Southeast Europe Journal of Soft Computing ISSN 2233 – 1859
Available online: http://scjournal.ius.edu.ba

Vol 12, No. 1, March 2023, pp. 13-41

Exploring Medical Breakthroughs: A Systematic Review of ChatGPT Applications in Healthcare

[1*]Fatima Muftić, [1]Merjem Kadunić,[1]Almina Mušinbegović and [1]Ali Abd Almisreb

[1]Faculty of Engineering and Natural Sciences, International University of Sarajevo,
Hrasnicka Cesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
*Corresponding Author: fmuftic@student.ius.edu.ba

Article Info

Article history:

Article received on 14 April 2023

Received in revised form 17 April 2023

Keywords:

ChatGPT; Artificial intelligence;
Natural language processing;
Systematic review; Medicine

ABSTRACT: ChatGPT, a large language model developed by OpenAI, has emerged as a powerful tool in the field of medicine. In this systematic review, we explore the potential of ChatGPT in various medical applications by analyzing articles related to medicine and healthcare. We carefully examined the methodologies, results, and conclusions of these articles to provide a comprehensive overview of the current evidence on the use of ChatGPT in the field of medicine. Through this review, we highlight how ChatGPT has been utilized to streamline and simplify complex tasks, improve patient care, enhance clinical decision-making, and facilitate communication among healthcare professionals. We also discuss the challenges and limitations of using ChatGPT in medicine, including concerns related to privacy, ethical considerations, and potential biases. Despite these challenges, ChatGPT has shown great promise in transforming the landscape of medicine and has the potential to revolutionize healthcare delivery. By synthesizing the findings from these articles, we aim to provide a critical and evidence-based evaluation of the current state of ChatGPT in medicine, and to identify areas for further research and development.

1. INTRODUCTION

The field of medicine is constantly evolving, with rapid advancements in technology driving innovations in patient care, clinical decision-making, and healthcare administration. Artificial intelligence systems promise many potential applications in medicine, such as differential diagnosis generation and selection, clinical decision support, and analysis of imaging-, physiologic-, and genomic-based data [1] AI has the potential to significantly impact the diagnosis of diseases by improving the accuracy, speed, and efficiency of decision-making. AI algorithms can process vast amounts of data, identify patterns, and make predictions that may be beyond the capabilities of human physicians [2].

One such innovation that has gained significant attention in recent years is the use of ChatGPT, a large language model developed by OpenAI. The OpenAI's chatbot gained more than 1 million users in the first few days after its launch and 100 million in the first 2 months, positioning itself as the fastest-growing consumer application in history [3]. ChatGPT is a powerful tool that utilizes natural language processing (NLP) techniques to generate human-like responses in real-time conversations. It has the ability to understand and respond to text-based inputs, making it an ideal solution for a wide range of medical tasks [4] ChatGPT could also provide tutoring and homework help by answering questions and providing explanations to help students understand complex concepts. Medical students must be able to evaluate the accuracy of medical information generated by AI

and to have the ability to create reliable, validated information for patients and the public. Therefore, it is necessary to determine how accurately ChatGPT can solve questions on medical examinations [5].

The goal of this review is to provide an overview of the growing body of literature that explores the potential of ChatGPT in medicine. We conducted a comprehensive search of articles related to medicine and healthcare, and analyzed the findings to highlight the ways in which ChatGPT has been utilized to simplify complex tasks, improve patient care [1], [6], [7], [8], [9] enhance clinical decision-making [10], [11], [12], and facilitate communication among healthcare professionals [13]. In this systematic review, we have specifically focused on articles that provide empirical results related to the utilization of ChatGPT in medicine. By thoroughly examining and synthesizing the findings of these studies, we aim to critically evaluate the current state of ChatGPT's performance in various medical tasks. Through rigorous analysis and discussion of the results, we will draw evidence-based conclusions regarding the efficacy and potential of ChatGPT in the field of medicine.

In the next chapter, we will provide a background on NLP and LLMs to better understand the technical aspects of ChatGPT and how it differs from other NLP models. By providing this background, researchers can evaluate the potential applications and limitations of ChatGPT in medicine, as well as identify gaps in the current literature that can inform future research directions. Chapter 3 will detail the methodology used to identify, screen, and select relevant studies for inclusion in the review. Chapter 4 will present the results of the review, synthesizing findings across studies to identify patterns, trends, and gaps in the literature. Chapter 5 will offer a critical discussion of the results, highlighting their implications for future research, practice, and policy. Finally, Chapter 6 will provide a conclusion, summarizing the key findings of the review and offering recommendations for future work in this area.

2. BACKGROUND

ChatGPT is an AI-powered chatbot developed by the artificial intelligence (AI) research company OpenAI and launched in November 2022. The chatbot uses a field of machine learning known as natural language processing (NLP) to generate responses to users' questions and prompts. In a gist, ChatGPT works in a conversational interface with its user, responds to follow-up questions, admits and corrects mistakes, rejects improper asks, and even challenges incorrect premises. GPT stands for "generative pre-trained transformers", which are capable of understanding and producing strings of complex thoughts and ideas. When a command is entered, ChatGPT pulls data from everywhere it can get its hands on, feeds it into a transformer model, then maps the relationships between different pieces of information and guesses what text belongs together in certain contexts [10].

2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) 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 [4].

2.2 Transformer

Traditional language models are programmed to use statistical techniques to predict the next word in a sentence, while ChatGPT uses transformer-based models that allow for the processing of vast amounts of data in parallel. The result is a revolution in the ability of these models to understand and generate text [14] Since introducing the concept of attention in deep learning models, the transformer is an established architecture with dominance in nearly all NLP benchmarks, including question answering, translation, and text classification [6]. Transformers have also been used for tasks beyond NLP, such as image and video processing, and they are an active area of research in the deep learning community [4].

2.3 Large Language Models (LLMs)

While the base-model architecture remained relatively unchanged throughout the years, significant progress was made by scaling the number of layers and internal dimensions resulting in so-called Large Language Models (LLMs) with billions of parameters, which

lead to increased model capacity and abilities [6] LLM represents artificial intelligence (AI) tools based on multi-layer recurrent neural networks that are trained on vast amounts of data to generate human-like text [14].

3. METHODOLOGY

It's important to adhere to the specific guidelines and reporting standards for systematic reviews, such as those outlined by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement, to ensure the highest quality and transparency in reporting. The search strategy used in the following systematic review is illustrated in Figure 1. Since most 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 systemic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRIMSA) guidelines. We use Google Scholar and PubMed as the only sources to search candidate publications. When searching on Google Scholar, filters were used that enable searching for articles ordered by relevance and published after 2022. The keywords we entered during the search were “chatgpt applications in medicine” which showed 2,000 results. PubMed showed 49 results, but after screening only 7 of them were selected. During the review of Google Scholar search results, we screened the first 13 pages, each containing 10 articles, resulting in a total of 130 articles screened out of 2000 results.

It is important to mention that every day there are more and more articles related to the use of ChatGPT in medicine and that these searches were made on April 15, 2023. The number of searches after this date will certainly increase, but we focus on the number that was recorded at the time of conducting our review.

The eligibility criteria involved any type of published scientific research or preprints (article, review, communication, editorial, opinion, etc.) addressing ChatGPT that fell under the medical field categories (surgical, healthcare, educational, etc.).

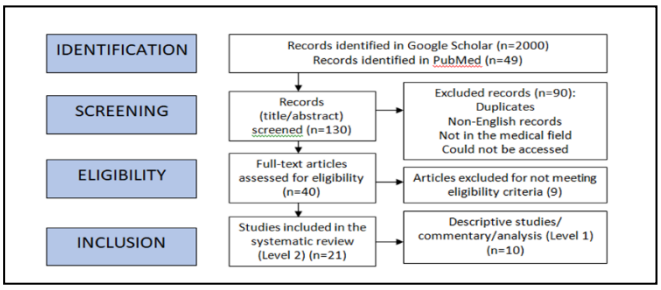

Figure 1: Flowchart of the record selection process based on PRIMSA guidelines.

The exclusion criteria included: non-English records, records addressing ChatGPT in subjects other than those related to medical fields, and articles from non-academic sources (e.g., newspapers, internet websites, magazines, etc.). We excluded 90 out of 130 screened articles, leaving us with 40 full-text articles assessed for eligibility.

After completing the eligibility assessment during the selection process (n=31), the articles obtained were categorized into two levels. The first level represents descriptive studies presented in the form of comments or analysis of the performance of ChatGPT in various fields of medicine. The second level represents performance evaluation or assessment studies of ChatGPT, as well as cross-selection observational studies. (As shown in Table 1.)

