Global Pediatric Health
Volume 11: 1–8
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2333794X241240327
journals.sagepub.com/home/gph
Sage

ChatGPT Yields a Passing Score on a Pediatric Board Preparatory Exam but Raises Red Flags

Mindy Le, BS[1] and Michael Davis, MD, MBA[1]

Abstract

Objectives. We aimed to evaluate the performance of a publicly-available online artificial intelligence program (OpenAI's ChatGPT-3.5 and -4.0, August 3 versions) on a pediatric board preparatory examination, 2021 and 2022 PREP® Self-Assessment, American Academy of Pediatrics (AAP). *Methods.* We entered 245 questions and answer choices from the Pediatrics 2021 PREP® Self-Assessment and 247 questions and answer choices from the Pediatrics 2022 PREP® Self-Assessment into OpenAI's ChatGPT-3.5 and ChatGPT-4.0, August 3 versions, in September 2023. The ChatGPT-3.5 and 4.0 scores were compared with the advertised passing scores (70%+) for the PREP® exams and the average scores (74.09%) and (75.71%) for all 10 715 and 6825 first-time human test takers. *Results.* For the AAP 2021 and 2022 PREP® Self-Assessments, ChatGPT-3.5 answered 143 of 243 (58.85%) and 137 of 247 (55.46%) questions correctly on a single attempt. ChatGPT-4.0 answered 193 of 243 (79.84%) and 208 of 247 (84.21%) questions correctly. *Conclusion.* Using a publicly-available online chatbot to answer pediatric board preparatory examination questions yielded a passing score but demonstrated significant limitations in the chatbot's ability to assess some complex medical situations in children, posing a potential risk to this vulnerable population.

Keywords
ChatGPT, pediatrics

Received November 19, 2023. Received revised January 28, 2024. Accepted for publication March 4, 2024.

What's Known on This Subject

The recent surge in popularity of artificial intelligence and publicly-available online chatbots has captured the attention of multiple stakeholders in the field of pediatrics. Care providers of children should be cognizant of the limitations and risks of these emerging technologies.

What This Study Adds

We evaluated the performance of a publicly-available online artificial intelligence program (OpenAI's ChatGPT-3.5 and -4.0, August 3 versions) on a pediatric board preparatory examination (2021 and 2022 PREP® Self-Assessment, American Academy of Pediatrics).

Introduction

Artificial intelligence (AI) is an evolving computer science field focused on the development of machines that can perform tasks that would typically require human intelligence, such as visual perception, speech recognition, decision-making, and language understanding.[1] AI is achieved through a combination of techniques, including machine learning, neural networks, and natural language processing (NLP).[2] NLP leverages the use of large language models (LLMs) to analyze and understand natural language text and speech. LLMs are complex neural network-based systems that are trained on massive amounts of text data to learn the patterns, grammar, syntax, and semantics of natural language.[3] LLMs can process and generate text in a way that simulates human language comprehension and production. It's important to note, however, that LLMs are not truly "understanding" language in the same way humans do; they are statistical models that predict the likelihood of words and phrases given the context they've been trained on. Despite

[1]University of Florida College of Medicine, Gainesville, FL, USA

Corresponding Author:
Mindy Le, University of Florida College of Medicine, 7046 SW Archer Road, Gainesville, FL 32608, USA.
Email: mindymle@ufl.edu

significant limitations, LLMs have revolutionized the field of NLP and are driving advancements in AI-powered language-related applications in healthcare such as voice assistants, chatbots, language translation, and sentiment analysis.[4,5]

