ChatGPT in Healthcare: A Taxonomy and Systematic Review

Jianning Li, Amin Dada, Jens Kleesiek, Jan Egger*

Institute of Artificial Intelligence in Medicine, University Hospital Essen (AöR), Girardetstraße, 45131 Essen, Germany.

*Corresponding author: jan.egger (at) uk-essen.de (J.E.)

March 2023

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.

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

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 [94] as up-to-date as ever.

Even though the application of natural language processing (NLP) in healthcare is not new [34, 101, 111, 77], 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 specialities. The sensation could largely be attributed to ChatGPT's barrier-free (browser-based) and user-friendly interface, allowing medical professionals and the general public without 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 specialities 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.

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

Ref.	Scenario	Category	Main Content	*Tag*
[93]	clinical workflow	editorial	discussion of the potential use, limitations and risks of ChatGPT in nursing practice	*Level 1*
[92]	medical research	perspective	comments about ChatGPT in scientific writing; Use ChatGPT to summarize and compare across papers	*Level 1*
[81]	medical research	editorial	generic comments on using ChatGPT in orthopaedic research	*Level 1*
[79]	medical research	letter to the editor	comments on using ChatGPT in scientific publications and generating research ideas	*Level 1*
[59]	miscellaneous	letter to the editor	comments on the potential use and pitfalls of ChatGPT in healthcare	*Level 1*
[106]	miscellaneous	editorial	discuss with ChatGPT about synthetic biology (e.g., applications, ethical regulations, history, research trends, etc.)	*Level 1*
[25]	medical research	editorial	comments on the pros and cons of using ChatGPT in medical research	*Level 1*
[65]	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*
[8]	medical education & research	letter to the editor	comments on the pros and cons of ChatGPT in medical education and research	*Level 1*
[35]	miscellaneous	primer	short comment on ChatGPT for Urologists	*Level 1*
[49]	consultation	correspondence	ChatGPT for antimicrobial consultation	*Level 1*
[48]	medical research	article (preprint)	comments on ChatGPT in peer-review	*Level 1*
[72]	miscellaneous	editorial	comments on ChatGPT in translational medicine	*Level 1*

[14]	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*
[73]	miscellaneous	article	comments on the ethics of using ChatGPT in Health Professions Education	*Level 1*
[60]	medical research	letter to the editor	brief comments on using ChatGPT in medical writing	*Level 1*
[80]	medical education	editorial	comment on ChatGPT in nursing education	*Level 1*
[11]	miscellaneous	commentary	comment on ChatGPT in translational medicine	*Level 1*
[113]	miscellaneous	editorial	comment on ChatGPT in healthcare	*Level 1*
[58]	medical research	editorial	comment on ChatGPT in medical writing	*Level 1*
[7]	medical research	editorial	comment on using ChatGPT for scientific writing in sports & exercise medicine	*Level 1*
[12]	medical research	perspective	comment on medical writing	*Level 1*
[91]	miscellaneous	review	systematic review on ChatGPT in healthcare	*Level 1*
[5]	medical research	editorial	comment on the hallucination issue of ChatGPT in medical writing	*Level 1*
[71]	medical research	editorial	ChatGPT draft an article on vaccine effectiveness	*Level 2*
[108]	medical research	review	review on ChatGPT in medical research, including use examples	*Level 2*
[9]	medical research	original article	use ChatGPT to compile a review article on Digital Twin in healthcare	*Level 2*
[82]	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*
[89]	clinical workflow	letter to the editor	ChatGPT gives diagnosis, prognosis and explanation for a clinical toxicology case of acute organophosphate poisoning	*Level 2*

[19]	medical research	editorial	ChatGPT answers questions about computational systems biology in stem cell research but its answers lack depth	*Level 2*
[40]	medical research	letter to the editor	use ChatGPT to search literature of a given topic,but majority of returned publications are fabricated	*Level 2*
[78]	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*
[75]	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*
[20]	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*
[47]	consultation	commentary	ChatGPT answers cancer-related questions information	*Level 2*
[15]	consultation	commentary	ChatGPT answersepilepsy-related questions	*Level 2*
[100]	consultation	article	comments on ChatGPT in diabetes self management and education (DSME)	*Level 2*
[31]	medical research	editorial	ChatGPT generates a curriculum about AI for medical students and a list of recommended readings)	*Level 2*

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), and Section 5 summarizes and concludes the review.