It's important to note that while Level 1 types of articles may provide valuable information or insights, they may not meet the rigorous standards of evidence synthesis and analysis typically required in a systematic review. We use Level 1 articles for a better understanding of the advantages and disadvantages of using ChatGPT in medicine [15] and its use in specific fields of medicine such as orthopedics and sports medicine [16], surgery [10], healthcare [3], patient care [13], clinical and translational medicine [17], diagnose precision [18], medical academic writing [19], radiology [11] and medical education [20].

In the following, we will use Level 2 articles that provide concrete values obtained in experimental studies, and there will be a discussion and an overview of the obtained results. Including only studies with results in a systematic review is essential because it ensures that the synthesis of evidence is based on empirical data rather than subjective commentary or analysis, thus enhancing the reliability and validity of the findings.

Table 1 Dividing articles into Level 1 and Level 2

Ref.	Category	Main Content	Tag
[1]	Observational study	ChatGPT and GPT-4 used to generate answers on a 500-question mock neurosurgical written boards examination	Level 2
[2]	Cross-sectional observational study	Using ChatGPT to respond to 100 randomly selected higher-order reasoning queries related to different systems in pathology categorized into 11 systems	Level 2
[3]	Commentary/analysis	The use of ChatGPTas a chatbot for communication, content generation, and other tasks in healthcare and scientific settings.	Level 1
[5]	Performance evaluation	Comparing the knowledge and interpretation ability of ChatGPT with those of medical students in Korea by administering a parasitology examination	Level 2
[6]	Cross-selection observational study	Assessing the simplification of radiology reports using ChatGPT	Level 2
[7]	Cross-sectional observational study	ChatGPT(GPT 3.5) used to generate answers to medical queries created by physicians	Level 2
[8]	Observational Study	Use of ChatGPT-3 to generate differential-diagnosis lists for clinical vignettes	Level 2
[9]	Performance evaluation	Use of ChatGPT to answer 164 questions related to cirrhosis and HCC	Level 2
[10]	Commentary/analysis	Use of ChatGPT in various applications related to surgery (clinical decision-making, appointment scheduling, medical record transcription, surgical planning)	Level 1
[11]	Descriptive study	Can AI take the position of a healthcare expert or if it can be utilized as a tool to enhance decision-making that is dependable and simple as ChatGPT interprets radiological images	Level 1
[12]	Cross-sectional observational study	376 questions from the USMLE-2022 sample exam were made available to the general public. The effectiveness of GPT3 model was assessed, where all inputs represented genuine out-of-training samples	Level 2
[13]	Commentary/analysis	Use of ChatGPT in facilitating communication between healthcare providers and patients, problem and sentiment analysis in the messages from the patient, clinical use and integration of AI into dental education	Level 1
[14]	Performance evaluation	Use of ChatGPT for answering multiple-choice questions related to nuclear medicine treatments and investigations	Level 2
[15]	Generalized review	Examples of conversation transcripts between Chat GPT-4 to demonstrate its behavior, capabilities and limitations	Level 1
[16]	Commentary/analysis	The use of ChatGPT for tasks such as literature review, data analysis, and hypothesis generation in the field of orthopedic and sports medicine	Level 1
[17]	Commentary/analysis	Use of ChatGPTs potential to revolutionize the way medical information is disseminated and processed in fields of medicine	Level 1
[18]	Descriptive study	Working with doctors to develop ChatGPT in the formation of a "Medical Dream Team," where the combined knowledge can result in more precise diagnoses, individualized treatment regimens, and better patient care	Level 1
[19]	Descriptive study	ChatGPT's potential advantages and disadvantages when used in academic writing, as a tool to enhance academic writing, its speedy text generation, grammar and style suggestions, and help with content organization	Level 1

[20]	Descriptive study	Impact of ChatGPT on medical education, scientific research, medical writing, ethical issues, diagnostic decision-making, automation possibilities, and criticisms	Level 1
[21]	Performance evaluation	Applying LLMs to solve Japanese medical licensing exam	Level 2
[22]	Observational study	Evaluating GPT-4 against medical United States Medical Licensing Examination (USMLE), & MultiMedQA suite of benchmark datasets Comparison: Performance of GPT-4 model	Level 2
[23]	Observational Study	Utilizing ChatGPT to provide the most probable diagnosis and top five most probable diagnoses for each case	Level 2
[24]	Performance evaluation	Use of ChatGPT, GPT-3 and InstructGPT to answer questions from medical knowledge data sets (AMBOSS-Step1, AMBOSS-Step2, NBME-Free-Step1, and NBME-Free-Step2)	Level 2
[25]	Descriptive evaluation study	Use of ChatGPT to answer questions related to vaccination and COVID-19 related topics	Level 2
[26]	Multilingual feasibility study	Use of GPT-4 to automate the conversion of free-text radiology reports into structured template	Level 2
[27]	Performance evaluation	Use of ChatGPT for answering ophthalmology-related questions	Level 2
[28]	Cross-sectional observational study	Using ChatGPT to converse with and obtain responses to first-order and second-order knowledge questions related to microbiology	Level 2
[29]	Cross-sectional observational study	ChatGPT answers questions from National testing of medical students in Italy	Level 2
[30]	Experimental study	A group of skilled gastroenterologists with knowledge of the related subject areas evaluating the use of ChatGPT in identifying top research questions	Level 2
[31]	Cross-sectional observational study	Evaluating ChatGPT's ability to solve higher-order questions with more complex medical biochemistry issues	Level 2

4. RESULTS

The initial search in Google Scholar and PubMed yielded a total of 2,000 results (Google Scholar) and 49 (PubMed). After applying the inclusion and exclusion criteria, a total of 31 eligible articles were included in the final review. Out of the 31 articles, 10 were categorized as Level 1 and 21 as Level 2. These articles cover various medical fields, such as pathology, ophthalmology, surgery, nuclear medicine, microbiology and more.

In this study, we examined each Level 2 article and performed a comprehensive assessment of its content. Our evaluation included generating concise summaries, conducting PICOS analysis, and determining the Outcome Level Assessment. The findings of these analyses have been systematically organized and presented in Table 2 for ease of reference.

PICOS, an acronym for Population, Intervention, Comparator, Outcomes, and Study design, is a widely utilized framework in evidence-based medicine that facilitates the process of formulating research questions, conducting literature reviews, and synthesizing the findings of relevant studies [32]. The PICOS approach enables researchers to delineate the key aspects of a study, which, in turn, allows for a more streamlined assessment and comparison of the gathered evidence [33].

For each Level 2 article included in our research, the following details were extracted:

1. Summary: A brief overview highlighting the main objectives, methodology, and findings of the article.
2. PICOS analysis: A thorough evaluation encompassing the population studied, the intervention implemented, the comparator or control group, the outcomes measured, and the study design employed in the research.
3. Outcome Level Assessment: Primary and Secondary outcomes of the study.
4. Risk of biases: Potential systematic errors in the design, conduct, or analysis of a study that may lead to misleading or inaccurate results.

The cumulative results from these analyses, as represented in Table 2, offer valuable insights into the various studies and contribute significantly to our understanding of the subject matter.

Table 2 Analysis of Level 2 articles

STUDY REFERENCE [2]

SUMMARY	This study used ChatGPT, a natural language processing model, to assess its ability in higher-order reasoning in the field of pathology. A total of 100 higher-order reasoning questions were randomly selected from a question bank and categorized by organ systems. The results showed that ChatGPT was able to solve the questions with a relational level of accuracy, and the majority of responses fell into the "relational" category in the structure of the observed learning outcome (SOLO) taxonomy. There was no significant difference in scores among questions from different organ systems. The inter-rater reliability among three expert pathologists was excellent. The study concluded that ChatGPT can be a helpful tool for academicians or students in solving reasoning-type questions in pathology, although further studies are needed to determine its accuracy level in future versions.
PICOS	**Population:** The utilization of ChatGPT for higher-order reasoning in the subject of pathology. **Intervention**: Using ChatGPT to respond to 100 randomly selected higher-order reasoning queries related to different systems in pathology categorized into 11 systems (e.g., general pathology, cardiovascular pathology, gastrointestinal pathology) **Comparison:** Score was compared by a one-sample median test with hypothetical values to find its accuracy. **Outcomes:** Accuracy of ChatGPT in providing responses which was evaluated using a scoring system based on a pre-defined answer key and the Structure of the Observed Learning Outcome (SOLO) taxonomy. **Study design:** Cross-sectional study using ChatGPT and evaluating the responses using scoring methods.
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** ChatGPT's median accuracy score for higher-order pathology questions was 4.08 (Q1-Q3: 4-4.33), significantly lower than 5 but close to 4. Average response time: 45.31±7.14 seconds. Hepatobiliary and nervous system pathology scores were closest to 5. SOLO taxonomy: 86 relational, 12 multistructural, 2 prestructural ($p<0.0001$). **Secondary outcomes:** Average response time, SOLO taxonomy distribution, inter-rater reliability (ICC=0.975 [95% CI: 0.965-0.983], F=40.26, $p<0.0001$), and performance comparison across organ systems (Kruskal Wallis test $p=0.55$).
RISK OF BIAS	Although the answer keys were prepared beforehand, a subjective evaluation bias still may present.