Recently, online interfaces have become publicly available that are capable of dispensing healthcare advice over a broad array of categories, although the quality and safety of these responses are in question.[6-8] A current example of this technology is OpenAI's ChatGPT, an online chatbot offered to the public free of charge beginning in November 2022.[9] ChatGPT was estimated to have reached 100 million monthly active users just 2 months after launch, making it the fastest-growing consumer application in history, according to UBS analysts.[10] Through a simple webpage interface, anyone with Internet access can easily pose text-based questions of varying length and complexity with responses that are surprisingly human-like.[11] ChatGPT responds nearly instantaneously to both the informal questions posed by patients and detailed scientific questions posed by physicians.[12] Answers are routinely highly-detailed, insightful, and convincing even when inaccurate.[11]

Elicited, in part, by the COVID-19 pandemic, virtual healthcare has proliferated[13] and this has been accompanied by record levels of physician workload and burnout.[14,15] There has been a concomitant rise in patient's expectations for electronic communication of medical advice *in lieu* of in person visits.[16] There are several potential benefits to the use of chatbots for addressing the medical concerns of caregivers for children. For one, chatbots are available anytime and can provide immediate answers, which can be especially helpful for parents who are dealing with a sick child late at night or on weekends when medical offices are closed.[17,18] Chatbots could also help reduce the work burden of healthcare providers, who often work long hours and may not have the time or resources to answer every question posed by caregivers in virtual messages.[19]

There is a duality of excitement and apprehension surrounding the application of chatbots in a variety of healthcare scenarios.[20-23] There is enthusiasm that chatbots will reduce administrative burden for healthcare providers while increasing access to useful health information for patients and the caregivers of young children. On the other hand, the quality of dispensed medical advice, as with any form of LLM output, requires supervision by relevant professionals to ensure accuracy and safety. The optimal method of quality assurance has not yet been determined. The purpose of this study was to test the "knowledge" of a publicly-available chatbot in the same way human pediatricians are tested for board certification. We chose the American Academy of Pediatrics (AAP) 2021 and 2022 PREP® Self-Assessment tools and OpenAI's ChatGPT for this purpose.

Methods

In compliance with fair use copyright law and with methods deemed exempt by our institutional review board, we were able to enter 245 questions and answer choices from the Pediatrics 2021 PREP® Self-Assessment (1 question was not available and 2 questions were excluded by ChatGPT due to a possible violation of content policy) and 247 questions and answer choices from the Pediatrics 2022 PREP® Self-Assessment into OpenAI's ChatGPT-3.5 (available free of charge) and ChatGPT-4.0 (available by paid subscription), August 3 versions, in September 2023. The 2021 version of PREP® Self-Assessment was chosen over more recent versions in order to precede ChatGPT's knowledge cutoff date, September 2021, in an attempt to match the chatbots training data temporally with the publishing date of the questions entered. Pediatrics 2022 PREP® Self-Assessment was also tested as a means of evaluating test materials that were only available after the ChatGPT knowledge cutoff date to ensure that ChatGPT did "cheat" by simply finding the answer key to the self-assessment questions posted on the Internet.

ChatGPT "Takes the Test"

A fresh chatbot session was created, the original full text of each question was entered individually, and the chatbot response was saved. One chatbot session was utilized for all questions for each PREP® test, with understanding that the chatbot is able to use previously-entered text and generated responses to affect future responses within the same session or "thread." This was felt to be representative a human test taker's experience of answering all questions in one uninterrupted session. Each self-assessment question and its 4 potential answer choices were entered into the online ChatGPT interface individually as the "prompt," maintaining formatting to the greatest degree possible in text-only format. The process was repeated for each individual question in each version of ChatGPT (-3.5 and -4.0).

Scoring ChatGPT's Performance

Responses were compared to the answer key accompanying the PREP® Self-Assessment. Media files (ie, photos, growth charts, graphs) were ignored by the text-only chatbot. Table formatting was lost when the questions were transferred to text-only format, however tables

Table 1. Performance of ChatGPT-3.5 and ChatGPT-4.0 on AAP Pediatrics 2021 PREP® Self-Assessment.