2 Background

2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) [22] 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. [109] introduced the Transformer model architecture, replacing previously widespread recurrent neural networks (RNN) [76], Long short-term memory networks (LSTM) [45] and Word2Vec [23]. 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 [24, 54, 114], adaptions of the network architecture [42, 64], and reduction of computational complexity [57, 64, 41]. 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 [95], and they are an active area of research in the deep learning community.

2.3 Large Language Models (LLMs)

Large language models (LLMs) [17] refer to massive Transformer models trained on extensive datasets. Substantial research has been conducted on scaling the size of Transformer models. The popular BERT model [26], which in 2019 achieved record-breaking performance on seven tasks in the Glue Benchmark [110], possesses 110 million parameters. On the other hand, GPT-3 [18] had 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 [107] was trained on 1.4 trillion tokens. Despite the success of the LLMs, 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 released their LLMs, such as the previously mentioned LLaMA from Meta, as a response. Figure 1 illustrates the evolution of LLMs.

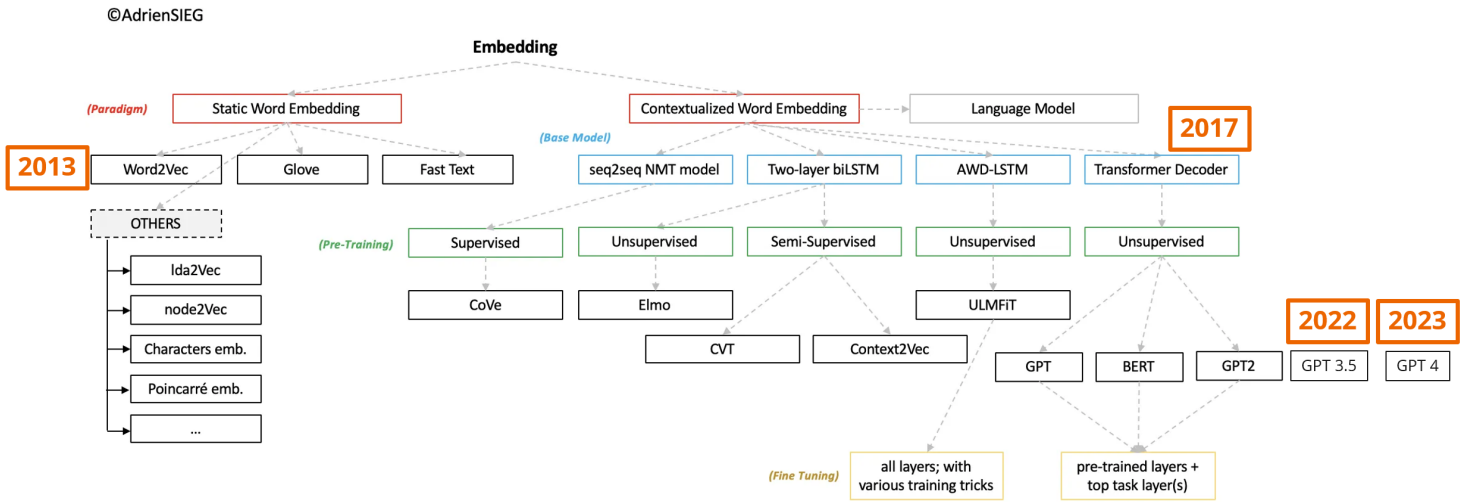

Figure 1: Evolution of large language models (LLMs) (adapted from [96]).

3 Methodology

The search strategy used in this systematic review is illustrated in Figure 2, following 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.

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 [69]. 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 [84]), without full text access (e.g., [62]), or whose main content has little to do with (or is not specific to) either ChatGPT (e.g., [46, 104, 33, 37]) or healthcare (e.g., [97, 103, 27, 6, 39, 13, 88, 21, 66, 115, 102, 43]) are excluded. Other representative exclusions include [44, 55], which deal with CPT-3, and [56, 30, 90, 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 [105, 115, 16, 50], 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 list of selected publications based on the inclusion (exclusion) criteria.

3.2 Taxonomy

We propose a taxonomy, as shown in Figure 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 specialities. 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