STUDY REFERENCE [5]

SUMMARY	The article most likely includes descriptive research that assesses ChatGPT's performance in a particular situation, such as when taking a parasitology exam. In order to assess whether ChatGPT's performance is comparable to that of human medical students, the study may compare the accuracy and proficiency of ChatGPT's responses to those of Korean medical students. The study's conclusions may shed light on ChatGPT's potential as a medium for education in the field of parasitology as well as on its superior capacity for interpretation and accurate response-giving compared to human medical students. However, it is impossible to offer a thorough explanation of the study's findings, methods, or conclusions without having access to the complete publication. Referring back to the source article, it is advised for thorough and accurate information.
PICOS	**Population:** 77 medical students, ChatGPT **Intervention:** Comparing the knowledge and interpretation ability of ChatGPT with those of medical students in Korea by administering a parasitology examination. **Comparison**: Firstly**,** the knowledge level of ChatGBT and students' performance were compared, then these results were compared with previous studies. **Outcomes:** The performance of ChatGPT was lower than that of medical students. The knowledge level of the items was not related to the right answer rate displayed by ChatGPT. However, there was a correlation between possible explanations and accurate solutions. **Study design:** Performance evaluation
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** The fact that 60.8% of the questions on ChatGPT had the correct answers did not necessarily mean that the students were performing poorly. The fact that medical students took the test four days following the session and had prior knowledge of parasitology may have contributed to their significantly higher average score (89.6%). **Secondary outcomes**: The input for the ChatGPT question items was not exactly the same as what was used for the medical students. The author has to re-describe this information because graphs, figures, and tables cannot be received by the conversation. Although the author has worked in the field of parasitology for 40 years (1982-2022) in Korea, the interpretation of the explanations and right answers may change depending on the viewpoints of other parasitologists. The area and the medical setting may also influence the best patient care techniques and circumstances.
RISK OF BIAS	The main bias in the study is that ChatGPT was unable to interpret figures, graphs, and tables as a student can, so the author had to describe these materials in text form. Additionally, the interpretation of explanations and correct answers may vary according to the perspectives of different parasitologists.

STUDY REFERENCE [6]

SUMMARY	The release of ChatGPT has gained widespread attention beyond the research community. Users are expected to apply ChatGPT to various tasks, including simplifying their own medical reports. This can empower patients, promote patient-centered care, and enhance patient satisfaction. To investigate this phenomenon, a case study was conducted where 15 radiologists were asked to assess the quality of radiology reports simplified by ChatGPT. While most radiologists agreed that the simplified reports were factually correct, complete, and not potentially harmful to patients, instances of incorrect statements, missed key medical findings, and potentially harmful passages were reported. This study highlights both the opportunities and challenges of using ChatGPT-like models for simplifying radiology reports. Opportunities identified in using ChatGPT for simplifying radiology reports include improved accessibility of personal medical information for patients, enabling them to better comprehend their health situation and prepare for future doctor-patient interactions. Challenges identified include the potential risks of errors and harmful conclusions in the simplified reports, as identified by radiologists. So, despite the potential of using large language models like ChatGPT to improve patient-centered care in radiology and other medical domains, further studies are needed.
PICOS	**Population:** Radiologists with varying levels of experience from clinic (Department of Radiology, University Hospital, LMU Munich) **Intervention:** Assessing the simplification of radiology reports using ChatGPT **Comparison:** N/A **Outcomes:** Radiologists' opinion on the quality of the simplified radiology reports through different ratings **Study design:** Cross-selection observational study
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** 51% of the participants highlighted incorrect passages, while only 22% and 36% of the radiologists listed missing relevant information and potentially harmful conclusions, respectively. Factual Correctness: Radiologists' ratings of the factual correctness of the simplified radiology reports generated with ChatGPT, measured on a scale (median = 2) based on their agreement with the statements made in the reports. Error categories identified include misinterpretation of medical terms, imprecise language, hallucination, odd language, and grammatical errors. Completeness: Radiologists' verification of the completeness of the simplified reports generated with ChatGPT, assessed based on their agreement (median = 2) with the reports containing the relevant medical information for the patients. Categories of missing key medical information identified include missed findings and unspecific location, suggesting the potential loss of medical context and preciseness in the simplification process. **Secondary outcomes:** Potential Harm: Radiologists' perception of the potential harm caused by the simplified reports generated with ChatGPT, measured on a scale (median = 4) based on their agreement with the statement that the reports might lead patients to draw wrong conclusions resulting in physical and/or psychological harm.
RISK OF BIAS	ChatGPT might have intrinsic biases due to imbalanced training data.

STUDY REFERENCE	[7]
SUMMARY	This study was conducted to address the accuracy and completeness of medical information generated by ChatGPT based on 284 medical questions prepared by 33 physicians from 17 different specialties. The physicians rated the accuracy of answers to the six-point Linkert scale and completeness to the three-point Linkert scale. The results revealed a median accuracy score of 5.5 (mean score of 4.8) and a median completeness score of 3 (mean score of 2.5). This study has shown that ChatGPT has a promise for providing accurate and comprehensive medical information, however, they are not completely reliable. The scope of conclusions of this study is limited due to sample size, number of questions, and cohort of physicians.
PICOS	**Population:** 33 physicians from 17 medical, surgical and pediatric specialties and GPT-3.5 Model **Intervention:** ChatGPT(GPT 3.5) used to generate answers to medical queries created by physicians **Comparison:** N/A **Outcome:** Across all questions (n=284), median accuracy score was 5.5 (between almost completely and completely correct) with mean score of 4.8 (between mostly and almost completely correct). The median completeness score was 3 (complete and comprehensive) with a mean score of 2.5. For questions rated easy, medium, and hard, median accuracy scores were 6, 5.5, and 5 (mean 5.0, 4.7, and 4.6; p=0.05). Accuracy scores for binary and descriptive questions were similar (median 6 vs. 5; mean 4.9 vs. 4.7; p=0.07). Of 36 questions with scores of 1-2, 34 were re-queried/re-graded 8-17 days (about 2 and a half weeks) later with substantial improvement (median 2 vs. 4; p<0.01). **Study Design:** Cross-sectional study
OUTCOME LEVEL ASSESSMENT	**Primary Outcome:** Accuracy and completeness of medical information generated by ChatGPT Median accuracy score: 5.5 (between almost completely and completely correct). Mean accuracy score: 4.8 (between mostly and almost completely correct). Median completeness score: 3 (complete and comprehensive). Mean completeness score: 2.5. **Secondary Outcome:** Comparison of accuracy and completeness scores for easy, medium, and hard questions, as well as binary and descriptive questions. Median accuracy scores for easy, medium, and hard questions: 6, 5.5, and 5 (mean 5.0, 4.7, and 4.6; p=0.05). Accuracy scores for binary and descriptive questions were similar (median 6 vs. 5; mean 4.9 vs. 4.7; p=0.07).
RISK OF BIAS	The study acknowledges several risks of bias associated with the use of ChatGPT, including limitations in the sample size and dataset used, as well as biases in selection and respondent factors. As technology continues to evolve, it is possible that the results of this study may change over time.

STUDY REFERENCE [8]

SUMMARY	The pilot study with the aim to evaluate the diagnostic accuracy of differential-diagnosis lists generated by ChatGPT-3 for clinical vignettes with common chief complaints. The study found that ChatGPT-3 can generate differential-diagnosis lists with good diagnostic accuracy. However, the study had several limitations, such as being vignette-based instead of based on real patients' cases, potential biases in AI chatbots, and a lack of transparency in the algorithm. The study suggests that general AI chatbots like ChatGPT-3 can generate well-differentiated diagnosis lists for common chief complaints, but the order of these lists can be improved in the future. Further studies should focus on evaluating more complex cases with well-trained AI chatbots for diagnoses and optimizing collaboration among physicians, patients, and AI in eHealth.
PICOS	**Population:** Clinical vignettes with common chief complaints created by general internal medicine physicians **Intervention:** Use of ChatGPT-3 to generate differential-diagnosis lists for clinical vignettes **Comparison:** Diagnostic accuracy of ChatGPT-3 was compared to the diagnoses made by physicians **Outcomes:** Rate of correct diagnosis by ChatGPT-3 within the ten differential-diagnosis lists: 93.3%.Rate of correct diagnosis by physicians within the five differential-diagnosis lists: 98.3%.Rate of correct diagnosis by ChatGPT-3 within the five differential-diagnosis lists: 83.3%.Rate of correct diagnosis by physicians in the top diagnosis: 53.3%.Rate of correct diagnosis by ChatGPT-3 in the top diagnosis: 93.3%.Rate of consistent differential diagnoses among physicians within the ten differential-diagnosis lists generated by ChatGPT-3: 70.5%. **Study design:** Pilot Study
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** The total rate of correct diagnoses within ten differential-diagnosis lists generated by ChatGPT-3. The ChatGPT-3 system demonstrated a high rate of correct diagnoses within the ten differential-diagnosis lists (93.3%). **Secondary outcomes**: The total rate of correct diagnoses within five differential-diagnosis lists generated by ChatGPT-3. The total rate of correct diagnosis within the five differential-diagnosis lists generated by ChatGPT-3 was 83.3%, with physicians still outperforming the AI system (98.3% vs. 83.3%, p=0.03).
RISK OF BIAS	Study is vignette-based design instead of real patients' cases, potential biases in AI chatbots and a lack of transparency in the ChatGPT-3 algorithm. Additionally, the risk of change in diagnostic accuracy due to incomplete medical information exists for general users.

