which ChatGPT will exert a significant impact. Thus, publishers introduce new regulations regarding the use of ChatGPT in scientific publications, in particular whether ChatGPT is eligible as an author and ChatGPT-generated texts are allowed. Answers from leading publishers like *Science* are in the negative [133,19]. *Nature* also bans ChatGPT authorship but takes a slightly more tolerant stance regarding ChatGPT-generated content, subject to a clear statement of whether, how and to what extent ChatGPT contributed to the submitted manuscript [130,34]. Main argument for the decision is that ChatGPT cannot properly source literature where its answers are derived from, causing unintentional plagiarism, nor can it take accountability as human authors do [133,34]. The decision is echoed by the academic community [74,124,144,85], agreeing that ChatGPT-generated content must be scrutinized by human experts before being used [74], as the generated content, such as references [133,14,50,38] could be fabricated. Lee, J.Y. et al. [85] reiterated from a legal (e.g., copy-right law) perspective the inappropriateness of listing ChatGPT as an author, emphasizing that a non-human cannot take legal responsibilities and consequences. However, banning ChatGPT from scientific writing is not easily enforceable, since ChatGPT is trained to produce human-like texts that even scientists and specifically-trained AI detector sometimes fail to detect [36,9]. In short, even though the prospect is promising [118,31,53,129], new regulations and substantial improvements are needed before ChatGPT can be safely and widely used for scientific writing, publishing, or medical research in general [133]. The *scenario* column in Table 1 corresponds to the taxonomy categorization. If the article concerns healthcare or a medical specialty in general, it is categorized as 'miscellaneous'. The *category* column indicates the type of the publications.

3.4. Reviews of Level 3 papers

Level 3 papers feature extensive experiments conducted to assess the suitability of ChatGPT for a medical specialty or clinical scenario. For open-ended (OE) questions, human experts are usually involved to assess the appropriateness of the answers. To quantify the subjective assessments, a scoring criteria and scheme (e.g., 5-point, 6-point or 10-point Likert scale) is usually required. For multiple choice questions, it is desirable to not only quantify the accuracies but to evaluate whether the 'justification' given by ChatGPT and the choice are in congruence. When it comes to comparisons (with humans or other language models), statistical analysis is usually performed. As shown in Table 2, many of *Level 3* papers are still pre-prints (under review) at the time of writing this review. Most of current ChatGPT evaluations are on 'medical education' (medical exams in particular), which requires no ethical approval to conduct. Representative works include [44,77], where the authors test ChatGPT in the US Medical Licensing Examination (USMLE). Even though the evaluations were carried out independently ([44] and [77] were published almost at the same time), similar results were reported, i.e., ChatGPT achieved only moderate passing performance. [44] further showed that ChatGPT outperformed two other language models, InstructGPT and GPT-3, in the exam. In both studies, ChatGPT was asked to give not only the answers but also the justifications, which were taken into consideration during evaluation (by physicians). [44] further found that ChatGPT performed better on fact-check questions than on complex 'know-how' type questions. It is worthy of noting that the exam contains questions from different medical specialties. However, Mbakwe, A.B. et al. [97] raised concerns that ChatGPT, a language model, passing the exam indicates the flawness of the exam system.[1] Besides USMLE, ChatGPT was also tested on the Chinese National Medical Licensing Examination [141] and the AHA BLS / CLS Exams 2016 [39], on both of which ChatGPT failed to achieve passing scores. ChatGPT achieved similar performance to students examinees on a Doctor of Veterinary Medicine (DVM) exam containing 288 parasitology exam questions. One major limitation of using ChatGPT in medical exams is that, current release of ChatGPT can only process text inputs, whereas some questions are diagram-/figure-based.[2] Such questions are either excluded or translated into text descriptions.

Besides the standard medical exams, ChatGPT achieved promising results on cancer-related questions [57,68]. In [68], ChatGPT's answers to common cancer myths and misconceptions were evaluated by expert reviewers and compared with the standard answers from the National Cancer Institute (NCI). Results showed that ChatGPT is able to achieve very high accuracies, showing that current ChatGPT is already a reliable source of cancer-related information for cancer patients [57]. Furthermore, [110] tested ChatGPT with 100 questions related to retina disease. The answers were evaluated based on a 5-point Likert scale by

[1] ChatGPT does not fulfill the 'USMLE Mission Statement', but still passes the exam.

[2] ChatGPT developers revealed that future versions of ChatGPT will have vision capabilities, and can comprehend images.

domain experts. It is found that ChatGPT answers with high accuracy on general questions, while the answers are less satisfactory, sometimes harmful, when it comes to treatment/prescription recommendations. On 85 multiple-choice questions concerning genetics/genomics, ChatGPT achieved similar performance to human respondents [35]. Interestingly, based on the test results, [35] also reached the conclusion that ChatGPT fares better on 'memorization (fact-lookup)' type questions than on those requiring critical thinking, similar to [110]. The performance of ChatGPT on these *question-answering* scenarios[3] shows its potential for medical consultation and education.

