Original Paper

Putting ChatGPT’s Medical Advice to the (Turing) Test: Survey Study

Oded Nov[1], PhD; Nina Singh[2], BSc; Devin Mann[2,3], MD

[1]Department of Technology Management, Tandon School of Engineering, New York University, New York, NY, United States
[2]Department of Population Health, Grossman School of Medicine, New York University, New York, NY, United States
[3]Medical Center Information Technology, Langone Health, New York University, New York, NY, United States

Corresponding Author:
Oded Nov, PhD
Department of Technology Management
Tandon School of Engineering
New York University
5 Metrotech, Brooklyn
New York, NY, 11201
United States
Phone: 1 646 207 7864
Email: onov@nyu.edu

Abstract

Background: Chatbots are being piloted to draft responses to patient questions, but patients’ ability to distinguish between provider and chatbot responses and patients’ trust in chatbots’ functions are not well established.

Objective: This study aimed to assess the feasibility of using ChatGPT (Chat Generative Pre-trained Transformer) or a similar artificial intelligence–based chatbot for patient-provider communication.

Methods: A survey study was conducted in January 2023. Ten representative, nonadministrative patient-provider interactions were extracted from the electronic health record. Patients’ questions were entered into ChatGPT with a request for the chatbot to respond using approximately the same word count as the human provider’s response. In the survey, each patient question was followed by a provider- or ChatGPT-generated response. Participants were informed that 5 responses were provider generated and 5 were chatbot generated. Participants were asked—and incentivized financially—to correctly identify the response source. Participants were also asked about their trust in chatbots’ functions in patient-provider communication, using a Likert scale from 1-5.

Results: A US-representative sample of 430 study participants aged 18 and older were recruited on Prolific, a crowdsourcing platform for academic studies. In all, 426 participants filled out the full survey. After removing participants who spent less than 3 minutes on the survey, 392 respondents remained. Overall, 53.3% (209/392) of respondents analyzed were women, and the average age was 47.1 (range 18-91) years. The correct classification of responses ranged between 49% (192/392) to 85.7% (336/392) for different questions. On average, chatbot responses were identified correctly in 65.5% (1284/1960) of the cases, and human provider responses were identified correctly in 65.1% (1276/1960) of the cases. On average, responses toward patients’ trust in chatbots’ functions were weakly positive (mean Likert score 3.4 out of 5), with lower trust as the health-related complexity of the task in the questions increased.

Conclusions: ChatGPT responses to patient questions were weakly distinguishable from provider responses. Laypeople appear to trust the use of chatbots to answer lower-risk health questions. It is important to continue studying patient-chatbot interaction as chatbots move from administrative to more clinical roles in health care.

(JMIR Med Educ 2023;9:e46939) doi: 10.2196/46939

KEYWORDS

artificial intelligence; AI; ChatGPT; large language model; patient-provider interaction; chatbot; feasibility; ethics; privacy; language model; machine learning

XSL•FO
RenderX

Introduction

Advances in large language models (LLMs) have enabled dramatic improvements in the quality of artificial intelligence (AI)–generated conversations. Recently, the launch of ChatGPT (Chat Generative Pre-trained Transformer; OpenAI) [1] has prompted a surge of interest in AI-based chatbots, both from the health care field [2,3] and the general public [4,5]. Several health care systems, including University of California San Diego Health and University of Wisconsin Health, have already announced pilots of using the underlying Generative Pre-trained Transformer (GPT) technology as a means of drafting initial responses to patient portal messages [6]. Other health care systems, including Stanford Health Care, are also preparing for pilots of GPT-drafted patient portal message responses [6].

This study assessed the feasibility of using ChatGPT or similar AI-based chatbots for answering patient portal messages directed at health care providers. ChatGPT is a chatbot created by OpenAI that is based on the LLM known as GPT [1]. At a high level, it was trained to predict the most probable next word using a large body of text data from the internet, and it was optimized to respond to user queries using reinforcement learning with human feedback on its responses to questions. Although it is generally able to generate humanlike and accurate text, LLMs such as ChatGPT have several limitations. These include biases from the underlying data (eg, social biases such as racism and sexism) [7,8], the ability to “hallucinate” information that is untrue [9], and the lack of mental models that would allow for true reasoning rather than simply probabilistic text generation (leading it to make errors in response to queries such as simple arithmetic problems) [10].

Using ChatGPT or similar AI-based chatbots to respond to patient portal messages is of interest given the recently launched pilots, the increasing burden of patient messages being delivered to providers [11], and the association between increased electronic health record (EHR) work and provider burnout [12,13]. Moreover, providers are generally not allocated time or reimbursement for answering patient messages. In an age when patients increasingly expect providers to be digitally accessible, it is likely that patient message load will continue to increase. As the technology behind AI-based chatbots matures, the time is ripe for exploring chatbots’ potential role in patient-provider communication.