STUDY REFERENCE [9]

SUMMARY	The study examines ChatGPT's performance in answering questions related to cirrhosis and HCC, with a focus on accuracy, completeness, and reproducibility. It collects questions from professional societies, institutions, and patient support groups, and the responses are graded by two transplant hepatologist reviewers. ChatGPT's performance is also compared to physicians and trainees using published questionnaires. The findings of the study suggest that ChatGPT can provide correct and reproducible responses to the majority of the questions, although most responses were categorized as correct but inadequate. The study also identified ChatGPT's limitations in providing accurate information based on specific cut-offs, treatment durations, and regional guidelines, highlighting its potential as an adjunct tool for patient education rather than a complete replacement of care from licensed healthcare providers.
PICOS	**Population:** Frequently asked questions with cirrhosis or HCC posted by well-regarded professional societies and institutions **Intervention:** Use of ChatGPT to answer 164 questions related to to cirrhosis and HCC **Comparison:** Performance of ChatGPT was compared to GPT-3 and InstructGPT on same datasets **Outcomes:** The outcomes assessed were the accuracy, reproducibility, and quality of the responses generated by ChatGPT when answering questions related to cirrhosis and HCC. ChatGPT regurgitated extensive knowledge of cirrhosis (79.1% correct) and HCC (74.0% correct), but only small proportions (47.3% in cirrhosis, 41.1% in HCC) were labeled as comprehensive. It performed better in areas of basic knowledge, lifestyle, and treatment than in diagnosis and preventive medicine. It correctly answered 76.9% of questions on quality measures but failed to specify decision-making cut-offs and treatment durations. ChatGPT lacked knowledge of regional guideline variations but provided practical and multifaceted advice to patients and caregivers. **Study design:** Descriptive analysis of ChatGPT's performance
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** The evaluation of the accuracy and reproducibility of ChatGPT's responses to questions related to cirrhosis and HCC **Secondary outcomes**: The identification of ChatGPT's limitations in providing accurate information based on specific cut-offs, treatment durations, and regional guidelines, as well as its potential to be used as an adjunct tool for patient education
RISK OF BIAS	Selection of questions, subjectivity in grading process, reviewer bias

STUDY REFERENCE	[12]
SUMMARY	On the USMLE, which consists of the Step 1, Step 2CK, and Step 3 tests, the effectiveness of a sizable language model known as ChatGPT was assessed. Without any extra instruction or reinforcement, ChatGPT passed all three exams with a score at or around the passing mark. Furthermore, ChatGPT's explanations displayed a high degree of concordance and insight. These findings imply that massive language models may be able to support clinical decision-making as well as medical education.
PICOS	**Population:** For Step 1 – second-year students, Step 2CK – fourth-year students and Step 3 – physicians, ChatGPT **Intervention:** 376 questions from the USMLE-2022 sample exam, which was released in June 2022, were made available to the general public. Therefore, for the GPT3 model, all inputs represented genuine out-of-training samples. **Comparison**: Analysis of ChatGBT students' performance distributing them by steps. **Outcomes:** The screening process for all sample exam questions deleted any that contained visual elements including clinical photos, medical photographs, and graphs. 350 USMLE pieces (Step 1: 119, Step 2CK: 102, Step 3: 122) were sent to encoding after filtering. This provides 90% power at = 0.05 to detect a 2.5% increase in accuracy compared to a baseline rate of 60 ±20% (σ), assuming a normal distribution for model performance. **Study design:** Cross-sectional study
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** ChatGPT's accuracy with ambiguous responses censored/included: USMLE Steps 1, 2CK, 3 (free-form queries): 75.0%/45.4%, 61.5%/54.1%, 68.8%/61.5% USMLE Steps 1, 2CK, 3 (multiple choice without justification): 55.8%/36.1%, 59.1%/56.9%, 61.3%/55.7% USMLE Steps 1, 2CK, 3 (multiple choice with justification): 64.5%/41.2%, 52.4%/49.5%, 65.2%/59.8% **Secondary outcomes:** No significant interactions between encoders and question prompt type. Physician agreement: high for open-ended (0.74-0.81) and almost flawless for multiple-choice (>0.9). ChatGPT shows high internal concordance.
RISK OF BIAS	The study has notable limitations, including a restricted input size that constrained the analysis. Exploring ChatGPT's performance across competency types or subject taxonomies and investigating AI failure modes could benefit medical educators and identify performance heterogeneities. Human adjudication, being time-consuming and error-prone, is less reliable than automated methods.

STUDY REFERENCE [14]

SUMMARY	ChatGPT was evaluated on its capacity to answer 50 nuclear medicine-related multiple-choice questions representative of the European Board Examination in Nuclear Medicine. With only 34% accuracy, its performance was deemed subpar, and it displayed confabulation when corrected. The study emphasized ChatGPT's limitations, the need for further robustness testing, and potential misuse in dishonestly passing online exams. The authors warned against relying solely on AI for medical interpretations, as incorrect answers could have harmful consequences. The article compared ChatGPT's performance to adversarial examples in AI, suggesting further investigation into language model robustness is needed. Caution should be exercised when utilizing AI models like ChatGPT in medical interpretations, and more research is necessary to understand their capabilities and limitations in healthcare fields such as nuclear medicine.
PICOS	**Population:** Candidates taking the Fellowship of the European Board Examination in Nuclear Medicine **Intervention**: Use of ChatGPT for answering multiple-choice questions related to nuclear medicine treatments and investigations **Comparison:** ChatGPT's performance with the expected performance based on random chance, measured as the difference between ChatGPT's accuracy and the mean probability of choosing the correct answer by random chance. **Outcomes:** Performance of ChatGPT in answering the multiple-choice questions **Study design:** Performance evaluation/assessment study using 50 example multiple-choice questions
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** Accuracy of ChatGPT in answering the multiple-choice questions, measured as the percentage of correct answers out of the total number of questions answered. In all 50 cases, ChatGPT provided a definitive answer. Marking these against the model answer provided in the training material revealed that ChatGPT was correct only 34% of the time (17/50). With 11 answers requiring the candidate to choose from fve possible responses and the remainder having four possible responses, the mean probability of choosing the correct by random chance was 0.24, suggesting that ChatGPT was likely able to draw on some knowledge rather than simply guessing. **Secondary outcomes:** Frequency of confabulation, measured as the number of instances where ChatGPT provided incorrect answers that were not part of the original question stem. Reproducibility of ChatGPT's answers, measured as the consistency of answers provided by ChatGPT when asked the same question multiple times and context sensitivity of the answers, measured as the variation in answers provided by ChatGPT when asked the same question with different phrasing or in different contexts. Robustness of ChatGPT's performance, measured as the ability to provide accurate answers in the presence of adversarial examples or misleading information.
RISK OF BIAS	The article did not acknowledge any potential biases in the research.