A few studies evaluate the use of ChatGPT in medical research, particularly in scientific writing [86] and generating research questions [80] and systematic review topics [47]. In [86], the authors use ChatGPT to generate full abstracts, providing only the title and result sections of the abstracts from 50 real scientific publications. Even though previous studies [36] have shown that scientists can not tell apart abstracts generated by ChatGPT from those written by humans, [86] found that the two groups can simply be differentiated based on Grammarly scores. Discriminative features of ChatGPT-generated texts include mixed use of English dialects and language perfectness e.g., very few typos, more unique words, proper prepositions usage and no misuse of conjunction and comma. These characteristics can be captured by Grammarly scores. The finding indicates that Grammarly could potentially be adopted by scientific journals to enforce the 'no-AI-generated-texts' policy. In [80], the authors use ChatGPT to identify research questions in gastroenterology. The answers generated by ChatGPT prove to be highly relevant but lack depth and novelty. In [47], ChatGPT is used to generate systematic review topics in plastic surgery. Similar to [80], ChatGPT-generated research topics are generally not novel. The *version* column in Table 2 shows the version of ChatGPT used for evaluation. [80] found that newer versions of ChatGPT tend to have better performance on the same questions. In contrast to using ChatGPT directly for writing, which is expressly banned by many scientific journals, exploring new research ideas/topics with the assistance of ChatGPT faces less ethical issues. However, [80,47] demonstrated that the current version of ChatGPT is not sufficiently qualified for such tasks. Humans still play dominating roles in ingenious and innovative research.

[113,112,5,90] evaluate the application of ChatGPT in clinical workflow. In [113], ChatGPT is used to decide the appropriate imaging procedure (e.g., Mammography, MRI, US, etc.) for breast cancer screening and breast pain, given a description of the patients' conditions. ChatGPT's responses were evaluated against the corresponding American College of Radiology (ACR) appropriateness criteria. Results showed that ChatGPT achieved moderate overall results, and its performance is noticeably better for breast cancer screening than breast pain. The finding is in accordance with previous discussions that ChatGPT is already highly accurate on cancer-related information [57,68]. The authors concluded that, even though ChatGPT showed impressive performance on the task, specialized AI tools are desired to support the clinical decision-making process more reliably. In a follow-up study [112], the authors tested ChatGPT with 36 clinical vignettes from the Merck Sharpe & Dohme (MSD), covering the entire clinical workflow (differential diagnosis, final diagnosis and subsequent clinical management of the patients). Overall, ChatGPT obtained a 71.8% accuracy in the test, and its performance on differential diagnosis is significantly lower than on final diagnosis. ChatGPT achieved the highest accuracy on a cancer vignette. The patients and their conditions in these vignettes are only hypothetical, which removes the ethical barrier to conduct the evaluation. In [5], ChatGPT is used to write patient clinic letters in 38 hypothetical clinical scenarios (e.g., basal cell carcinoma, malignant melanoma, etc.), where ChatGPT communicates the diagnosis results and treatment advice to the patients in a friendly and easily-understandable manner. The letters are evaluated from the perspective of factual correctness and humanness by clinicians, and ChatGPT achieved high scores on both criteria. In [90], ChatGPT is supplied with seven types of clinical decision support (CDS) alerts (e.g., pediatrics bronchiolitis, immunization, postoperative anesthesia nausea and vomiting, etc.) and asked to give suggestions. However, ChatGPT's answers, even though highly relevant to the alerts, were not adequately acceptable by the standard of CDS experts.

[3] Exams are essentially also *question-answering*.

4. Results

The following presents the answers to the four research questions (RQ1-RQ4) based on the discussion in Section 3.

4.1. Medical applications of ChatGPT

According to Table 1, Table 2 and the taxonomy (Fig. 3), it is straightforward to see that ChatGPT is mostly evaluated in medical education, consultation and research, as well as in various scenarios in the clinical workflow, such as diagnosis, decision-making and clinical documentation (patient letter, medical note, discharge summary, etc.). However, it is important to note these 'applications' are carried out in a 'laboratory environment', by providing ChatGPT questions from standard medical exams (question banks), CSD alerts from Epic EHR or clinical vignettes from Merck Sharpe & Dohme (MSD), through its QA interface. None of the reviewed publications have reported an actual deployment of ChatGPT in clinical settings and practices. Furthermore, due to the current strict policies on AI-generated content imposed by publishers, the unsolved ethical issues as well as its incapability in generating novel research topics, using ChatGPT for medical literature and research remains experimental as well. For medical consultation, the fact that ChatGPT is already capable of providing highly accurate cancer-related information can not be generalized to all medical specialties, since reliable sources of cancer information, such as the National Cancer Institute (NCI), are publicly accessible and could have already been part of ChatGPT's training set. Its qualification as a medical consultant remains to be further evaluated. Overall, the out-of-the-box performance of ChatGPT in healthcare is only moderate, which does not meet the high clinical standards. Improvements such as specialization [113] and standardization of evaluation are needed. Like other emerging technologies, a stable evaluation system that objectively reflects the applicability of ChatGPT in a clinical scenario must be established, in order for the tool to be deployed reliably in clinical practices. As the *Level 3* paper revealed (Section 2.3), the current evaluation of ChatGPT heavily relies on human input and lacks objectivity and scalability. Quantitative metrics reflecting the experts' qualitative evaluations of ChatGPT's performance can be computed automatically, and they are therefore desirable for ChatGPT's clinical integration [147]. A professional version of ChatGPT specialized in a medical specialty is promising for clinical use after passing the said quantitative evaluation (to be discussed in Section 5.1).

4.2. Strengths and limitations of ChatGPT in healthcare

Strengths The QA design of ChatGPT's interface makes it easy to be integrated into existing clinical workflow, providing feedback in real-time. ChatGPT can not only give answers to specific questions but provide 'justifications' to its answers. Sometimes, ChatGPT's 'justifications' and answers to open-ended question contain novel insights and perspectives, which might inspire novel research ideas. ChatGPT also shows superior performance in healthcare compared to other general large language models, such as InstructGPT, GPT-3.5.