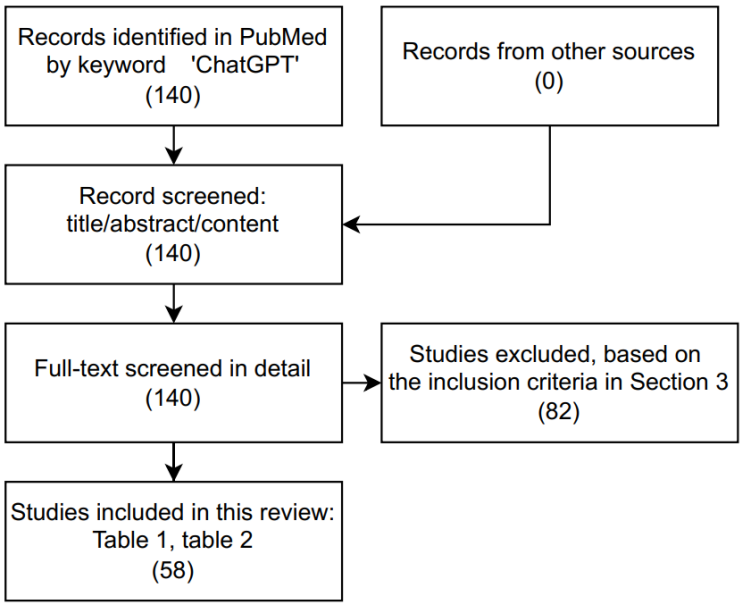

Figure 2: Search strategy used in this systematic review.

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 specialities, may benefit from a translation system for patients/visitors who are non-native language speakers ('translation'). A triage system [10] 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 speciality 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 Figure 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.

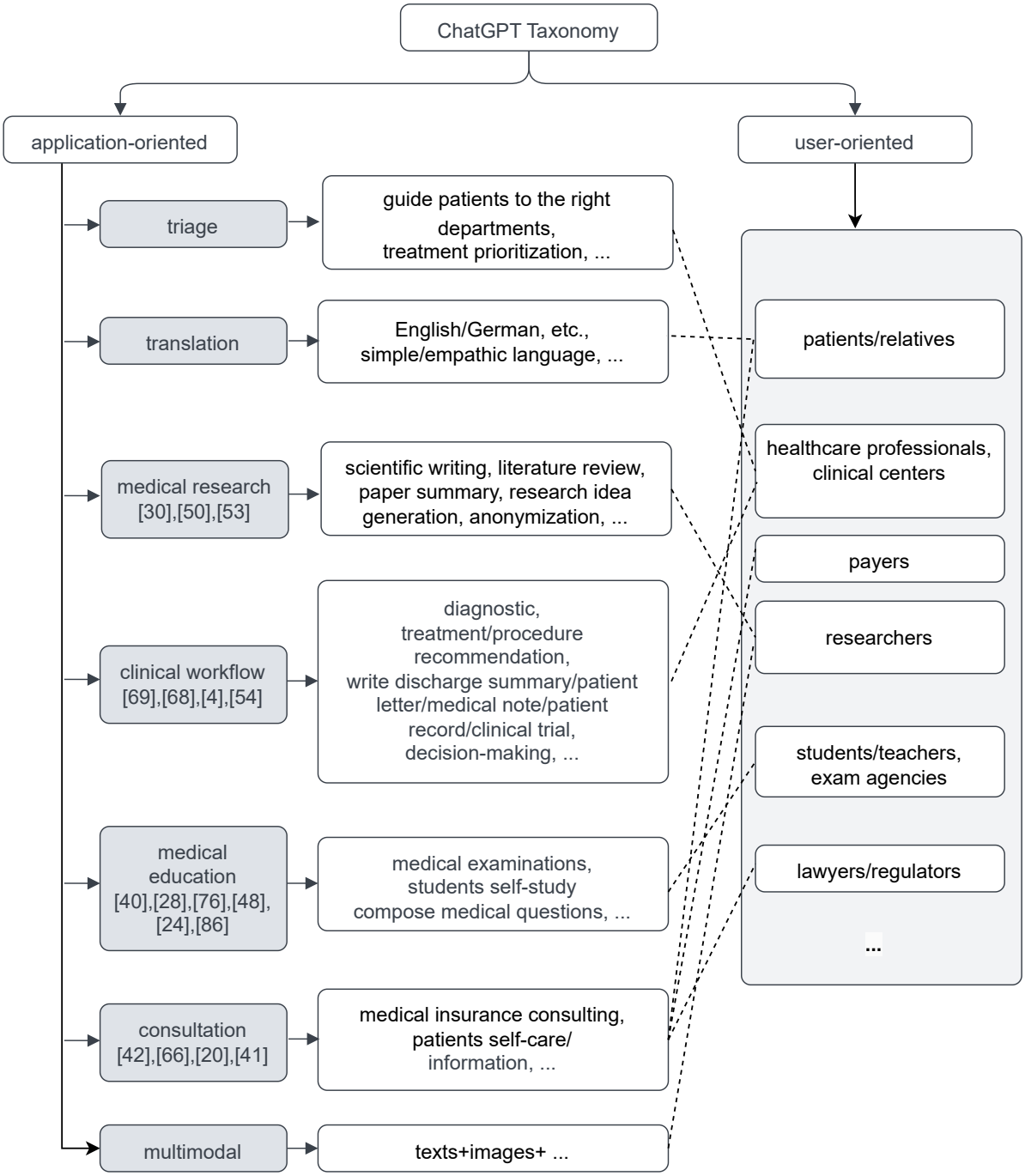

Figure 3: Application- and user-oriented Taxonomy used in the ChatGPT review. The references shown in the application boxes are the *Level 3* publications.