STUDY REFERENCE [21]

SUMMARY	This study evaluated performance of LLMs on Japanese Medica Licensing Examinations. The research question aimed to determine how well these AI models could perform on non-English languages, particularly Japanese. Questions were sourced from past exam papers (2018-2023), and evaluation matrix included both required and general sections along with prohibited choices. Automatic evaluations were done by exact matching because almost all questions are multiple-choice questions with few exceptions that require numbers. The results show that GPT-4 managed to pass all exams, but both GOT-4 and ChatGPT scored below student majority which demonstrated their limitations. The study also highlighted the challenges in multilingually and tokenization when applying LLM-s to non-English languages.
PICOS	**Population:** Large Language Data Models – Chat GPT, GPT-3, GPT-4, ChatGPT-EN **Intervention:** Applying LLMs to solve Japanese medical licensing exam **Comparison:** Comparison of LLMs and Student Majority **Outcome:** Chat GPT-4 managed to pass all exams (2018-2022), but it underperformed student majority baseline **Study Design:** Retrospective performance evaluation on Japanese Medical Licensing Exams
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** Exam performance scores 2018: ChatGPT (266), ChatGPT-EN (281), GPT-3 (209), GPT-4 (382), students (472), passing (368) 2019: ChatGPT (250), ChatGPT-EN (274), GPT-3 (210), GPT-4 (385), students (470), passing (369) 2020: ChatGPT (266), ChatGPT-EN (263), GPT-3 (208), GPT-4 (387), students (471), passing (375) 2021: ChatGPT (297), ChatGPT-EN (277), GPT-3 (203), GPT-4 (398), students (477), passing (369) 2022: ChatGPT (287), ChatGPT-EN (327), GPT-3 (217), GPT-4 (392), students (482), passing (371) 2023 (no prohibited choices): ChatGPT (260), ChatGPT-EN (301), GPT-3 (199), GPT-4 (391), passing (380) **Secondary outcome:** Limitations, model behavior, comparisons GPT-4 underperformed compared to students, emphasizing the need for evaluation in specialized domains. The study reveals LLMs' challenges with translation, tokenization, and specialized knowledge application. GPT-4 struggled with difficult questions, but open-book approaches or retrieval augmentation may improve performance on context-dependent questions.
RISK OF BIAS	Several limitations were identified, such as reproducibility and potential data leakage, language coverage, and scope of evaluation. To address these concerns, the IGAKU QA benchmark and model outputs were released for future research.

STUDY REFERENCE	[22]
SUMMARY	This study evaluates the text-only version of GPT-4, a large language model, on medical competency examinations (USMLE) and MultiMedQA benchmark datasets. GPT-4 is not specialized for medical problems and was not prompted with specialized techniques. The evaluation involves USMLE practice materials, text-based questions, and questions with media elements. GPT-4 exceeds the USMLE passing score by 20 points, outperforming earlier models (GPT-3.5), and models fine-tuned for medical knowledge (Med-PaLM). GPT-4 shows improved calibration and potential uses in medical education, assessment, and clinical practice, while considering challenges of accuracy and safety. Limitations include biases in training data, limited generalizability to other medical tasks, and concerns regarding erroneous recommendations and biases. Further research and careful review are required to address these risks.
PICOS	**Population:** Medical examinations, benchmark datasets and AI models **Intervention**: Evaluating GPT-4 against medical United States Medical Licensing Examination (USMLE), & MultiMedQA benchmark datasets **Comparison:** Performance of GPT-4 model to other AI models including GPT-3.5, Flan-PaLM, Med-PaLM **Outcomes:** GPT-4 exceeds the passing score on USMLE by over 20 points and outperforms other AI models and it outperforms GPT-3.5 and Flan-PaLM 540B on every dataset except PubMedQA. Results also show that GPT-4 performs best on questions that contain only text, it still performs well on questions with media elements, obtaining 70-80% prediction accuracies for these questions on both exams. **Study design:** Experimental study
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** GPT-4 outperformed GPT-3.5 and ChatGPT on USMLE questions, showing a 30-percentage-point increase compared to GPT-3.5. It excelled over GPT-3.5 and Flan-PaLM 540B on all datasets except PubMedQA. **Secondary outcomes:** GPT-4 displayed better calibration than GPT-3.5, with 93% accuracy at 0.96 average probabilities, vital in high-stakes domains like medicine. It achieved 70-80% accuracy on USMLE questions with media elements without seeing images, using logical reasoning and test-taking strategies. Limitations and extensions include exploring richer prompting strategies, memorization effects, and focusing on multiple-choice questions. GPT-4 shows potential in medical applications but requires further research on accuracy, biases, and risks. LLMs' progress has broad implications for various knowledge-intensive professions.
RISK OF BIAS	Risks of bias in this study may include reflection of biases in training data, limited generalizability to other medical tasks, limited qualitative analysis of GPT-4's behavior. Authors mentioned that there are risks of erroneous generations, that could negatively impact patient care.

STUDY REFERENCE [23]

SUMMARY	This study is focused on evaluating the diagnostic competence of ChatGPT in providing diagnoses for neurological cases and comparison of its performance to medical doctors (MDs) and expert neurologists. The analysis includes synthetically generated 200 scenarios covering acute and non-acute neurological cases. The results showed that ChatGPT's diagnostic accuracy surpassed MDs and was nearly identical to expert neurologists in the top five most probable diagnoses. ChatGPT demonstrated comparable diagnostic accuracy in acute neurological cases and showed promise in diagnosing rare and unsolved cases. While the AIs performance is promising, it should be considered as an augmentation tool, and its suggestions must be further evaluated by medical experts.
PICOS	**Population:** Study consists of 200 scenarios covering acute (85) and non-acute (115) neurological cases **Intervention:** Utilizing ChatGPT to provide the most probable diagnosis and top five most probable diagnoses for each case **Comparison:** The diagnostic accuracy of ChatGPT was compared to medical doctors and expert neurologists **Outcomes:** ChatGPT achieved similar diagnostic accuracy to expert neurologists and surpassed the accuracy of general medical doctors. In certain cases where experts failed to provide the correct diagnosis, ChatGPT successfully diagnosed the disease in 40% of the cases. **Study design:** Observational Study
OUTCOME LEVEL ASSESSMENT	**Primary outcome**: ChatGPT's diagnostic accuracy was 68.5%, exceeding MDs (57.08% ± 4.8%) but below expert neurologists (81.58% ± 2.34%). **Secondary outcomes:** ChatGPT, MDs, and experts had similar accuracy in acute cases. In non-acute cases, ChatGPT (66.09%) outperformed MDs (48.41% ± 6.43%) and closely matched experts (64.49% ± 8.84%). It diagnosed 40% of unsolved cases experts missed and had a 60% success rate in its top five diagnoses. Incorrect diagnoses had a 26.98% similarity between ChatGPT and humans, increasing to 37.87% with ChatGPT's top five diagnoses.
RISK OF BIAS	The quality of the data provided by expert neurologists, selection bias in the choice of cases, and potential overfitting of ChatGPT due to the training data.

STUDY REFERENCE [24]

SUMMARY	In this study, the authors evaluated the performance of ChatGPT, in the medical domain by testing it on four unique medical knowledge competency data sets derived from USMLE Step 1 and Step 2 licensing exams. The results demonstrated that ChatGPT achieved an accuracy level comparable to that of a third-year medical student, with a threshold of 60% often considered the benchmark passing standards for both Step 1 and Step 2 exams. The authors also found that even when ChatGPT provided incorrect answers, its responses contained logical explanations for the answer selection. Additionally, the study highlighted the potential of ChatGPT as an innovative tool for small group education in medicine and as a virtual medical tutor. However, it is important to note the limitations and potential risks of bias in this study, including the selection of questions and subjectivity in evaluating the model's responses.
PICOS	**Population:** Performance of AI models, especially ChatGPT **Intervention:** Use of ChatGPT, GPT-3 and InstructGPT to answer questions from medical knowledge data sets (AMBOSS-Step1, AMBOSS-Step2, NBME-Free-Step1, and NBME-Free-Step2) **Comparison:** Performance of ChatGPT was compared to GPT-3 and InstructGPT on same datasets **Outcomes:** The outcomes assessed were the accuracy, coherence, and logical reasoning of the AI models' responses to medical questions. Of the 4 data sets, AMBOSS-Step1, AMBOSS-Step2, NBME-Free-Step1, and NBME-Free-Step2, ChatGPT achieved accuracies of 44% (44/100), 42% (42/100), 64.4% (56/87), and 57.8% (59/102), respectively. **Study design:** Descriptive, comparative analysis
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** The performance of ChatGPT on the medical knowledge data sets, measured by accuracy. **Secondary outcomes:** Comparison of ChatGPT's performance with GPT-3 and InstructGPTS, as well as evaluation of ChatGPT's potential as a virtual medical tutor and small group education tool
RISK OF BIAS	Selection bias in the choice of questions from the AMBOSS and NBME data sets, subjectivity in the qualitative evaluation of ChatGPT's responses, version of ChatGPT is not up to date model.