Limitations The current release of ChatGPT can only take input and give feedback in texts, so that ChatGPT cannot handle questions requiring the interpretation of images. ChatGPT is incapable of 'reasoning'

like an expert system, and the 'justifications' provided by ChatGPT is merely a result of predicting the next words according to probability. It is possible that ChatGPT makes a correct choice, but gives completely nonsensical explanations. Accuracy of ChatGPT's answers depends largely on the quality of its training data, and the information ChatGPT is trained on decides how ChatGPT would respond to a question. However, ChatGPT itself cannot distinguish between real and fake information fed into it, so that its answers could be highly misleading, biased and dangerous when it comes to healthcare. For example, one of the most concerning issues of current release of ChatGPT, as confirmed by the reviewed publications, is that it can 'fabricate' information and convey it in a persuasive tone. Therefore, its answers should always be fact-checked by human experts before adoption. Furthermore, ChatGPT's answers, even if can be highly relevant, stay most of the time superficial and lack depth and novelty. Most importantly, ChatGPT is not fine-tuned for healthcare by design, and should not be used as such without specialization. Last but not least, the use of ChatGPT is not without barriers. Reformulating the prompt to the same question might change ChatGPT's answer as well. Proper formulation of prompts is another factor to obtaining desirable answers from ChatGPT. Last but not least, ChatGPT is a proprietary product, and therefore feeding sensitive patient information into its interface in order to obtain a feedback might violate privacy regulations.

4.3. Research gaps and future works

Prior to the deployment of any product in clinical settings, extensive evaluations of the product in a laboratory environment are required to identify the limitations and improve the product iteratively. Since ChatGPT was released no more than half a year ago, it has only been tested in a limited number of scenarios (Table 2). ChatGPT clearly is still at an experimental stage, and clinical deployment faces substantial unsolved technical and regulatory challenges. The *Level 3* publications provide a sound paradigm on how ChatGPT should be continued to be evaluated in different specialties, for future works to follow. However, before further pursuing the direction, researchers should be aware that, even though these evaluations provide, at best, a general picture of ChatGPT's capability in a medical specialty, little contribution to the improvement of the underlying language model is made. The limitations identified through these evaluations have also long been known in NLP research and are not specific to ChatGPT. Most importantly, whether or not ChatGPT has achieved good performance in an application scenario, it is unlikely that the ChatGPT with general knowledge will be clinically deployed in the future. Specialized AI models in healthcare, which the NLP community has long been working on, are more promising for practical and reliable clinical applications, compared to ChatGPT.

4.4. Categorization of publications based on a taxonomy

Finally, we have shown in our review that existing publications on ChatGPT in healthcare can be compactly grouped according to applications and target user groups. Thus, we come up with a application- and user-oriented taxonomy to categorize the selected publications, as discussed in Section 3.

5. Discussion

In this systematic review, we review published works (from Nov. 2022 to Mar. 2023) that used ChatGPT within the healthcare sector. In doing so, we extract publications from *PubMed* using the keyword 'ChatGPT' and propose a two-sided taxonomy (application-oriented and user-oriented) to categorize these publications, which we see as a building block for new publications on ChatGPT in healthcare. Even though the current taxonomy is already quite inclusive, it can be easily extended to emerging new applications or user groups. This first taxonomy is not limited to ChatGPT, rather it can also be applied to other (existing or upcoming) NLP models, like Bard from Google. On the one hand, the taxonomy helps interested readers to identify relevant works. On the other hand, it also helps identify areas where ChatGPT has not yet been applied to. The barrier-free user interface, the ability to produce human-like texts and the breadth of its knowledge on a variety of topics are the key reasons why ChatGPT has amassed a phenomenally large user base shortly after its release. Besides the architectural design of the LLM, the immeasurable human efforts invested in training the LLM through reinforcement learning contribute greatly to its impressive performance in human-like conversations. Even though ChatGPT technically represents the productization of a NLP model by OpenAI, rather than a fundamental technological advance or breakthrough, it is undeniable that ChatGPT is a living embodiment of state-of-the-art NLP techniques. The efforts devoted to making the product a reality still greatly push forward the field as a whole. Speaking from the perspective of a tech product, existing publications on ChatGPT's healthcare applications boil down to 'reviews and testing of a new NLP product in healthcare'. However, the product is not intended for medical applications by design, and it is therefore not unexpected that most 'test reports' evaluated ChatGPT as 'unqualified' or 'of merely passing grade' for healthcare. However, the reported limitations (see Section 4) of ChatGPT are not specific to the product, but are applicable to language models in general, as discussed in Section 2. These limitations can mostly be addressed by improving the underlying language model through NLP innovations. Nevertheless, the fact that ChatGPT is monetized[4] and therefore not (fully) open-sourced makes it difficult for the community to pinpoint the issues and come up with specific solutions for future improvement. In particular, the sources of datasets used for training the language model, which determine the type of questions and topics of the conversations ChatGPT can handle, remain unclear. As suggested by van Dis et al. [34], the community should invest in truly open LLMs that perform on par with proprietary NLP products like ChatGPT, in order to fully address these limitations. Currently, for healthcare applications, specialized AI models trained on biomedical datasets, such as BioGPT [92], are always more desirable than ChatGPT.