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 speciality and/or scenario;
- *Level 2*: Comments with one or more example use cases of ChatGPT in a specific medical speciality 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 speciality- 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. [91]. 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 speciality journals like nuclear medicine [3, 59], plastic surgery [79, 38], synthetic biology [106] and orthopaedic [81]. These publications usually deliver high-level comments about the potential impact and pitfalls of ChatGPT in healthcare [113], 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 [105, 16]. *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 [103, 27]. 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 [105, 27]. The decision is echoed by the academic community [58, 97, 115, 66], agreeing that

ChatGPT-generated content must be scrutinized by human experts before being used [58], as the generated content, such as references [105, 12, 40, 31] could be fabricated. Lee, J.Y. et al. [66] 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 [29, 7]. In short, even though the prospect is promising [92, 25, 43, 102], new regulations and substantial improvements are needed before ChatGPT can be safely and widely used for scientific writing, publishing, or medical research in general [105]. The *scenario* column in Table 1 corresponds to the taxonomy categorization. If the article concerns healthcare or a medical speciality 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 speciality 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 [36, 61], where the authors test ChatGPT in the US Medical Licensing Examination (USMLE). Even though the evaluations were carried out independently ([36] and [61] were published almost at the same time), similar results were reported, i.e., ChatGPT achieved only moderate passing performance. [36] 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). [36] 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 specialities. However, Mbakwe, A.B. et al. [74] 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 [112] and the AHA BLS / CLS Exams 2016 [32], on both of which ChatGPT failed to achieve passing scores.

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

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 [47, 53]. In [53], 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 [47]. Furthermore, [83] tested ChatGPT with 100 questions related to retina disease. The answers were evaluated based on a 5-point Likert scale by 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 [28]. Interestingly, based on the test results, [28] also reached the conclusion that ChatGPT fares better on 'memorization (fact-lookup)' type questions than on those requiring critical thinking, similar to [83]. 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 [67] and generating research questions [63] and systematic review topics [38]. In [67], 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 [29] have shown that scientists can not tell apart abstracts generated by ChatGPT from those written by humans, [67] 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 [63], the authors use ChatGPT to identify research questions in gastroenterology. The answers generated by ChatGPT proves to be highly relevant but lack depth and novelty. In [38], ChatGPT is used to generate systematic review topics in plastic surgery. Similar to [63], ChatGPT-generated research topics are generally not novel. The *version* column in Table 2 shows the version of ChatGPT used for evaluation. [63] 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

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

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

banned by many scientific journals, exploring new research ideas/topics with the assistance of ChatGPT faces less ethical issues. However, [63, 38] demonstrated that the current version of ChatGPT is not sufficiently qualified for such tasks. Humans still play dominating roles in ingenious and innovative research.

[87, 86, 4, 68] evaluate the application of ChatGPT in clinical workflow. In [87], 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 [47, 53]. 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 [86], 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 [4], 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 [68], 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.

Table 2: Summary of *Level 3* papers.

Ref.	Scenario	Summary	Results/Conclusion	Version	Journal
[87]	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
[86]	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
[51]	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
[36]	medical education	ChatGPT takes US Medical Licensing Examination (USMLE)	ChatGPT achieved passing score	Dec. 15, 2022	JMIR
[4]	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
[99]	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

[68]	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
[61]	medical education	ChatGPT takes USMLE (June 2022)	ChatGPT achieved passing score, and its explanations contain novel insights	-	PLOS Digital Health
[32]	medical education	ChatGPT takes life-support exams (AHA BLS / CLS Exams 2016)	ChatGPT did not reach passing score	Jan. 9 and 30, 2023	Resuscitation
[53]	consultation	ChatGPT provides cancer-related information and feedback on cancer misconceptions	ChatGPT provides highly accurate cancer information	Dec. 15, 2022	JNCI Cancer Spectrum
[67]	medical research	compared 50 ChatGPT-generated abstracts with real abstracts from scientific publications	Grammarly can detect ChatGPT-generated abstracts with high accuracy	-	AJOG
[83]	consultation	evaluate ChatGPT using 100 questions about retinal diseases	ChatGPT is highly accurate on general questions but less accurate for treatment options	-	Acta Ophthalmologica
[28]	consultation	compare ChatGPT with humans on 85 genetics/genomics questions	ChatGPT and humans perform similarly	-	preprint
[52]	consultation medical education	ChatGPT answers 284 question from various medical specialities	ChatGPT achieved overall high accuracies	-	preprint