AAP pediatrics 2021 PREP®	ChatGPT-3.5	ChatGPT-4.0
Self-assessment questions[a]	Correct responses	Correct responses
With media (n = 53)	30 (56.6%)	40 (75.47%)
Without media (n = 190)	113 (59.47%)	153 (80.52%)
With table(s) (n = 33)	17 (51.51%)	28 (84.84%)
Without table(s) (n = 210)	126 (60%)	165 (78.57%)
With table(s) >2 columns (n = 5)	0 (0%)	4 (80%)
All questions (n = 243)	143 (58.85%)	193 (79.84%)

[a]Passing score 70%+, average score 74.09%, n = 10715.

with 2 columns retained appropriate juxta positioning of data and labels. Tables with 3 or more columns lost appropriate ordering of data. Individual questions were scored and a cumulative percentage score was determined for both versions of ChatGPT and both 2021 and 2022 PREP® questions. Scores were determined for questions with and without media and/or tables. The performances of ChatGPT-3.5 and ChatGPT-4.0 on AAP Pediatrics 2021 and 2022 PREP® Self-Assessments were then assessed by question category for individual pediatric specialties.

Ethical Approval and Informed Consent

Informed consent was not required and this study was classified as exempt by our Institutional Review Board because the data were deemed sufficiently generic and did not contain identifiable information.

Results

ChatGPT seemed to understand the intent of our prompts: to answer a multiple-choice question with a single answer and to provide explanation for the choice. Given that every prompt had 4 possible answer choices, the results were significantly better than chance. Because ChatGPT was not able to interpret media, such as images or charts, these media were excluded. ChatGPT-3.5's performance was negatively affected when a board question included a media file that could not be interpreted, however, ChatGPT-4.0's performance was minimally affected for these prompts. Performance was negatively affected for ChatGPT-3.5 when a table was included in the prompt (51.51% correct), especially for questions with multiple tables or tables with >2 columns (0% correct). ChatGPT-4.0 did not seem to have difficulty interpreting tables with lost formatting. Performance of ChatGPT-4.0 was similar when unformatted tables were present or absent, even when tables contained >2 columns.

All responses returned within a few seconds and were accompanied by one or more supporting paragraphs of explanation. Reasoning was provided for the selected answer choices and there was also discussion at to why the other answers were not chosen. All explanations were thorough and descriptive, even when incorrect.

For the AAP PREP® Self-Assessments, media files (ie, photos, growth charts, graphs) accompanied 100 (PREP® 2021: 53, PREP® 2022: 47) of the questions and these files were ignored by the text-only chatbot. Of the 100 questions with media, 59 (59%) [PREP® 2021: 30/53 (56.6%), PREP® 2022: 29/47 (61.7%)] were correctly answered by ChatGPT-3.5 and 80 (80%) [PREP® 2021: 40/53 (75.47%), PREP® 2022: 40/47 (85.1%)] were answered correctly by ChatGPT-4.0 (Tables 1 and 2). Table formatting was lost when the questions were transferred to text-only format, however tables with 2 columns retained appropriate juxta positioning of data and labels. Tables with 3 or more columns lost appropriate ordering of data. There were 76 (PREP® 2021: 33, PREP® 2022: 43) questions with data tables and of these, ChatGPT-3.5 answered 36 [PREP® 2021: 17/33 (51.51%) and 2022: 19/43 (44.18%)] correctly and ChatGPT-4.0 answered 63 [PREP® 2021: 28/33 (84.84%) and 2022: 35/43 (81.39%)] correctly. When results were separated by pediatric specialty, ChatGPT-4.0 yielded improvement over ChatGPT-3.5 in 23 out of 36 categories for PREP® 2021 and 25 out of 36 categories for PREP® 2022 (Table 3). In only 2 out of 36 categories for PREP® 2021 and 3 out of 36 categories for PREP® 2022 did ChatGPT-3.5 outperform ChatGPT-4.0.