5.1. ChatGPT medical professional edition

In this section, we discuss the factors that should be taken into consideration while designing a future *ChatGPT Medical Professional Edition*, drawing insights from our review. In short, a *ChatGPT Medical Professional Edition* should not aim for an 'artificial general intelligence' that can handle any medical situations, in terms of both interface design, features and capabilities. Instead, it should be highly specialized and tailored to different medical specialties, situations and end users. (1) First and foremost, specialized training corpora should be carefully curated. The medical big data involved in training the language model should be specialized to the purpose, as discussed in Section 2.4. Considering that the knowledge base of different medical specialties could be overlapping, the training data can come from several closely related medical domains, and they should be up-to-date; (2) Second, the interface design and features should be specialized depending on its target users. If the target users are the general public, the interface should not involve sophisticated medical terminologies. Instead, plain words comprehensible to those without prior medical knowledge should be used. An example of this is a ChatGPT-based self-diagnosis system that gives advice to common symptoms [78,103]. It can also provide initial guidance on what professional medical help they should seek (similar to triage). However, it should clearly state that the final diagnosis and the subsequent treatment (or medication) options should be based on (human) professional advice. If the target users are medical professionals, it is important to make sure that the interface and features can be

[4] OpenAI has already introduced a subscription plan for ChatGPT (Plus).

seamlessly integrated into their existing workflow, without requiring complex setup.

5.2. Limitations of the review

It should be noted that, despite the aim of our review being to provide a general picture of the applications of ChatGPT in healthcare, the fact that a large portion of the reviewed papers are reviews themselves or editorial comments (as those of the *Level 1* and *Level 2* papers in Table 1) might limit or even bias our understanding of its actual clinical capability and applicability. In particular, the review papers and editorial comments generally portray an optimistic outlook on ChatGPT's application in healthcare, despite mentioning its potential pitfalls, whilst most of the application-oriented publications (i.e., the *Level 3* papers in Table 2) reported only 'passing performance' of ChatGPT in a limited medical application scenarios. Therefore, it is not feasible to extrapolate the findings drawn from these publications, especially the limitations of ChatGPT for a particular application and the proposed solutions, to the entire healthcare sector. Furthermore, the lack of a standardized and quantitative evaluation scheme, as discussed in Section 4.1, makes their findings inherently subjective and the results across different publications not directly comparable.

5.3. GPT-4 in healthcare

GPT-4 was released in March 2023. The enhanced version of ChatGPT is capable of processing not only texts, but also images, and it significantly outperforms existing LLMs, including ChatGPT, in a variety of standard NLP benchmarks, according to its official technical report [107]. In general, existing publications of GPT-4 in healthcare showed a similar pattern to those of ChatGPT, which consist of primarily short comments (e.g., [4,93,82]), reviews (e.g., [6]) and evaluations (e.g., [21,122,63]). A recent study reported that GPT-4 performs better or on par with state-of-the-art radiology-specific models on several common radiology tasks (e.g., summarization, disease classification, entity extraction from radiology report) [89]; GPT-4 also passed the USMLE with distinction [104], which stands in contrast to the 'passing performance' of ChatGPT in the same exam [77]. The study also reported that GPT-4 outperforms Med-PaLM [136], which is specifically fine-tuned on medical knowledge; However, like ChatGPT, GPT-4 has also had pitfalls. One study revealed that non-radiology physicians are more likely to adopt GPT-4 generated impressions of radiology reports, even if they might be false and harmful, due to better language coherence compared to the impressions written by radiologists [132]. Another study evaluated GPT-4's multimodal capabilities, and reported that its performance of identifying anatomies and pathologies from radiological images is still unreliable [21]. In [147], the authors also reported suboptimal performance of the visual feature of GPT-4, by using it to generate impressions from chest X-ray images.

6. Conclusion

To conclude, our review provides a general picture of the capability of the current release of ChatGPT in healthcare. By and large, the training set and the underlying language model decide the quality (accuracy, unbiasedness, humanness, etc.) of the responses of an AI chat bot to certain questions. Therefore, this review concludes that healthcare researchers in particular should retract from the AI hype generated by the product and focus their attention on NLP research in general and developing/evaluating specialized language models for healthcare applications.

CRediT authorship contribution statement

Jianning Li: Conceptualization, Writing – original draft, Writing – review & editing. **Amin Dada:** Investigation, Writing – original draft, Writing – review & editing. **Behrus Puladi:** Investigation, Supervision. **Jens Kleesiek:** Conceptualization, Investigation, Supervision. **Jan Egger:** Conceptualization, Investigation, Supervision.

Declaration of competing interest

The authors declare no conflict of interests.

Acknowledgements

This work was supported by the REACT-EU project KITE (Plattform für KI-Translation Essen, EFRE-0801977, https://kite.ikim.nrw/), the k-Radiomics project from the Bruno and Helene Jöster Foundation (https://k-radiomics.ikim.nrw/), "NUM 2.0" (FKZ: 01KX2121) and the Cancer Research Center Cologne Essen (CCCE). Behrus Puladi was funded by the Medical Faculty of RWTH Aachen University as part of the Clinician Scientist Program.