[112]	medical education	ChatGPT takes Chinese National Medical Licensing Examination	ChatGPT's performance on the exam is well below passing level	-	preprint
[63]	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
[38]	medical research	ChatGPT generates systematic review topics in plastic surgeries	ChatGPT performs moderately in generating novel systematic review ideas	-	Aesthetic Surgery Journal
[98]	consultation medical education	evaluate ChatGPT using 100 OE questions about pathology	ChatGPT scored around 80%	Jan. 30, 2023	Cureus

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 (Figure 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 question samples 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. 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 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 specialities, 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.

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 continued to be evaluated in different specialities, 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 speciality, 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 and Conclusion

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. An automatic processing of multimodal input, like text and images, is an exciting development for future healthcare. In example, Contrastive Language-Image Pre-Training (CLIP) [85], a neural network trained on large-scale *image-text* pairs, possesses both vision and language capabilities, and is therefore a promising research direction towards AI-assisted multimodal healthcare. In general, a physician takes also several sources of information into account when making diagnosis and treatment decisions, such as the written reports and image acquisitions from a patient. ChatGPT-4, a enhanced version of ChatGPT released recently, is able to analyse and summarize images and texts, as seen from a live demo given by its developers.

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

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

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. [27], 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 [70], are always more desirable than ChatGPT.

As discussed in this review (Section 3), these evaluation studies on ChatGPT's performance in healthcare provide a general picture of the capability of the current release of ChatGPT. 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.

Acknowledgments

This work was supported by the REACT-EU project KITE (Plattform für KI-Translation Essen, EFRE-0801977, `https://kite.ikim.nrw/`) and the Cancer Research Center Cologne Essen (CCCE).

References

[1] Chiwon Ahn. "Exploring ChatGPT for information of cardiopulmonary resuscitation". In: *Resuscitation* 185 (2023).

[2] Haris M Akhter, Jeffrey S Cooper, and Jeffrey Cooper. "Acute Pulmonary Edema After Hyperbaric Oxygen Treatment: A Case Report Written With ChatGPT Assistance". In: *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] Stephen R Ali et al. "Using ChatGPT to write patient clinic letters". In: *The Lancet Digital Health* (2023).

[5] Hussam Alkaissi and Samy I McFarlane. "Artificial hallucinations in ChatGPT: implications in scientific writing". In: *Cureus* 15.2 (2023).

[6] Lauren B Anderson et al. "Generative AI as a Tool for Environmental Health Research Translation". In: *medRxiv* (2023), pp. 2023–02.

[7] Nash Anderson 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". In: *BMJ Open Sport & Exercise Medicine* 9.1 (2023), e001568.

[8] Taha Bin Arif, Uzair Munaf, and Ibtehaj Ul-Haque. "The future of medical education and research: Is ChatGPT a blessing or blight in disguise?" In: *Medical Education Online* 28.1 (2023), p. 2181052.

[9] Ömer Aydın and Enis Karaarslan. "OpenAI ChatGPT generated literature review: Digital twin in healthcare". In: *Emerging Computer Technologies 2* (2022), pp. 22–31.

[10] Adam Baker et al. "A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis". In: *Frontiers in artificial intelligence* 3 (2020), p. 543405.

[11] Christian Baumgartner. "The potential impact of ChatGPT in clinical and translational medicine". In: *Clinical and Translational Medicine* 13.3 (2023).

[12] Som Biswas. "ChatGPT and the Future of Medical Writing". In: *Radiology* (2023), p. 223312.

[13] Som S Biswas. "Potential Use of Chat GPT in Global Warming". In: *Annals of Biomedical Engineering* (2023), pp. 1–2.

[14] Som S Biswas. "Role of Chat GPT in Public Health". In: *Annals of Biomedical Engineering* (2023), pp. 1–2.

[15] Christian M Boßelmann, Costin Leu, and Dennis Lal. "Are AI language models such as ChatGPT ready to improve the care of individuals with epilepsy?" In: *Epilepsia* (2023).

[16] Jeffrey Brainard. "Journals take up arms against AI-written text". In: *Science (New York, NY)* 379.6634 (2023), pp. 740–741.