Recent studies have had health care professionals judge ChatGPT’s responses to health-related questions [14-16], with findings such as 84% of answers to cardiovascular disease prevention questions being appropriate [15] and ChatGPT overall scoring higher for quality and empathy than health care providers [16]. Fewer studies have examined patient attitudes toward ChatGPT providing responses to health-related questions [17]. Here, we sought to understand how patients may perceive AI chatbot–generated responses to their questions. We reported on the ability of members of the public to distinguish between AI- and provider-generated responses to patients’ health questions. Further, we characterized participants’ trust in chatbots’ functions. Finally, we discussed the possible implications of the adoption of AI-based chatbots in patient messaging portals.

Notably, we were not trying to distinguish whether AI- or human-generated responses are a better solution for patients. Rather, we studied whether patients can tell that the response is coming from AI versus a provider and whether they trust AI, which are separate questions.

Methods

Overview

Ten representative, nonadministrative patient-provider interactions from one of the authors were extracted from the patient-provider interaction module of the EHR. All identifying details were removed, and typos in the provider’s response were fixed. Patients’ questions were entered into ChatGPT on January 19, 2023, with a request to respond using approximately the same word count as the provider’s response (see Textbox 1). Chatbot response text that recommended consultation with the patient’s health care provider was removed. The response accuracy of the human and ChatGPT responses were not evaluated to provide as close as possible to an in-the-wild experience for participants.

The 10 questions and responses were presented to a US-representative sample of 430 people aged 18 years and older who were recruited on Prolific, a crowdsourcing platform for academic studies. Participants provided written informed consent to take part in the study.

Participants were informed that 5 of the responses were written by a human provider and 5 were generated by an AI-based chatbot. For each participant, each patient question was followed by either a provider- or ChatGPT-generated response. Participants were asked to determine which responses were written by a provider and which were generated by a chatbot. The setup of 5 human responses versus 5 chatbot responses follows Fisher’s [18] seminal work on experimental design, which recommends an equal distribution of items and that participants be told in advance of the distribution. In doing so, we (1) establish a uniform prior belief in the probability associated with each advice source, (2) promote independent decision-making by participants regarding individual responses without considering other questions, and (3) avoid any influence that could sway participants’ preferences toward a specific advice source. The order of the 10 questions and answers, as well as the order of the choices presented to participants, were randomized. Participants were incentivized financially to distinguish between human and chatbot responses (US $2 baseline compensation, with up to a US $3 bonus for answering questions correctly).

Participants were then asked questions about their trust in chatbots’ use in patient-provider communication using a Likert scale from 1-5 (see the *Results* section). They were asked about their trust in chatbots to provide different types of services (logistical information, preventative care advice, diagnostic advice, and treatment advice), their trust in AI chatbots to answer health questions compared to a Google search, and their

XSL•FO
RenderX

overall trust in AI chatbots to help them make better health decisions.

With respect to distinguishability, a chi-squared test for proportions was used to determine if there was a difference in the proportion of correct identification by men versus women.

A chi-square test for goodness of fit was used to investigate whether there were variations in the proportion of correct identification across different participant age groups. Similarly, differences across age and gender in participants' response to the survey's trust questions were analyzed using ANOVA. Across all tests, results were considered significant if *P*<.05.

Textbox 1. Sample patient question (A) and answers (B and C).

A. Patient question

You indicated that based on lab tests conducted last month, my blood sugar levels were at a pre-diabetic stage. Could you provide me with steps I can take to reduce those levels, including any specific dietary recommendations. Thanks very much.

B. Human response

It comes down to weight loss with some physical activity which can help as well. How you achieve weight loss really depends on each person. Rules of thumb are to reduce carbohydrate intake both from simple sugars as well as starches like bread pasta and rice. In addition, limiting any juices or sweets helps as well. But most importantly is really increasing the amount of healthy fruit and vegetables and other high nutrient density foods and just limiting the portion sizes of the foods as well. Many people appreciate a more structured program for reducing the risk of diabetes and if you look at the federal government's National diabetes prevention program they have lots of online materials and there are often group classes at YMCA's or other locations in the community.

C. Chatbot response

To reduce your pre-diabetic blood sugar levels, the first step is to focus on making lifestyle changes. Specifically, you should focus on your diet and exercise. Eating a diet high in whole, unprocessed foods such as fruits, vegetables and whole grains, while limiting your intake of added sugars, saturated fats, and processed foods will help. Incorporating more protein-rich foods such as lean meats, fish, and beans and eating small, frequent meals throughout the day rather than a few large ones can also help. In terms of exercise, try to aim for at least 30 minutes of moderate physical activity per day, such as brisk walking, cycling or swimming. It is also very important to maintain a healthy weight, as being overweight or obese can increase your risk of developing type 2 diabetes. I recommend speaking with a registered dietitian or nutritionist who can help you create a personalized eating plan.