References

[1] Chiwon Ahn, Exploring ChatGPT for information of cardiopulmonary resuscitation, Resuscitation 185 (2023).
[2] Haris M. Akhter, Jeffrey S. Cooper, Jeffrey Cooper, Acute pulmonary edema after hyperbaric oxygen treatment: a case report written with ChatGPT assistance, Cureus 15 (2) (2023).
[3] Ian L. Alberts, et al., Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?, in: European Journal of Nuclear Medicine and Molecular Imaging, 2023, pp. 1–4.
[4] Hassam Ali, Generative pre-trained transformer 4 in healthcare: challenges, opportunities, and recommendations, Med. Adv. 1 (2) (2023).
[5] Stephen R. Ali, et al., Using ChatGPT to write patient clinic letters, in: The Lancet Digital Health, 2023.
[6] Fahad K. Aljindan, et al., Utilization of ChatGPT-4 in plastic and reconstructive surgery: a narrative review, Plast. Reconstr. Surg., Glob. Open 11 (10) (2023) e5305.
[7] Hussam Alkaissi, Samy I. McFarlane, Artificial hallucinations in ChatGPT: implications in scientific writing, Cureus 15 (2) (2023).
[8] Lauren B. Anderson, et al., Generative AI as a Tool for Environmental Health Research Translation, medRxiv, 2023, pp. 2023–02.
[9] Anderson Nash, et al., AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation, BMJ Open Sport Exerc. Med. 9 (1) (2023) e001568.
[10] Taha Bin Arif, Uzair Munaf, Ibtehaj Ul-Haque, The future of medical education and research: is ChatGPT a blessing or blight in disguise?, Med. Educ. Online 28 (1) (2023) 2181052.
[11] Ömer Aydın, Enis Karaarslan, OpenAI ChatGPT generated literature review: digital twin in healthcare, in: Emerging Computer Technologies 2, 2022, pp. 22–31.
[12] Adam Baker, et al., A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis, Front. Artif. Intell. 3 (2020) 543405.
[13] Christian Baumgartner, The potential impact of ChatGPT in clinical and translational medicine, Clin. Transl. Med. 13 (2023) 3.
[14] Som Biswas, ChatGPT and the future of medical writing, Radiology (2023) 223312.
[15] Som S. Biswas, Potential use of chat GPT in global warming, in: Annals of Biomedical Engineering, 2023, pp. 1–2.
[16] Som S. Biswas, Role of chat GPT in public health, in: Annals of Biomedical Engineering, 2023, pp. 1–2.
[17] Wout Bittremieux, et al., A learned embedding for efficient joint analysis of millions of mass spectra, Nat. Methods 19 (6) (2022) 675–678.
[18] Christian M. Boßelmann, Costin Leu, Dennis Lal, Are AI language models such as ChatGPT ready to improve the care of individuals with epilepsy?, in: Epilepsia, 2023.
[19] Jeffrey Brainard, Journals take up arms against AI-written text, Science (New York, NY) 379 (6634) (2023) 740–741.
[20] Thorsten Brants, et al., Large language models in machine translation, in: Proc. of the 2007 Joint Conf. on EMNLP-CoNLL, 2007, pp. 858–867.
[21] Dana Brin, et al., Assessing GPT-4 Multimodal Performance in Radiological Image Analysis, medRxiv, 2023, pp. 2023–11.
[22] Tom Brown, et al., Language models are few-shot learners, in: H. Larochelle, et al. (Eds.), Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901, https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
[23] Irène Buvat, Wolfgang Weber, Nuclear medicine from a novel perspective: Buvat and Weber talk with OpenAI's ChatGPT, J. Nucl. Med. (2023).
[24] Patrick Cahan, Barbara Treutlein, A conversation with ChatGPT on the role of computational systems biology in stem cell research, Stem Cell Rep. 18 (1) (2023) 1–2.