STUDY REFERENCE [25]

SUMMARY	The study aimed to evaluate ChatGPT's responses on vaccination topics, including the origins of SARS-CoV-2, COVID-19 vaccine conspiracies, and compulsory vaccination. The study utilized a qualitative descriptive approach to analyze ChatGPT's content and assess its correctness, clarity, and conciseness. The authors concluded that ChatGPT provided largely correct, clear, and concise content, with good inter-rater agreement in the evaluation of its responses. The study demonstrated the potential of AI-powered systems like ChatGPT to provide accurate information on topics related to public health, which is crucial during a pandemic to address misconceptions and misinformation. However, it is important to keep in mind that AI systems have limitations, and information should always be verified through reliable sources.
PICOS	**Population:** Open ended questions related to vaccination and COVID-19 that were based on previous studies **Intervention:** Use of ChatGPT to answer questions related to vaccination and COVID-19 related topics **Comparison:** ChatGPT's answers were compared to scientifically accurate information and consensus within scientific community **Outcomes:** Quality of ChatGPT's responses, assessed by correctness, clarity, and conciseness, as well as any bias in the AI's responses. ChatGPT dismissed conspiracy theories about the origin of SARS-CoV-2 and COVID-19 vaccines and remained neutral regarding compulsory vaccination. **Study design:** Descriptive evaluation
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** Correctness, clarity and conciseness in ChatPT responses. ChatGPT's responses dismiss conspiracy theories and provide a neutral stance on compulsory vaccination, detailing its pros and cons. **Secondary outcomes:** Assessment of biases in generated content that is done using Cohen's kappa.
RISK OF BIAS	Subjection evaluation of content, descriptive nature of study, AI itself can produce biased data

STUDY REFERENCE	[26]
SUMMARY	This feasibility study evaluates the use of GPT-4 to automate the conversion of free-text radiology reports into structured templates. The study assessed GPT-4's performance using 170 detailed CT and MRI scan reports in English and 583 chest X-ray reports in German. The study found that GPT-4 successfully transformed all 170 free-text radiology reports into valid JSON files without error, identified all key findings without loss of accuracy, and selected appropriate report templates for the report text and main findings. On a chest X-ray classification benchmark, GPT-4 outperformed medBERT.de in detecting three out of four pathological findings and one therapeutic device category.
PICOS	**Population:** Radiology reports from various body regions and examinations. **Intervention:** Use of GPT-4 to automate the conversion of free-text radiology reports into structured template **Comparison:** GPT-4's performance was compared to the medBERT.de chest x-ray classification benchmark and state-of-the-art models in the German language. **Outcomes:** GPT-4 demonstrated the ability to effectively transform all free-text radiology reports into structured templates without any errors, identified all key findings accurately, and outperformed the existing state-of-the-art in certain tasks. **Study design:** Multilingual feasibility study.
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** Effective transformation of free-text radiology reports into structured templates **Secondary outcomes:** Identification of key findings, selection of the most appropriate report template, performance on the medBERT.de chest X-ray benchmark
RISK OF BIAS	GPT-4's restricted access, which requires potentially sensitive data to be shared with third parties, conflicts with privacy laws. The study uses fictitious CT and MRI reports, which may not fully represent real-world radiology report complexities.

STUDY REFERENCE [27]

SUMMARY	The article assesses ChatGPT's ability to answer ophthalmology multiple-choice questions from two popular Ophthalmic Knowledge Assessment Program (OKAP) exam banks: the American Academy of Ophthalmology's Basic and Clinical Science Course (BCSC) Self-Assessment Program and the OphthoQuestions online question bank. Tested on easy-to-moderate difficulty questions covering various ophthalmic areas, ChatGPT (GPT-3.5 series, January 9 version) scored 55.8% and 42.7% on the simulated exams, with performance varying across subspecialties. The study suggests domain-specific pretraining may be needed to enhance LLM performance in ophthalmic subspecialties. The article also introduces AI and deep learning's application in ophthalmology, specifically in natural language processing, and discusses ChatGPT, a generic LLM optimized for dialogue. It emphasizes the novelty of evaluating LLMs in ophthalmology and outlines the methods, including ChatGPT and the utilized question banks.
PICOS	**Population:** Ophthalmology trainees or residents (United States and Canada). **Intervention**: Use of ChatGPT for answering ophthalmology-related questions. **Comparison:** Performance of ophthalmology trainees or residents without using ChatGPT. **Outcomes:** Accuracy of ChatGPT in answering multiple-choice questions from the two ophthalmology question banks: the American Academy of Ophthalmology's Basic and Clinical Science Course (BCSC) Self-Assessment Program and the OphthoQuestions online question bank, with a total of 520 questions. **Study design:** Performance evaluation/assessment study of ChatGPT's performance against human ophthalmology trainees or residents in a controlled setting.
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** Accuracy of ChatGPT in providing correct answers to ophthalmology-related questions, as measured by the percentage of correct responses compared to gold standard answers or expert opinions. ChatGPT achieved an accuracy of 55.8% on the simulated OKAP exam using the BCSC testing set and 42.7% on the OphthoQuestions testing set (humans score 74% on the BCSC question bank and 61% on OphthoQuestions, with first-year residents scoring an average of 53% on OphthoQuestions). The highest accuracy was in General Medicine (75%) and the second highest was in Fundamentals (60%) and Cornea (60%). ChatGPT did not perform as well in Neuro-ophthalmology (25%), Glaucoma (37.5%), and Pediatrics and Strabismus (42.5%). **Secondary outcomes:** Time taken to receive answers from ChatGPT compared to traditional methods. Trainees' or residents' perception of ChatGPT's usefulness and reliability as a learning tool, as assessed through surveys or questionnaires. Trainees' or residents' performance in formal assessments, such as the OKAP exam or other ophthalmology examinations, before and after using ChatGPT.
RISK OF BIAS	While performing the testing a new session was started in ChatGPT for every question to reduce memory retention bias. But as the performance of ChatGPT improves, it will be necessary to include protecting vulnerable populations from biases and evaluating the potential harm or risk of acting on the answers provided by ChatGPT.

STUDY REFERENCE [28]

SUMMARY	The study, conducted by the All India Institute of Medical Sciences in February 2023, analyzed ChatGPT's ability to answer first- and second-order microbiology questions based on the competency-based medical education (CBME) curriculum. Expert microbiologists assessed content validity and scored ChatGPT's responses. The results indicate that ChatGPT has potential as an automated microbiology question-answering tool but requires further improvements in training and development. ChatGPT could assist medical students in self-directed learning by providing personalized access to relevant information.
PICOS	**Population:** Medical students studying microbiology as part of their competency-based medical education (CBME) curriculum. **Intervention**: Using ChatGPT to converse with and obtain responses to first-order and second-order knowledge questions related to microbiology. **Comparison:** The study compared the accuracy of ChatGPT in answering first-order and second-order knowledge questions in microbiology. **Outcomes:** Assessing the capability of ChatGPT in answering microbiology questions based on a set of 96 questions validated by expert microbiologists. **Study design:** Cross-sectional observational study conducted at the department of Microbiology, Pathology, and Physiology.
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** ChatGPT accurately answered 80% of first- and second-order microbiology questions, with an average score of 4.04 ±0.37 (median 4.17) on a 0-5 scale. No significant difference in performance between question types (Mann-Whitney $p=0.4$) and overall score significantly below the maximum (one-sample median test $p<0.0001$). **Secondary outcomes:** Comparisons of ChatGPT's accuracy for different question types, module-wise performance analysis, potential as an automated question-answering tool, and capability to assist self-directed learning through qualitative feedback analysis.
RISK OF BIAS	In order to reduce biases and increase the accuracy and reliability of the evaluation the study employed three evaluators to assess an answer. Different evaluators may have different biases based on their personal experiences, perspectives, or preferences. But the scoring of the responses was subjective; hence, evaluation bias may be present even after taking the average of three scores from three evaluators.

STUDY REFERENCE [29]

SUMMARY	The aim of experiment performed with ChatGPT is to see if it could take and potentially pass examinations used for undergraduate admissions is described in this publication. The Italian Medical School Admission Test (IMSAT) and the Cambridge Bio Medical Admission Test (BMAT) were the two exams selected for biomedical undergraduate entrance. On the two tests, ChatGPT performed significantly differently. There was a theory developed that may account for these variations, and it is currently the focus of additional research. The most unexpected finding of this preliminary study is that ChatGPT received an IMSAT score that qualified it for instant admission to the sixth-best medical school in Italy.
PICOS	**Population:** Medical students, IMSAT, BMAT, ChatGPT **Intervention:** ChatGPT answers questions from National testing of medical students in Italy **Comparison:** The analysis of two different exams (IMSAT and BMAT). **Outcomes:** The ChatGPT performance was different in each of tests because of focus on different variations and parameters of exam. **Study design:** Cross-sectional study using ChatGPT and evaluating the students using scoring methods.
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** According to the findings of research, ChatGPT performed better overall on the IMSAT test (62% of accurate answers) than the BMAT test (49% of correct answers). In terms of the IMSAT exam, ChatGPT's score (46.3) is significantly higher than the minimal score (33.4) necessary to be admitted to an Italian medical school for the academic year 2022-2023. **Secondary outcomes**: It is reasonable to claim that the knowledge and skills demanded by the first IMSAT test are essentially the same as those demanded by the Writing Skills section of the BMAT. They are mostly concerned with language comprehension and proper usage in non-specialist circumstances.
RISK OF BIAS	The questions that are brought up by this kind of research are important and ought to be addressed in a wider context. In a time when communication with intelligent systems of the ChatGPT type will become more frequent and unavoidable, the most pressing question is what knowledge and skills should be required of students about to enter university level studies.