[17] Thorsten Brants et al. "Large language models in machine translation". In: (2007).

[18] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[19] Patrick Cahan and Barbara Treutlein. "A conversation with ChatGPT on the role of computational systems biology in stem cell research". In: *Stem Cell Reports* 18.1 (2023), pp. 1–2.

[20] Marco Cascella et al. "Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios". In: *Journal of Medical Systems* 47.1 (2023), pp. 1–5.

[21] Joyjit Chatterjee and Nina Dethlefs. "This new conversational AI model can be your friend, philosopher, and guide... and even your worst enemy". In: *Patterns* 4.1 (2023), p. 100676.

[22] KR Chowdhary. "Natural language processing". In: *Fundamentals of artificial intelligence* (2020), pp. 603–649.

[23] Kenneth Ward Church. "Word2Vec". In: *Natural Language Engineering* 23.1 (2017), pp. 155–162.

[24] Kevin Clark et al. "Electra: Pre-training text encoders as discriminators rather than generators". In: *arXiv preprint arXiv:2003.10555* (2020).

[25] 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.

[26] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[27] Eva AM van Dis et al. "ChatGPT: five priorities for research". In: *Nature* 614.7947 (2023), pp. 224–226.

[28] Dat Duong and Benjamin D Solomon. "Analysis of large-language model versus human performance for genetics questions". In: *medRxiv* (2023), pp. 2023–01.

[29] Holly Else. "Abstracts written by ChatGPT fool scientists". In: *Nature* 613.7944 (2023), pp. 423–423.

[30] Fábio Caleça Emidio et al. "Rectal Bezoar: A Rare Cause of Intestinal Obstruction". In: *Cureus* 15.3 (2023).

[31] 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". In: *JMIR Medical Education* 9.1 (2023), e46885.

[32] Nino Fijačko et al. "Can ChatGPT pass the life support exams without entering the American heart association course?" In: *Resuscitation* 185 (2023).

[33] Caitlin R Francis et al. "Arf6 Regulates Endocytosis and Angiogenesis by Promoting Filamentous Actin Assembly". In: *bioRxiv* (2023), pp. 2023–02.

[34] Carol Friedman, George Hripcsak, et al. "Natural language processing and its future in medicine". In: *Acad Med* 74.8 (1999), pp. 890–5.

[35] Andrew T Gabrielson, Anobel Y Odisho, and David Canes. "Harnessing Generative AI to Improve Efficiency Among Urologists: Welcome ChatGPT". In: *The Journal of Urology* (2023), pp. 10–1097.

[36] 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". In: *JMIR Medical Education* 9.1 (2023), e45312.

[37] Rachel S Goodman et al. "On the cusp: Considering the impact of artificial intelligence language models in healthcare". In: *Med* 4.3 (2023), pp. 139–140.

[38] Rohun Gupta et al. "Application of ChatGPT in Cosmetic Plastic Surgery: Ally or Antagonist". In: *Aesthetic Surgery Journal* (2023), p. 042.

[39] John E Hallsworth et al. "Scientific novelty beyond the experiment". In: *Microbial Biotechnology* (2023).

[40] Michael Haman and Milan Školnık. "Using ChatGPT to conduct a literature review". In: *Accountability in Research* (2023), pp. 1–3.

[41] Pengcheng He, Jianfeng Gao, and Weizhu Chen. "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: `https://openreview.net/forum?id=sE7-XhLxHA`.

[42] Pengcheng He et al. "DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION". In: *International Conference on Learning Representations*. 2021. URL: `https://openreview.net/forum?id=XPZIaotutsD`.

[43] Elisa L Hill-Yardin et al. "A Chat (GPT) about the future of scientific publishing". In: *Brain, behavior, and immunity* (2023), S0889–1591.

[44] 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". In: *International Journal of Environmental Research and Public Health* 20.4 (2023), p. 3378.

[45] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[46] Andreas Holzinger et al. "AI for life: Trends in artificial intelligence for biotechnology". In: *New Biotechnology* 74 (2023), pp. 16–24.

[47] Ashley M Hopkins et al. "Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift". In: *JNCI Cancer Spectrum* 7.2 (2023), pkad010.

[48] Mohammad Hosseini and Serge PJM Horbach. "Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other Large Language Models in scholarly peer review". In: (2023).

[49] Alex Howard, William Hope, and Alessandro Gerada. "ChatGPT and antimicrobial advice: the end of the consulting infection doctor?" In: *The Lancet Infectious Diseases* (2023).