[25] Marco Cascella, et al., Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios, J. Med. Syst. 47 (1) (2023) 1–5.
[26] Joyjit Chatterjee, Nina Dethlefs, This new conversational AI model can be your friend, philosopher, and guide... and even your worst enemy, Patterns 4 (1) (2023) 100676.
[27] K.R. Chowdhary, Natural language processing, in: Fundamentals of Artificial Intelligence, 2020, pp. 603–649.
[28] Kenneth Ward Church, Word2Vec, Nat. Lang. Eng. 23 (1) (2017) 155–162.
[29] Kevin Clark, et al., Electra: pre-training text encoders as discriminators rather than generators, arXiv preprint, arXiv:2003.10555, 2020.
[30] Amin Dada, et al., On the impact of cross-domain data on German language models, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 13801–13813.
[31] Jari Dahmen, et al., Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword, in: Knee Surgery, Sports Traumatology, Arthroscopy, 2023, pp. 1–3.
[32] Franck Dernoncourt, Ji Young Lee, Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts, arXiv preprint, arXiv:1710.06071, 2017.
[33] Jacob Devlin, et al., Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint, arXiv:1810.04805, 2018.
[34] Eva A.M. van Dis, et al., ChatGPT: five priorities for research, Nature 614 (7947) (2023) 224–226.
[35] Dat Duong, Benjamin D. Solomon, Analysis of large-language model versus human performance for genetics questions, medRxiv, 2023, pp. 2023–01.
[36] Holly Else, Abstracts written by ChatGPT fool scientists, Nature 613 (7944) (2023) 423.
[37] Fábio Caleça Emidio, et al., Rectal bezoar: a rare cause of intestinal obstruction, Cureus 15 (3) (2023).
[38] Gunther Eysenbach, et al., The role of chatgpt, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers, JMIR Med. Educ. 9 (1) (2023) e46885.
[39] Nino Fijačko, et al., Can ChatGPT pass the life support exams without entering the American heart association course?, Resuscitation 185 (2023).
[40] Caitlin R. Francis, et al., Arf6 Regulates Endocytosis and Angiogenesis by Promoting Filamentous Actin Assembly, bioRxiv, 2023, pp. 2023–02.
[41] Johann Frei, Ludwig Frei-Stuber, Frank Kramer, GERNERMED++: semantic annotation in German medical NLP through transfer-learning, translation and word alignment, J. Biomed. Inform. 147 (2023) 104513.
[42] Carol Friedman, George Hripcsak, et al., Natural language processing and its future in medicine, Acad. Med. 74 (8) (1999) 890–895.
[43] Andrew T. Gabrielson, Anobel Y. Odisho, David Canes, Harnessing generative AI to improve efficiency among urologists: welcome ChatGPT, J. Urol. (2023) 10–1097.
[44] Aidan Gilson, et al., How does CHATGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment, JMIR Med. Educ. 9 (1) (2023) e45312.
[45] Rachel S. Goodman, et al., On the cusp: considering the impact of artificial intelligence language models in healthcare, Med 4 (3) (2023) 139–140.
[46] Yu Gu, et al., Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthc. 3 (1) (2021) 1–23.
[47] Rohun Gupta, et al., Application of ChatGPT in cosmetic plastic surgery: ally or antagonist, Aesthet. Surg. J. (2023) 042.
[48] Vivek Gupta, et al., SUMPUBMED: summarization dataset of PubMed scientific article, in: Proceedings of the 2021 Conference of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, 2020, https://vgupta123.github.io/docs/121_paper.pdf.
[49] John E. Hallsworth, et al., Scientific novelty beyond the experiment, in: Microbial Biotechnology, 2023.
[50] Michael Haman, Milan Školník, Using ChatGPT to conduct a literature review, in: Accountability in Research, 2023, pp. 1–3.
[51] Pengcheng He, Jianfeng Gao, Weizhu Chen, DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, 2023, https://openreview.net/forum?id=sE7-XhLxHA.
[52] Pengcheng He, et al., Deberta: decoding-enhanced Bert with disentangled attention, in: International Conference on Learning Representations, 2021, https://openreview.net/forum?id=XPZIaotutsD.
[53] Elisa L. Hill-Yardin, et al., A chat (GPT) about the future of scientific publishing, in: Brain, Behavior, and Immunity, 2023, S0889–1591.
[54] Takanobu Hirosawa, et al., Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study, Int. J. Environ. Res. Public Health 20 (4) (2023) 3378.
[55] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[56] Andreas Holzinger, et al., AI for life: trends in artificial intelligence for biotechnology, New Biotechnol. 74 (2023) 16–24.
[57] Ashley M. Hopkins, et al., Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift, JNCI Cancer Spectr. 7 (2) (2023) pkad010.