ChatGPT with Vision, a version of ChatGPT that can process and respond to images, was released to the public in April 2023 and has a reported knowledge cutoff date of April 2023. The PREP® 2021 and 2022 questions with images (including photos, charts, and diagrams) were tested with the new version of ChatGPT-4.0 in

Table 2. Performance of ChatGPT-3.5 and ChatGPT-4.0 on AAP Pediatrics 2022 PREP® Self-Assessment.

AAP pediatrics 2022 PREP®	ChatGPT-3.5	ChatGPT-4.0
Self-assessment questions[a]	Correct responses	Correct responses
With media (n = 47)	29 (61.7%)	40 (85.1%)
Without media (n = 200)	108 (54%)	168 (84%)
With table(s) (n = 43)	19 (44.18%)	35 (81.39%)
Without table(s) (n = 204)	118 (57.84%)	173 (84.8%)
With table(s) >2 columns (n = 5)	1 (20%)	4 (80%)
All questions (n = 247)	137 (55.46%)	208 (84.21%)

[a] Passing score 70%+, average score 75.71%, n = 6825.

November 2023 by including JPEG versions of images along with the text of each question in the ChatGPT prompt. Of the 100 questions with media entered into ChatGPT with Vision, 68 (68%) [PREP® 2021: 36/53 (67.92%), PREP® 2022: 32/47 (68.08%)] were answered correctly, 15 (15%) [PREP® 2021: 6/53 (11.32%), PREP® 2022: 9/47 (19.14%)] were answered incorrectly, and for the remaining 17 (17%) questions [PREP® 2021: 11/53 (20.75%), PREP® 2022: 6/47 (12.76%)] the chatbot declined to provide a specific answer for 13 questions, and could not interpret 3 video files and 1 audio file.

The 39 incorrect responses generated by ChatGPT-4.0 on PREP® 2022, were reviewed by the authors and were classified by perceived level of risk: minimal risk (22/39, 56.4%), moderate risk (13/39, 33.3%), or high risk (4/39, 10.3%). The incorrect answers classified as high risk included responses that: (1) missed the risk of possible congenital cytomegalovirus infection and inappropriately attributed an elevated conjugated bilirubinemia to breastmilk jaundice, (2) recommended oral antibiotic therapy instead of intravenous therapy for symptomatic lyme-associated complete heart block, (3) recommended nasogastric feedings instead of nasojejunal feedings in an intubated child at risk for aspiration, and (4) recommended lymph node biopsy as an additional, but unnecessary, diagnostic step in a case of hemophagocytic lymphohistiocytosis.

Discussion

AI-driven analysis has demonstrated impressive prowess in healthcare, especially by facilitating diagnosis of complex disease and analyzing massive data sets.[24,25] In pediatrics, studies show that AI can effectively help diagnose common childhood diseases,[26] assist in decisions related to pediatric surgery,[27] and detect child physical abuse.[28] A developing domain of AI involves the ability of powerful computers to mimic human conversation by processing massive data sets of text collected from the Internet.[3] These so-called large language models (LLM)'s are able to respond to open-ended textual queries without the need for specific training in the given task and have improved exponentially in the past few years.[29] An example of such technology is ChatGPT, an AI-powered conversational bot created by extensively refining a LLM. OpenAI's ChatGPT is a publicly-available online artificial intelligence program designed to optimize language models for dialog.[9] ChatGPT-3.5 is available publicly and free of charge. Version 4.0 of ChatGPT requires a paid subscription.[9]

Despite AI's exciting potential, there is concern that these new technologies may be used for nefarious purposes. With publicly-available versions of AI chatbots, students of all disciplines may use AI-assisted technology to cheat on academic assignments and exams. ChatGPT has performed at or above the passing level on law school exams,[30] business management courses,[31] and the United States Medical Licensing Exam (USMLE).[32-34] The results of our study suggest that ChatGPT-4.0 could be used to respond to questions on the pediatric board certification exam, which is now offered in an "open book" online format.