STUDY REFERENCE [30]

SUMMARY	Gastroenterology (GI) is a field that is constantly changing. The most urgent and significant research questions must be identified. to assess ChatGPT's potential for defining GI research priorities and serve as a foundation for more study. On four important GI topics—inflammatory bowel illness, microbiome, artificial intelligence in GI, and enhanced endoscopy in GI—we conducted ChatGPT queries. On a scale of 1 to 5, with 5 being the most significant and pertinent to ongoing GI research, a group of knowledgeable gastroenterologists independently assessed and rated the produced research questions. ChatGPT produced pertinent and understandable research questions. However, the panel of gastroenterologists did not think the queries were very unique. The Large Language Models (LLMs) could be a beneficial tool for determining the most important areas of GI research, more needs to be done to make the produced research topics more innovative.
PICOS	**Population:** 3 gastroenterologists, ChatGPT **Intervention:** Each topic has its own set of five research questions, for a total of 20 research questions. A group of skilled gastroenterologists with knowledge of the related subject areas then evaluated and scored each of these questions separately. **Comparison**: The research questions produced by ChatGPT were compared to those being investigated currently in the field of GI, as determined by a thorough literature review. **Outcomes:** Excellent performance was shown by ChatGPT in terms of clarity and relevance, as well as respectable results in terms of specificity and originality. **Study design:** Experimental study
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** According to the results of the expert review, ChatGPT was successful in producing research questions that were very pertinent to the field of inflammatory bowel disease (IBD), with the majority of these questions obtaining the highest relevance rating of 5, and a mean grade of 4.9±0.26. ChatGPT did exceptionally well in terms of clarity, with the majority of questions obtaining a rating of 4 or 5, and a mean grade of 4.8±0.41. The ChatGPT achieved a mean grade of 2.86 0.64 for specificity, which is a passable performance. However, all scores for originality were extremely poor, with a mean of 1.07±0.26. The outcomes for microbiome-related subjects were comparable to those for IBD. Similar to IBD, grades for relevance and clarity were practically at their highest, while those for originality were at their lowest. **Secondary outcomes**: Question 1 was the same for both themes. Originality, clarity, and specificity had mean SD values of 4.93±0.26, 1.13±0.35, 4.93±0.26, and 3.13±0.64, respectively. The results for improved endoscopy and artificial intelligence follow a similar pattern, with excellent relevance and clarity, good specificity, and low originality. The average results for AI on each of the aforementioned metrics are 5±0, 4.33±0.89, 3.2±0.67, and 1.87±0.99, respectively. The average scores for relevance, clarity, specificity, and originality for advanced endoscopy were 4.89, 4.47, 0.74, 3.2, and 1.73, respectively.
RISK OF BIAS	To ensure generalizability, further research with larger, diverse expert panels is needed. ChatGPT's performance was assessed through subjective judgments, which may be biased. Objective metrics, such as citation frequency or impact factor, would offer a more precise evaluation.

STUDY REFERENCE [31]

SUMMARY	The goal of this study was to ascertain whether ChatGPT can handle higher-order medical biochemistry-related issues. This survey was conducted online using ChatGPT's 14 March 2023 edition, which is now available for free to registered users. It was given 200 questions about medical biochemistry reasoning that call for higher-order thinking. These inquiries were chosen at random from the institution's question bank and categorized in accordance with the competency modules of the Competency-Based Medical Education (CBME) curriculum. The responses were gathered and saved for future analysis. Two knowledgeable academic biochemists scored the responses from zero to five. Using fictitious numbers, a one-sample Wilcoxon signed rank test was used to assess the score's correctness.
PICOS	**Population:** Medical biochemistry students, Chat GPT **Intervention:** The purpose in this study was to find out if ChatGPT can deal with more complex medical biochemistry issues. **Comparison**: Comparing ChatGPT (version March 14, 2023) with older one by setting 200 questions from biochemistry fields. **Outcomes:** The findings of this study suggest that ChatGPT, with a median score of four out of five, has the potential to be an effective tool for answering problems demanding higher-order thinking in medical biochemistry. **Study design:** Cross-sectional study
OUTCOME LEVEL ASSESSMENT	**Primary outcome:** 200 questions demanding higher-order thinking were answered by AI software, with a median score of 4.0 (Q1=3.50, Q3=4.50). Using a Wilcoxon signed rank test on a single sample, the outcome was lower than the hypothetical maximum of five (p=0.001) and close to four (p=0.16). There was no difference in the answers to the medical biochemistry-related questions from the various CBME modules (Kruskal-Wallis p=0.39). Two professors of biochemistry who participated in the scoring had excellent inter-rater reliability (ICC=0.926 (95% CI: 0.814-0.971); F=19; p0.001). **Secondary outcomes**: Using descriptive statistical tests, the results were presented as number, mean, median, standard deviation, and first and third quartiles which represented detailed analysis of results mentioned above.
RISK OF BIAS	This study has a number of limitations. First, a grading system was used with a range from 0 to 5. Even though the answer keys had been prepared previously, there might still have been a subjective bias in the evaluation that was out of research control. Other institutions may have different questions; the ones which were utilized in study came from question bank. For a more broadly applicable result, future research may need to be multicentric. Future research should take this into account because even a small change to a query could cause ChatGPT to respond differently.

5. DISCUSSION

5.1 Overview of Findings

The review of research articles encompassed a wide array of medical topics such as neurology [1][23], radiology [6][11][26], ophthalmology [27], nuclear medicine [14], microbiology [28], and pathology [2], cirrhosis [9], parasitology [5], biochemistry [31] and gastroenterology [30] among others. Potential applications of ChatGPT and future models were identified in these diverse medical domains, including student self-learning [10][24][28][2], medical report generation [1][4][14][6][11][26][27], and patient diagnosis [1][6][7][8][9]. ChatGPT has demonstrated promising results across these applications.

In addition, ChatGPT was found to remain neutral in questions related to COVID-19 and vaccination, while avoiding the spread of conspiracies [25]. Some articles explored ChatGPT's performance on medical examinations [1][4][5][10][21][31].

Despite the promising findings, most of the articles emphasized the need for future research, addressing potential biases in data and the model's limited knowledge up to 2021. The authors cautioned that ChatGPT should not yet be considered a reliable source of information, highlighting various challenges and limitations associated with the model.

5.2 Challenges and Limitations

The reviewed articles identified several challenges and limitations of ChatGPT in medical applications, including:

- Struggling with lengthy questions [1]: The ChatGPT model faced difficulties with accurately processing and responding to lengthy or complex questions, although the GPT-4 model showed enhanced ability.
- Inability to incorporate imaging data [1][4][21][27]: Both ChatGPT and GPT-4 struggled with image-based questions, limiting their effectiveness in certain medical scenarios.
- Sensitivity to wording of prompt [15][22][31]: The model's sensitivity to question phrasing or prompting might lead to inconsistent or less accurate responses, emphasizing the need for refining its natural language understanding capabilities.
- Ethical and privacy concerns [14][6][11][7][21][26]: Ensuring the responsible use of ChatGPT in healthcare settings while respecting patient privacy is crucial.
- Tokenization and multilingualism [21][22]: The model needs to better understand and process text in multiple languages and various medical terminologies for its broader adoption in global healthcare settings.
- Bias in data [6][8][9]: Addressing potential biases in training data is essential to improve the model's accuracy and reliability in diverse medical contexts.

Moreover, there are limitations related to the design of the reviewed studies that need to be considered, including:

- Modest sample size [7]: Smaller sample sizes may limit the statistical power and the generalizability of the study findings.
- Bias in evaluation [7][9][25][28][2][29][31]: Potential biases in the evaluation process can impact the reliability of the results and conclusions drawn from the studies.
- Selection of questions [1][7][9][29][31][10]: The choice of questions used in the studies may affect the performance of ChatGPT, and a more diverse set of questions could lead to different results.
- Scope of evaluation [6][21][22][29]: The scope of the evaluation might be limited, and future studies should include a broader range of scenarios and contexts to test the performance of ChatGPT.
- Use of synthetic data [23][8][26]: The use of synthetic data in some studies might not accurately represent real-world data, which can limit the applicability of the findings.
- Content validation [8][31][4]: Ensuring that the content generated by ChatGPT is valid and accurate is essential for its practical applications in healthcare.
- Limited generalizability [1][22][23][8][31]: Most studies have limited generalizability because results may vary depending on the specific questions and scenarios tested, which

makes it difficult to generalize the findings to other contexts.

5.3 Future Research Directions

Given the identified challenges and limitations, future research should focus on the following aspects.

Improving the model's language understanding and processing capabilities to address multilinguality and tokenization issues.

Developing techniques to enhance the model's ability to handle lengthy or complex questions more effectively.