[58] Mohammad Hosseini, Serge P.J.M. Horbach, Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other Large Language Models in scholarly peer review, Res. Integr. Peer Rev. 8 (1) (2023) 4.
[59] Alex Howard, William Hope, Alessandro Gerada, ChatGPT and antimicrobial advice: the end of the consulting infection doctor?, in: The Lancet Infectious Diseases, 2023.
[60] Guangwei Hu, Challenges for enforcing editorial policies on AI-generated papers, in: Accountability in Research, 2023.
[61] Kexin Huang, Jaan Altosaar, Rajesh Ranganath, Clinicalbert: modeling clinical notes and predicting hospital readmission, arXiv preprint, arXiv:1904.05342, 2019.
[62] Sun Huh, Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study, J. Educ. Eval. Health Prof. 20 (2023) 1.
[63] Naoki Ito, et al., The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: evaluation study, JMIR Med. Educ. 9 (2023) e47532.
[64] Alistair E.W. Johnson, et al., MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (1) (2019) 317.
[65] Alistair E.W. Johnson, et al., MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.
[66] Alistair Johnson, et al., Mimic-iv, PhysioNet, Available online at: https://physionet.org/content/mimiciv/1.0/. (Accessed 23 August 2021), 2020.
[67] Douglas Johnson, et al., Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Research square preprint, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10002821/, 2023.
[68] Skyler B. Johnson, et al., Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information, JNCI Cancer Spectr. 7 (2) (2023) pkad015.
[69] Mandar Joshi, et al., Spanbert: improving pre-training by representing and predicting spans, Trans. Assoc. Comput. Linguist. 8 (2020) 64–77.
[70] David Jungwirth, Daniela Haluza, Artificial intelligence and public health: an exploratory study, Int. J. Environ. Res. Public Health 20 (5) (2023) 4541.
[71] Jared Kaplan, et al., Scaling laws for neural language models, arXiv preprint, arXiv:2001.08361, 2020.
[72] Karkra Rohan, et al., Recurrent strokes in a patient with metastatic lung cancer, Cureus 15 (2) (2023).
[73] Nikita Kitaev, Lukasz Kaiser, Anselm Levskaya, Reformer: the efficient transformer, in: International Conference on Learning Representations, 2020, https://openreview.net/forum?id=rkgNKkHtvB.
[74] Felipe C. Kitamura, ChatGPT is shaping the future of medical writing but still requires human judgment, Radiology (2023) 230171.
[75] Jens Kleesiek, et al., An opinion on ChatGPT in health care - written by humans only, J. Nucl. Med. (2023), https://doi.org/10.2967/jnumed.123.265687.
[76] Malcolm Koo, The importance of proper use of ChatGPT in medical writing, Radiology (2023) 230312.
[77] Tiffany H. Kung, et al., Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models, Digit. Health 2 (2) (2023) e0000198.
[78] Tomoyuki Kuroiwa, et al., The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study, J. Med. Internet Res. 25 (2023) e47621.
[79] Adi Lahat, Eyal Klang, Can advanced technologies help address the global increase in demand for specialized medical care and improve telehealth services?, J. Telemed. Telecare (2023) 1357633X231155520.
[80] Adi Lahat, et al., Evaluating the use of large language model in identifying top research questions in gastroenterology, Sci. Rep. 13 (1) (2023) 4164.
[81] Zhenzhong Lan, et al., Albert: a lite BERT for self-supervised learning of language representations, in: International Conference on Learning Representations, 2020, https://openreview.net/forum?id=H1eA7AEtvS.
[82] Aaron Lawson McLean, Artificial intelligence in surgical documentation: a critical review of the role of large language models, in: Annals of Biomedical Engineering, 2023, pp. 1–2.
[83] Augustin Lecler, Loïc Duron, Philippe Soyer, Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT, in: Diagnostic and Interventional Imaging, 2023.
[84] Jinhyuk Lee, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[85] Ju Yoen Lee, Can an artificial intelligence chatbot be the author of a scholarly article?, Sci. Ed. 10 (1) (2023) 7–12.
[86] Gabriel Levin, et al., Identifying ChatGPT-written OBGYN abstracts using a simple tool, Am. J. Obstet. Gynecol. (2023).
[87] Jianquan Li, et al., Huatuo-26M, a large-scale Chinese medical QA dataset, arXiv preprint, arXiv:2305.01526, 2023.
[88] Jie Li, Gaihong Yu, Zhixiong Zhang, RCMR 280k: refined corpus for move recognition based on PubMed abstracts, Data Intell. 5 (3) (2023) 511–536.
[89] Qianchu Liu, et al., Exploring the boundaries of GPT-4 in radiology, in: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.
[90] Siru Liu, et al., Assessing the Value of ChatGPT for Clinical Decision Support Optimization, medRxiv, 2023, pp. 2023–02.