ChatGPT has demonstrated near expert-level medical question answering[35] and when tested on an online social media healthcare forum, ChatGPT answers were rated as significantly higher quality and more empathetic than physician responses.[36] Nonetheless, there is ample evidence to suggest that ChatGPT is not ready to replace physicians.[12,37,38] ChatGPT's ability to scan the "collective knowledge" of the Internet and produce comprehensive answers to complex medical questions is impressive on the surface, yet the "average" and/or most common answers found in the public domain are not always correct.[3] Undesirable bias is expected to be transferred, and perhaps amplified, by chatbots.[39,40] Moreover, the ethical implications of AI's deployment in healthcare must be carefully considered in order to mitigate its potential harms, particularly for the most vulnerable.[41]

Table 3. Performance of ChatGPT-3.5 and ChatGPT-4.0 on AAP Pediatrics 2021 and 2022 PREP® Self-Assessments by Pediatric Specialty.

	AAP pediatrics 2021 PREP®		AAP pediatrics 2022 PREP®	
Pediatric specialty	ChatGPT-3.5 (%)	ChatGPT-4.0 (%)	ChatGPT-3.5 (%)	ChatGPT-4.0 (%)
Adolescent Medicine and Gynecology	5/8 (62.5)	6/8 (75)	6/11 (54.5)	9/11 (81.8)
Allergic and Immunologic Disorders	2/2 (100)	2/2 (100)	5/7 (71.4)	5/7 (71.4)
Behavioral and Mental Health Issues	3/4 (75)	4/4 (100)	7/7 (100)	6/7 (85.7)
Blood and Neoplastic Disorders	2/8 (25)	6/8 (75)	3/8 (37.5)	6/8 (75)
Cardiovascular Disorders	7/10 (70)	7/10 (70)	5/8 (62.5)	5/8 (62.5)
Child Abuse and Neglect	1/1 (100)	1/1 (100)	1/1 (100)	0/1 (0)
Collagen Vascular and Other Multisystem Disorders	1/3 (33.3)	2/3 (66.7)	1/3 (33.3)	3/3 (100)
Critical Care	1/2 (50)	2/2 (100)	2/2 (100)	2/2 (100)
Disorders of Cognition, Language, and Learning	5/6 (83.3)	4/6 (66.7)	2/4 (50)	3/4 (75)
Disorders of the Eye	4/5 (80)	5/5 (100)	4/4 (100)	3/4 (75)
Ear, Nose, and Throat Disorders	7/10 (70)	8/10 (80)	10/14 (71.4)	12/14 (85.7)
Emergency Care	3/5 (60)	3/5 (60)	2/4 (50)	4/4 (100)
Endocrine Disorders	3/8 (37.5)	7/8 (87.5)	4/8 (50)	8/8 (100)
Ethics for Primary Pediatricians	2/8 (25)	3/8 (37.5)	1/5 (20)	5/5 (100)
Fetus and Newborn Infant	9/14 (64.3)	11/14 (78.6)	3/9 (33.3)	5/9 (55.6)
Fluid and Electrolyte Metabolism	3/5 (60)	5/5 (100)	4/7 (57.1)	6/7 (85.7)
Gastrointestinal Disorders	6/7 (85.7)	6/7 (85.7)	5/12 (41.7)	12/12 (100)
Genetics and Dysmorphology	4/8 (50)	8/8 (100)	3/5 (60)	5/5 (100)
Genital System Disorders	2/5 (40)	2/5 (40)	1/1 (100)	1/1 (100)
Growth and Development	2/3 (66.7)	2/3 (66.7)	3/5 (60)	3/5 (60)
Infectious Diseases	11/26 (42.3)	21/26 (80.8)	13/24 (54.2)	21/24 (87.5)
Metabolic Disorders	1/4 (25)	3/4 (75)	4/5 (80)	5/5 (100)
Musculoskeletal Disorders	4/5 (80)	4/5 (80)	1/5 (20)	3/5 (60)
Neurologic Disorders	7/10 (70)	9/10 (90)	7/13 (53.8)	10/13 (76.9)
Nutrition and Nutritional Disorders	7/12 (58.3)	9/12 (75)	3/6 (50)	5/6 (83.3)
Patient Safety and Quality Improvement	1/1 (100)	1/1 (100)	4/4 (100)	4/4 (100)
Pharmacology and Pain Management	2/4 (50)	3/4 (75)	2/5 (40)	5/5 (100)
Poisoning and Environmental Exposure to Hazardous Substances	3/7 (42.9)	6/7 (85.7)	1/2 (50)	2/2 (100)
Preventive Pediatrics	3/3 (100)	3/3 (100)	5/13 (38.5)	10/13 (76.9)
Psychosocial Issues and Child Abuse	1/2 (50)	2/2 (100)	3/6 (50)	6/6 (100)
Renal and Urologic Disorders	4/7 (57.1)	7/7 (100)	4/8 (50)	7/8 (87.5)
Research and Statistics	6/8 (75)	6/8 (75)	2/2 (100)	2/2 (100)
Respiratory Disorders	5/10 (50)	7/10 (70)	4/8 (50)	6/8 (75)
Skin Disorders	7/9 (77.8)	8/9 (88.9)	6/13 (46.2)	11/13 (84.6)
Sports Medicine and Physical Fitness	4/7 (57.1)	6/7 (85.7)	3/5 (60)	5/5 (100)
Substance Abuse	5/6 (83.3)	4/6 (66.7)	3/3 (100)	3/3 (100)
Total	143/243 (58.8)	193/243 (79.4)	137/247 (55.5)	208/247 (84.2)