Investigating and refining prompting strategies to ensure more consistent and accurate responses.

Assessing and minimizing the risk of errors when advising patients, potentially by incorporating input from healthcare professionals during the development and implementation of the model

Furthermore, future research should continue exploring potential applications of ChatGPT and its successors in various medical fields, while closely monitoring and evaluating its performance, ethical implications, and potential risks.

In conclusion, ChatGPT has demonstrated promising results across a wide range of medical applications. However, several challenges and limitations need to be addressed before it can be considered a reliable source of information in healthcare. Continued collaboration between researchers, healthcare professionals, and policymakers is essential to harness the full potential of AI-powered chatbots like ChatGPT, ultimately improving patient outcomes and transforming the future of medicine.

6. CONCLUSION

In conclusion, our systematic review highlights the promising potential of ChatGPT in various medical applications, including simplifying complex tasks, enhancing patient care, improving clinical decision-making, and facilitating communication among healthcare professionals. While the studies reviewed demonstrate the versatility of ChatGPT in diverse medical fields, they also identify several challenges and limitations that need to be addressed to ensure the safe and effective implementation of this technology in healthcare.

Despite the limitations, the evidence gathered through this review emphasizes the transformative potential of ChatGPT in the field of medicine. Future research should focus on addressing the identified challenges and refining the performance of ChatGPT to optimize its use in clinical practice. As we continue to explore the possibilities and potential applications of AI in medicine, it is imperative that researchers, healthcare professionals, and policymakers work together to navigate the complex ethical, privacy, and safety concerns that accompany these advancements.

By critically evaluating the current state of ChatGPT in medicine, this review serves as a foundation for further exploration and development in this domain. As the field of AI continues to grow and evolve, so will the potential impact of technologies like ChatGPT on healthcare delivery. With ongoing research and collaboration, the integration of ChatGPT and similar AI-powered tools may well revolutionize the way we practice medicine, ultimately leading to improved patient outcomes and a transformed healthcare landscape.

ACKNOWLEDGEMENTS

During the preparation of this manuscript, we used OpenAI's ChatGPT to assist with paraphrasing and improving the wording of selected sections.

REFERENCES

[1] R. Ali *et al.*, “Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations”, doi: 10.1101/2023.03.25.23287743.

[2] R. K. Sinha, A. Deb Roy, N. Kumar, and H. Mondal, “Applicability of ChatGPT in Assisting to Solve Higher Order Problems in Pathology,” *Cureus*, Feb. 2023, doi: 10.7759/cureus.35237.

[3] J. Homolak, “Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma,” *Croat Med J*, vol. 64, no. 1, 2023.

[4] J. Li, A. Dada, J. Kleesiek, and J. Egger, “ChatGPT in Healthcare: A Taxonomy and Systematic Review,” 2023, doi: 10.1101/2023.03.30.23287899.

[5] S. 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*, vol. 20, p. 1, Jan. 2023, doi: 10.3352/jeehp.2023.20.1.

[6] K. Jeblick *et al.*, "ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports," Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.14882

[7] D. Johnson *et al.*, "Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model," 2023, doi: 10.21203/rs.3.rs-2566942/v1.

[8] T. Hirosawa, Y. Harada, M. Yokose, T. Sakamoto, R. Kawamura, and T. Shimizu, "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*, vol. 20, no. 4, Feb. 2023, doi: 10.3390/ijerph20043378.

[9] Y. Hui Yeo *et al.*, "Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma." [Online]. Available: https://orcid.org/0000-0002-4608-6896

[10] K. Bhattacharya, A. S. Bhattacharya, N. Bhattacharya, V. D. Yagnik, P. Garg, and S. Kumar, "ChatGPT in Surgical Practice—a New Kid on the Block," *Indian Journal of Surgery*. Springer, 2023. doi: 10.1007/s12262-023-03727-x.

[11] Ö. Cite ; Aydin and E. Karaarslan, "OpenAI ChatGPT interprets Radiological Images: GPT-4 as a Medical Doctor for a Fast Check-Up Enhancing Security In RFID View project Blockchain-based Secure Systems View project Enis Karaarslan Mugla Üniversitesi as a Medical Doctor for a Fast Check-Up OpenAI ChatGPT interprets Radiological Images: GPT-4 as a Medical Doctor for a Fast Check-Up Enis KARAARSLAN," APA, 2023. [Online]. Available: https://www.researchgate.net/publication/369744392

[12] T. H. Kung *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, p. e0000198, Feb. 2023, doi: 10.1371/journal.pdig.0000198.

[13] A. Thurzo, M. Strunga, R. Urban, J. Surovková, and K. I. Afrashtehfar, "Impact of Artificial Intelligence on Dental Education: A Review and Guide for Curriculum Update," *Education Sciences*, vol. 13, no. 2. MDPI, Feb. 01, 2023. doi: 10.3390/educsci13020150.

[14] I. L. Alberts *et al.*, "Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?," *European Journal of Nuclear Medicine and Molecular Imaging*. Springer Science and Business Media Deutschland GmbH, 2023. doi: 10.1007/s00259-023-06172-w.

[15] J. M. Drazen *et al.*, "Chatbot for Medicine," 2023.

[16] J. Dahmen *et al.*, "Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword," *Knee Surgery, Sports Traumatology, Arthroscopy*. Springer Science and Business Media Deutschland GmbH, Apr. 01, 2023. doi: 10.1007/s00167-023-07355-6.

[17] C. Baumgartner, "The opportunities and pitfalls of ChatGPT in clinical and translational medicine," *Clin Transl Med*, vol. 13, no. 3, Mar. 2023, doi: 10.1002/ctm2.1206.

[18] I. S. Abdullah, A. Loganathan, and R. W. Lee, "ChatGPT & Doctors: The Medical Dream Team," Health Sciences Commons. [Online]. Available: https://hsrc.himmelfarb.gwu.edu/smhs_URGENT_Matters

[19] I. Dergaa, K. Chamari, P. Zmijewski, and H. Ben Saad, "From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing," *Biol Sport*, 2023, doi: 10.5114/biolsport.2023.125623.

[20] O. Temsah *et al.*, "Overview of Early ChatGPT's Presence in Medical Literature: Insights From a Hybrid Literature Review by ChatGPT and Human Experts.," *Cureus*, vol. 15, no. 4, p. e37281, Apr. 2023, doi: 10.7759/cureus.37281.

[21] J. Kasai, Y. Kasai, K. Sakaguchi, Y. Yamada, and D. Radev, "Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.18027

[22] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on Medical Challenge Problems," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.13375

[23] B. Nógrádi *et al.*, "ChatGPT M.D.: Is there any room for generative AI in neurology and other medical areas?" [Online]. Available: https://ssrn.com/abstract=4372965

[24] A. 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*, vol. 9, 2023, doi: 10.2196/45312.

[25] M. Sallam *et al.*, "ChatGPT Output Regarding Compulsory Vaccination and COVID-19 Vaccine Conspiracy: A Descriptive Study at the Outset of a Paradigm Shift in Online Search for Information," *Cureus*, Feb. 2023, doi: 10.7759/cureus.35029.

[26] L. C. Adams *et al.*, "Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study Article type: Original Research." [Online]. Available: https://github.com/kbressem/gpt4-structured-reporting.

[27] F. Antaki, S. Touma, D. Milad, J. El-Khoury, and R. Duval, "Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings," *medRxiv*, 2023.

[28] D. Das *et al.*, "Assessing the Capability of ChatGPT in Answering First- and Second-Order Knowledge Questions on Microbiology as per Competency-Based Medical Education Curriculum," *Cureus*, Mar. 2023, doi: 10.7759/cureus.36034.

[29] M. Giunti, F. Giulia Garavaglia, R. Giuntini, S. Pinna, and G. Sergioli, "CHATGPT PROSPECTIVE STUDENT AT MEDICAL SCHOOL." [Online]. Available: https://ssrn.com/abstract=4378743

[30] A. Lahat, E. Shachar, B. Avidan, Z. Shatz, B. S. Glicksberg, and E. Klang, "Evaluating the use of large language model in identifying top research questions in gastroenterology," *Sci Rep*, vol. 13, no. 1, p. 4164, Mar. 2023, doi: 10.1038/s41598-023-31412-2.

[31] A. Ghosh and A. Bir, "Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry," *Cureus*, Apr. 2023, doi: 10.7759/cureus.37023.

[32] A. Liberati *et al.*, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration.," *BMJ*, vol. 339, 2009, doi: 10.1136/bmj.b2700.

[33] C. M. da C. Santos, C. A. de M. Pimenta, and M. R. C. Nobre, "The PICO strategy for the research question construction and evidence search," *Rev Lat Am Enfermagem*, vol. 15, no. 3, 2007, doi: 10.1590/s0104-11692007000300023.