[50] Guangwei Hu. "Challenges for Enforcing Editorial Policies on AI-generated Papers". In: *Accountability in Research* (2023).

[51] 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". In: *Journal of Educational Evaluation for Health Professions* 20 (2023), p. 1.

[52] Douglas Johnson et al. "Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model". In: (2023).

[53] Skyler B Johnson et al. "Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information". In: *JNCI Cancer Spectrum* 7.2 (2023), pkad015.

[54] Mandar Joshi et al. "Spanbert: Improving pre-training by representing and predicting spans". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 64–77.

[55] David Jungwirth and Daniela Haluza. "Artificial Intelligence and Public Health: An Exploratory Study". In: *International Journal of Environmental Research and Public Health* 20.5 (2023), p. 4541.

[56] Rohan Karkra et al. "Recurrent Strokes in a Patient With Metastatic Lung Cancer". In: *Cureus* 15.2 (2023).

[57] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. "Reformer: The Efficient Transformer". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=rkgNKkHtvB.

[58] Felipe C Kitamura. "ChatGPT is shaping the future of medical writing but still requires human judgment". In: *Radiology* (2023), p. 230171.

[59] Jens Kleesiek et al. "An opinion on ChatGPT in Healthcare - written by humans only". In: *Journal of Nuclear Medicine* (2023). DOI: 10.2967/jnumed.123.265687.

[60] Malcolm Koo. "The Importance of Proper Use of ChatGPT in Medical Writing". In: *Radiology* (2023), p. 230312.

[61] Tiffany H Kung et al. "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models". In: *PLOS Digital Health* 2.2 (2023), e0000198.

[62] Adi Lahat and Eyal Klang. "Can advanced technologies help address the global increase in demand for specialized medical care and improve telehealth services?" In: *Journal of Telemedicine and Telecare* (2023), pp. 1357633X231155520–1357633X231155520.

[63] Adi Lahat et al. "Evaluating the use of large language model in identifying top research questions in gastroenterology". In: *Scientific Reports* 13.1 (2023), p. 4164.

[64] Zhenzhong Lan et al. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

[65] Augustin Lecler, Loıc Duron, and Philippe Soyer. "Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT". In: *Diagnostic and Interventional Imaging* (2023).

[66] Ju Yoen Lee. "Can an artificial intelligence chatbot be the author of a scholarly article?" In: *science editing* 10.1 (2023), pp. 7–12.

[67] Gabriel Levin et al. "Identifying ChatGPT-written OBGYN abstracts using a simple tool". In: *American Journal of Obstetrics & Gynecology MFM* (2023).

[68] Siru Liu et al. "Assessing the Value of ChatGPT for Clinical Decision Support Optimization". In: *medRxiv* (2023), pp. 2023–02.

[69] Saskia Locke et al. "Natural language processing in medicine: a review". In: *Trends in Anaesthesia and Critical Care* 38 (2021), pp. 4–9.

[70] Renqian Luo et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining". In: *Briefings in Bioinformatics* 23.6 (2022).

[71] Calum Macdonald et al. "Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis". In: *Journal of Global Health* 13 (2023).

[72] 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).

[73] Ken Masters. "Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158". In: *Medical Teacher* (2023), pp. 1–11.

[74] Amarachi B Mbakwe et al. "ChatGPT passing USMLE shines a spotlight on the flaws of medical education". In: *PLOS Digital Health* 2.2 (2023), e0000205.

[75] Larry R Medsker and LC Jain. "Nuclear Medicine from a Novel Perspective: Buvat and Weber Talk with OpenAI's ChatGPT". In: *Journal of Nuclear Medicine* (2023).

[76] Larry R Medsker and LC Jain. "Recurrent neural networks". In: *Design and Applications* 5 (2001), pp. 64–67.

[77] Stéphane Meystre and Peter J Haug. "Natural language processing to extract medical problems from electronic clinical documents: performance evaluation". In: *Journal of biomedical informatics* 39.6 (2006), pp. 589–599.

[78] Sreenivasulu Reddy Mogali. "Initial impressions of ChatGPT for anatomy education". In: *Anatomical Sciences Education* (2023).

[79] 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.

[80] Siobhan O'Connor et al. "Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?" In: *Nurse Education in Practice* 66 (2022), pp. 103537–103537.

[81] 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.

[82] Sajan B Patel and Kyle Lam. "ChatGPT: the future of discharge summaries?" In: *The Lancet Digital Health* 5.3 (2023), e107–e108.

[83] Ivan Potapenko et al. "Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT." In: *Acta Ophthalmologica* (2023).