In this study, both versions of ChatGPT (-3.5 and -4.0) answered complex multiple-choice pediatric board questions in a manner similar to a human test taker, with a single answer and a detailed explanation for each choice. In some responses, ChatGPT commented on its lack of information (ie, lack of access to images) but still chose an answer based on the information provided. With rare exception, a specific answer was chosen for each prompt and a plausible explanation was generated by ChatGPT, even when the answer was incorrect.

ChatGPT demonstrated some interesting abilities. The responses were delivered in a variety of styles (ie, paragraph form, bullet points, etc.) and seemed reasonably human-like. For example, one chatbot response would provide the final answer first, followed by a detailed explanation. In other responses, the explanation would build up to the final answer. Although beyond the scope of this study, it was interesting to note that ChatGPT would sometimes change the answer choice when prompted a second time with the same question/answer text and would willingly change the answer when told, "That answer is incorrect. Try again." On several occasions, ChatGPT was noted to recognize a lack of necessary information. When a media file was referenced but not available in the text-only prompt, the chatbot would respond with comments such as, "Unfortunately, I cannot directly visualize the plasma concentration curves mentioned in your question (Item Q164). However, I can provide an interpretation based on the given answer choices and the typical scenarios in which they would be applicable." On one occasion, the chatbot astutely provided the following conditional response in a case of a child with a radio-opaque foreign body in the esophagus: "Unfortunately, I cannot directly view the radiograph (Item Q118) that you mentioned. If the radiograph shows a button battery in the esophagus or other potentially harmful objects in a location that poses a risk, then the correct answer would be: A. endoscopic removal of object. If the radiograph shows a coin or another inert object in the stomach without signs of obstruction or other complications, then: B. reassurance to his mother might be appropriate." When a second explanatory prompt was entered by a human such as, "The X-ray suggests that the foreign body is a battery in the mid-esophagus," then the chatbot committed to the correct answer. ChatGPT with Vision, a version of ChatGPT that can process and respond to images, was tested separately for the questions that included images, charts, or graphs. Despite the additional information provided, performance was negatively affected when these media files were included in the prompt.