[91] Saskia Locke, et al., Natural language processing in medicine: a review, Trends Anaesth. Crit. Care 38 (2021) 4–9.
[92] Renqian Luo, et al., BioGPT: generative pre-trained transformer for biomedical text generation and mining, Brief. Bioinform. 23 (6) (2022).
[93] Shaoting Luo, et al., Revolutionizing pediatric orthopedics: GPT-4, a groundbreaking innovation or just a fleeting trend?, Int. J. Surg. 109 (11) (2023) 3694–3697.
[94] Calum Macdonald, et al., Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis, J. Glob. Health 13 (2023).
[95] Douglas L. Mann, Artificial intelligence discusses the role of artificial intelligence in translational medicine: a JACC: basic to translational science interview with ChatGPT, in: Basic to Translational Science, 2023.
[96] Ken Masters, Ethical use of artificial intelligence in health professions education: AMEE guide no. 158, in: Medical Teacher, 2023, pp. 1–11.
[97] Amarachi B. Mbakwe, et al., ChatGPT passing USMLE shines a spotlight on the flaws of medical education, Digit. Health 2 (2) (2023).
[98] Larry R. Medsker, L.C. Jain, Recurrent neural networks, Des. Appl. 5 (2001) 64–67.
[99] Stéphane Meystre, Peter J. Haug, Natural language processing to extract medical problems from electronic clinical documents: performance evaluation, J. Biomed. Inform. 39 (6) (2006) 589–599.
[100] Sreenivasulu Reddy Mogali, Initial impressions of ChatGPT for anatomy education, in: Anatomical Sciences Education, 2023.
[101] Daniel Najafali, et al., Let's chat about chatbots: additional thoughts on ChatGPT and its role in plastic surgery along with its ability to perform systematic reviews, in: Aesthetic Surgery Journal, 2023, p. 056.
[102] Yuta Nakamura, et al., Clinical comparable corpus describing the same subjects with different expressions, in: MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation, IOS Press, 2022, pp. 253–257.
[103] Bhawna Nigam, Naman Mehra, M. Niranjanamurthy, Self-diagnosis in healthcare systems using AI chatbots, in: IoT and AI Technologies for Sustainable Living: A Practical Handbook, 2022, p. 79.
[104] Harsha Nori, et al., Capabilities of gpt-4 on medical challenge problems, arXiv preprint, arXiv:2303.13375, 2023.
[105] Siobhan O'Connor, et al., Open artificial intelligence platforms in nursing education: tools for academic progress or abuse?, Nurse Educ. Pract. 66 (2022) 103537.
[106] Matthieu Ollivier, et al., A deeper dive into ChatGPT: history, use and future perspectives for orthopaedic research, in: Knee Surgery, Sports Traumatology, Arthroscopy, 2023, pp. 1–3.
[107] OpenAI, GPT-4 technical report, ArXiv, https://arxiv.org/abs/2303.08774, 2023.
[108] Sajan B. Patel, Kyle Lam, ChatGPT: the future of discharge summaries?, Lancet Digit. Health 5 (3) (2023) e107–e108.
[109] Obioma Pelka, et al., Radiology objects in COntext (ROCO): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, in: Proceedings, vol. 3, Springer, 2018, pp. 180–189.
[110] Ivan Potapenko, et al., Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT, in: Acta Ophthalmologica, 2023.
[111] Paco Prada, Nader Perroud, Gabriel Thorens, Artificial intelligence and psychiatry: questions from psychiatrists to ChatGPT, Rev. Med. Suisse 19 (818) (2023) 532–536.
[112] Arya S. Rao, et al., Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow, medRxiv, 2023.
[113] Arya S. Rao, et al., Evaluating ChatGPT as an adjunct for radiologic decision-making, medRxiv, 2023.
[114] Matthias C. Rillig, et al., Risks and benefits of large language models for the environment, in: Environmental Science & Technology, 2023.
[115] Mary Sabry Abdel-Messih, Maged N. Kamel Boulos, ChatGPT in clinical toxicology, JMIR Med. Educ. 9 (2023) e46876.
[116] Abdullah Saeed, et al., Pacemaker malfunction in a patient with congestive heart failure and hypertension, Cureus J. Med. Sci. 15 (2) (2023).
[117] Sallam Malik, ChatGPT utility in health care education, research, and practice: systematic review on the promising perspectives and valid concerns, Healthcare 11 (6) (2023) 887, MDPI.
[118] Michele Salvagno, Fabio Silvio Taccone, Alberto Giovanni Gerli, et al., Can artificial intelligence help for scientific writing?, Crit. Care 27 (1) (2023) 1–5.
[119] Anthony Scerri, Karen H. Morin, Using chatbots like ChatGPT to support nursing practice, J. Clin. Nurs. (2023).
[120] William B. Schwartz, Medicine and the computer, N. Engl. J. Med. 283 (23) (1970) 1257–1264, https://doi.org/10.1056/NEJM197012032832305, PMID: 4920342.
[121] Javier Selva, et al., Video transformers: a survey, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.
[122] Yat-Fung Shea, et al., Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis, JAMA Netw. Open 6 (8) (2023) e2325000.
[123] Adrien Sieg, FROM pre-trained word embeddings TO pre-trained language models — focus on BERT, in: Towards Data Science, 2019.
[124] Bob Siegerink, et al., ChatGPT as an author of academic papers is wrong and highlights the concepts of accountability and contributorship, Nurse Educ. Pract. 68 (2023) 103599.
[125] Ranwir K. Sinha, et al., Applicability of ChatGPT in assisting to solve higher order problems in pathology, Cureus 15 (2) (2023).
[126] Jan Šlapeta, Are ChatGPT and other pretrained language models good parasitologists?, in: Trends in Parasitology, 2023.
[127] Gerald Gui Ren Sng, et al., Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education, in: Diabetes Care, 2023, p. dc230197.
[128] Peter Spyns, Natural language processing in medicine: an overview, Methods Inf. Med. 35 (04/05) (1996) 285–301.
[129] Chris Stokel-Walker, AI bot ChatGPT writes smart essays-should academics worry?, Nature (2022).
[130] Chris Stokel-Walker, ChatGPT listed as author on research papers: many scientists disapprove, Nature 613 (7945) (2023) 620–621.
[131] Martin Strunga, et al., Artificial intelligence systems assisting in the assessment of the course and retention of orthodontic treatment, Healthcare 11 (5) (2023) 683, MDPI.
[132] Zhaoyi Sun, et al., Evaluating GPT-4 on impressions generation in radiology reports, Radiology 307 (5) (2023) e231259.
[133] Holden H. Thorp, ChatGPT is fun, but not an author, Science 379 (6630) (2023) 313.
[134] Yaojun Tong, Lixin Zhang, Discovering the next decade's synthetic biology research trends with ChatGPT, Synth. Syst. Biotechnol. 8 (2) (2023) 220.
[135] Hugo Touvron, et al., LLaMA: open and efficient foundation language models, arXiv:2302.13971 [cs.CL], 2023.
[136] Tao Tu, et al., Towards generalist biomedical AI, arXiv preprint, arXiv:2307.14334, 2023.
[137] Raju Vaishya, Anoop Misra, Abhishek Vaish, ChatGPT: is this version good for healthcare and research?, Diabetes Metab. Syndr. Clin. Res. Rev. (2023) 102744.
[138] Ashish Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[139] Alex Wang, et al., GLUE: a multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, Nov. 2018, pp. 353–355, https://aclanthology.org/W18-5446.
[140] Jing Wang, et al., Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on PubMed, J. Med. Internet Res. 22 (1) (2020) e16816.
[141] Xinyi Wang, et al., ChatGPT Performs on the Chinese National Medical Licensing Examination, medRxiv, 2023.
[142] Will ChatGPT Transform Healthcare?, Nat. Med. 29 (2023) 505–506, https://doi.org/10.1038/s41591-023-02289-5.
[143] Zhilin Yang, et al., Xlnet: generalized autoregressive pretraining for language understanding, Adv. Neural Inf. Process. Syst. 32 (2019).
[144] Nicole Shu Ling Yeo-Teh, Bor Luen Tang, Letter to editor: NLP systems such as ChatGPT cannot be listed as an author because these cannot fulfill widely adopted authorship criteria, Account. Res. (2023), https://doi.org/10.1080/08989621.2023.2177160, in press.
[145] Zheng Yuan, et al., Improving biomedical pretrained language models with knowledge, in: Proceedings of the 20th Workshop on Biomedical Language Processing, 2021, pp. 180–190.
[146] Guangtao Zeng, et al., MedDialog: large-scale medical dialogue datasets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 9241–9250.
[147] Sebastian Ziegelmayer, et al., Evaluation of GPT-4's chest X-ray impression generation: a reader study on performance and perception, J. Med. Internet Res. 25 (2023) e50865.