[84] Paco Prada, Nader Perroud, and Gabriel Thorens. "Artificial intelligence and psychiatry: questions from psychiatrists to ChatGPT". In: *Revue Medicale Suisse* 19.818 (2023), pp. 532–536.

[85] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[86] Arya S Rao et al. "Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow". In: *medRxiv* (2023), pp. 2023–02.

[87] Arya S Rao et al. "Evaluating ChatGPT as an adjunct for radiologic decision-making". In: *medRxiv* (2023), pp. 2023–02.

[88] Matthias C Rillig et al. "Risks and Benefits of Large Language Models for the Environment". In: *Environmental Science & Technology* (2023).

[89] Mary Sabry Abdel-Messih and Maged N Kamel Boulos. "ChatGPT in Clinical Toxicology". In: *JMIR Medical Education* 9 (2023), e46876.

[90] Abdullah Saeed et al. "Pacemaker Malfunction in a Patient With Congestive Heart Failure and Hypertension". In: *Cureus Journal of Medical Science* 15.2 (2023).

[91] Malik Sallam. "ChatGPT Utility in Health Care Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns". In: *Healthcare*. Vol. 11. 6. MDPI. 2023, p. 887.

[92] Michele Salvagno, Fabio Silvio Taccone, Alberto Giovanni Gerli, et al. "Can artificial intelligence help for scientific writing?" In: *Critical Care* 27.1 (2023), pp. 1–5.

[93] Anthony Scerri and Karen H Morin. "Using chatbots like ChatGPT to support nursing practice". In: *Journal of Clinical Nursing* (2023).

[94] William B. Schwartz. "Medicine and the Computer". In: *New England Journal of Medicine* 283.23 (1970). PMID: 4920342, pp. 1257–1264. DOI: 10.1056/NEJM197012032832305.

[95] Javier Selva et al. "Video transformers: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[96] Adrien Sieg. "FROM Pre-trained Word Embeddings TO Pre-trained Language Models — Focus on BERT". In: *Towards Data Science* (2019).

[97] Bob Siegerink et al. "ChatGPT as an author of academic papers is wrong and highlights the concepts of accountability and contributorship." In: *Nurse Education in Practice* 68 (2023), pp. 103599–103599.

[98] Ranwir K Sinha et al. "Applicability of ChatGPT in assisting to solve higher order problems in pathology". In: *Cureus* 15.2 (2023).

[99] Jan Šlapeta. "Are ChatGPT and other pretrained language models good parasitologists?" In: *Trends in Parasitology* (2023).

[100] Gerald Gui Ren Sng et al. "Potential and Pitfalls of ChatGPT and Natural-Language Artificial Intelligence Models for Diabetes Education." In: *Diabetes Care* (2023), pp. dc230197–dc230197.

[101] Peter Spyns. "Natural language processing in medicine: an overview". In: *Methods of information in medicine* 35.04/05 (1996), pp. 285–301.

[102] Chris Stokel-Walker. "AI bot ChatGPT writes smart essays-should academics worry?" In: *Nature* (2022).

[103] Chris Stokel-Walker. "ChatGPT listed as author on research papers: many scientists disapprove". In: *Nature* ().

[104] Martin Strunga et al. "Artificial Intelligence Systems Assisting in the Assessment of the Course and Retention of Orthodontic Treatment". In: *Healthcare*. Vol. 11. 5. MDPI. 2023, p. 683.

[105] Holden H Thorp. "ChatGPT is fun, but not an author". In: *Science* 379.6630 (2023), pp. 313–313.

[106] Yaojun Tong and Lixin Zhang. "Discovering the next decade's synthetic biology research trends with ChatGPT". In: *Synthetic and Systems Biotechnology* 8.2 (2023), p. 220.

[107] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[108] Raju Vaishya, Anoop Misra, and Abhishek Vaish. "ChatGPT: Is this version good for healthcare and research?" In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* (2023), p. 102744.

[109] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[110] 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*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://aclanthology.org/W18-5446.

[111] Jing Wang et al. "Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on PubMed". In: *Journal of medical Internet research* 22.1 (2020), e16816.

[112] Xinyi Wang et al. "ChatGPT Performs on the Chinese National Medical Licensing Examination". In: *medRxiv* (2023).

[113] "Will ChatGPT transform healthcare?" In: *Nat Med* 29 (2023), pp. 505–506. DOI: 10.1038/s41591-023-02289-5.

[114] Zhilin Yang et al. "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* 32 (2019).

[115] Nicole Shu Ling Yeo-Teh and 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". In: *Accountability in research* just-accepted (2023).