The passing score for ChatGPT-4.0 on both the 2021 and 2022 PREP® exams demonstrates an impressive ability to process complex medical questions on a pediatric board preparatory exam. Nonetheless, we found that the chatbot's explanation for incorrect answers was sometimes accompanied by erroneous logic or seemingly nonsensical information that could be misleading to caregivers of children. Several, but not all, of the chatbot responses in our study were accompanied by a disclaimer such as, "However, the final decision should be made by the medical team based on the child's specific clinical situation and needs."

Performance on a preparatory exam may not translate to actual board exam performance or to real-life knowledge. We think that it is unlikely that human pediatricians will be replaced by computers any time in the near future. The field of pediatrics is complex and requires face-to-face, hands-on interaction between patient, caregivers, and pediatrician and a deep understanding of science and human nature. The potential for AI to augment pediatric healthcare is exciting but evolving. It is likely that AI will continue to provide useful tools for healthcare in the future, but these innovations must be used with discretion and under direct human supervision.

Conclusion

Reliance on ChatGPT or similar tools to guide treatment of sick children and to provide caregiver advice could pose unacceptable risk to this vulnerable population. It is expected that there will be rapid technological advances in the emerging field of NLP that may improve the usability and safety of chatbots in clinical settings in the future, however these technologies should be used with extreme caution at this time.

Author Contributions

ML: Contributed to conception and design; Drafted the manuscript; Gave final approval; Agrees to be accountable for all aspects of work ensuring integrity and accuracy. MD: Contributed to conception and design; Contributed to analysis; Drafted the manuscript; critically revised the manuscript; Gave final approval; Agrees to be accountable for all aspects of work ensuring integrity and accuracy.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Clinical Trial Registration (if any)

N/A.

ORCID iD

Michael Davis https://orcid.org/0000-0001-5585-727X

References

1. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull*. 2021;139(1):4-15. doi:10.1093/bmb/ldab016
2. Zhou B, Yang G, Shi Z, Ma S. Natural language processing for smart healthcare. *IEEE Rev Biomed Eng*. 2024;17:4-18. doi:10.1109/RBME.2022.3210270
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8

4. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023;330(9):866-869. doi:10.1001/jama.2023.14217
5. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0
6. Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (Turing) test: survey study. *JMIR Med Educ*. 2023;9:e46939. doi:10.2196/46939
7. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Snakebite advice and counseling from artificial intelligence: an acute venomous snakebite consultation with ChatGPT. *Cureus*. 2023;15(6):e40351. doi:10.7759/cureus.40351
8. Seth I, Cox A, Xie Y, et al. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. *Aesthet Surg J*. 2023;43(10):1126-1135. doi:10.1093/asj/sjad140
9. About OpenAI. OpenAI © 2015–2024. Accessed September 1, 2023. https://openai.com/about/
10. Hu K. ChatGPT sets record for fastest-growing user base – analyst note. Reuters. Archived from the original on February 3, 2023. Retrieved August 20, 2023.
11. ChatGPT Mar 23 Version. OpenAI © 2015–2024. Accessed September 1, 2023. https://chat.openai.com/chat/
12. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Med Educ*. 2023;9:e46599. doi:10.2196/46599
13. Bhaskar S, Nurtazina A, Mittoo S, Banach M, Weissert R. Editorial: telemedicine during and beyond COVID-19. *Front Public Health*. 2021;9:662617. doi:10.3389/fpubh.2021.662617
14. Watson AG, McCoy JV, Mathew J, Gundersen DA, Eisenstein RM. Impact of physician workload on burnout in the emergency department. *Psychol Health Med*. 2019;24(4):414-428. doi:10.1080/13548506.2018.1539236
15. Grow HM, McPhillips HA, Batra M. Understanding physician burnout. *Curr Probl Pediatr Adolesc Health Care*. 2019;49(11):100656. doi:10.1016/j.cppeds.2019.100656
16. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: Scoping Review. *J Med Internet Res*. 2020;22(12):e24087. doi:10.2196/24087
17. Wong J, Foussat AC, Ting S, et al. A chatbot to engage parents of preterm and term infants on parental stress, parental sleep, and infant feeding: usability and feasibility study. *JMIR Pediatr Parent*. 2021;4(4):e30169. doi:10.2196/30169
18. Kadariya D, Venkataramanan R, Yip H, et al. kBot: knowledge-enabled personalized chatbot for asthma self-management. *Proceedings - 2018 IEEE international conference on smart computing*, 2019:138-143. doi: 10.1109/smartcomp.2019.00043
19. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. *J Intern Med*. 2018;283(6):516-529. doi:10.1111/joim.12752
20. Dash D, Thapa R, Banda J, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. 2023;*arXiv:2304.13714v3*. doi:10.48550/arXiv.2304.13714
21. Cooling C. World Health Day 2023: How AI and ChatGPT Are Revolutionizing Telemedicine and Remote Patient Care. 2023. Accessed September 17, 2023. https://www.techopedia.com/world-health-day-2023-how-ai-and-chatgpt-are-revolutionizing-telemedicine-and-remote-patient-care
22. Harskamp R, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). *Acta Cardiol*. 2024;1-9. doi:10.1080/00015385.2024.2303528
23. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. *Neurosurg*. 2023;93:1090-1098. doi:10.1227/neu.0000000000002551
24. He F, Page JH, Weinberg KR, Mishra A. The development and validation of simplified machine learning algorithms to predict prognosis of hospitalized patients with COVID-19: multicenter, retrospective study. *J Med Internet Res*. 2022;24(1):e31549. doi:10.2196/31549
25. Ren Y, Loftus TJ, Datta S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. *JAMA Netw Open*. 2022;5(5):e2211973. doi:10.1001/jamanetworkopen.2022.11973
26. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. 2019;25(3):433-438. doi:10.1038/s41591-018-0335-9
27. Xiao D, Meyers P, Upperman JS, Robinson JR. Revolutionizing healthcare with ChatGPT: an early exploration of an AI language model's impact on medicine at large and its role in pediatric surgery. *J Pediatr Surg*. 2023;58(12):2410-2415. doi:10.1016/j.jpedsurg.2023.07.008
28. Shahi N, Shahi AK, Phillips R, et al. Using deep learning and natural language processing models to detect child physical abuse. *J Pediatr Surg*. 2021;56(12):2326-2332. doi:10.1016/j.jpedsurg.2021.03.007
29. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. 2020;*arXiv:2005.14165v4*. doi:10.48550/arXiv.2005.14165
30. Choi J, Hickman K, Monahan A, Schwarcz D. ChatGPT goes to law school. *J Legal Educ*. 2023;71:387. doi:10.2139/ssrn.4335905
31. Terwiesch C. Would Chat GPT3 Get a Wharton MBA? A prediction based on its performance in the operations management course, Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, 2023. Accessed September 17, 2023.

32. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit Heal*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
33. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. doi:10.2196/45312
34. Nori H, King N, McKinney S, et al. Capabilities of GPT-4 on medical challenge problems. *arXiv:2303.13375v2*, 2023. doi:10.48550/arXiv.2303.13375
35. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *arXiv:2305.09617v1*, 2023. doi:10.48550/arXiv.2305.09617
36. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838
37. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184
38. Looi MK. Sixty seconds on . . . chatgpt. *BMJ*. 2023;380:05. doi:10.1136/bmj.p205
39. Brinker TJ, Hekler A, Enk AH, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur J Cancer*. 2019;119:11-17. doi:10.1016/j.ejca.2019.05.023
40. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. *JAMA*. 2023;330(23):2255-2257. doi:10.1001/jama.2023.22557
41. Murphy K, Di Ruggiero E, Upshur R, et al. Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics*. 2021;22(1):14. doi:10.1186/s12910-021-00577-8