Ethical Considerations

This study was certified and filed as a Quality Improvement study per NYU Langone Health's Quality Improvement self-certification protocol. As a Quality Improvement study, institutional review board approval is not needed.

Results

Overall, 426 participants filled out the full survey. After removing participants who spent less than 3 minutes on the survey, 392 survey responses were used in the analysis. Of the 392 respondents, 53.3% (n=209) were women, and the average age was 47.1 (SD 16.0) years.

The responses to patient questions varied widely in the participants' ability to identify whether they were written by a human or chatbot, ranging between 49% (192/392) to 85.7% (336/392) for different questions. Each participant received a score between 0-10 based on the number of responses they identified correctly (Multimedia Appendix 1). On average, chatbot responses were identified correctly in 65.5% (1284/1960) of the cases, and human provider responses were identified correctly in 65.1% (1276/1960) of the cases. No substantial differences were found in response distinguishability or trust by demographic characteristics.

On average, patients trusted chatbots (Table 1), yet trust was lower as the health-related complexity of the task in the questions increased. Logistical questions (eg, scheduling appointments and insurance questions) had the highest trust rating (mean Likert score 3.94, SD 0.92), followed by preventative care (eg, vaccines and cancer screenings; mean Likert score 3.52, SD 1.10). Diagnostic and treatment advice had the lowest trust ratings (mean Likert scores 2.90, SD 1.14 and 2.89, SD 1.12, respectively). No significant correlations were found between trust in health chatbots and demographics or the ability to correctly identify chatbot versus human responses (all *P*>.05).

XSL•FO
RenderX

Table 1. Health chatbot trust questions and responses.

Question	Patients with Likert response ≥4 (n=392), n (%)	Likert response (range 1-5), mean (SD)
I could trust answers from a health chatbot about logistical questions (such as scheduling appointments, insurance questions, medication requests).	312 (79.6)	3.94 (0.92)
I could trust a chatbot to provide advice about preventative care, such as vaccines, or cancer screenings.	248 (63.3)	3.52 (1.10)
I could trust a chatbot to provide diagnostic advice about symptoms.	152 (38.8)	2.90 (1.14)
I could trust a chatbot to provide treatment advice.	150 (38.3)	2.89 (1.12)
AI[a] chatbots can be a more trustworthy alternative to Google to answer my health questions.	232 (59.2)	3.56 (1.02)
Health chatbots could help me make better decisions.	236 (60.2)	3.49 (0.91)

[a] AI: artificial intelligence.

Discussion

Principal Findings

Patients increasingly expect *consumer-grade* health care experiences that mirror their experiences with the rest of their digital life. They want omnichannel and interactive communication, frictionless access to care, and personalized education. The resulting overwhelming volume of patient portal messages highlights an opportunity for chatbots to assist health care providers, one that is already being acted upon by several large health care systems [6]. Early research on provider perception of these chatbot-generated responses has revealed high degrees of appropriateness [15] and has even revealed higher quality and empathy ratings than human-generated responses [16]. However, whether patients view chatbot communication as comparable to communication with human providers requires empirical investigation [19-21].

In this study of a US-representative sample, compared to the benchmark of 50% representing random distinguishability and 100% representing perfect distinguishability, laypeople found responses from an AI-based chatbot to be weakly distinguishable from those from a human provider. Notably, there was very little difference between the distinguishability rate of human versus chatbot responses (65.5 vs 65.1%).

It is likely that in the near future, the level of indistinguishability we found will represent a lower bound of performance, as chatbots trained on medical data specifically, or prompted with medical queries, will likely be less distinguishable [14]. Another possible future development is for chatbots to reach a superhuman level as seen in other medical domains [22]. The emerging group of vendors designing optimized prompt libraries for health systems is likely to further improve chatbots' performance on health-related questions (eg, DocsGPT [23]). It is important to note that products based on LLMs, such as ChatGPT, merely provide text that resembles good medical advice, and it is only with the addition of medical knowledge that useful health care provider–level advice could be provided.

Respondents' trust in chatbots' functions were mildly positive. Notably, there was a lower level of trust in chatbots as the medical complexity of the task increased, with the highest acceptance for administrative tasks such as scheduling appointments and the lowest acceptance for treatment advice. This is broadly consistent with prior studies [17,24]. In particular, a recent study of user intentions to use ChatGPT for self-diagnosis found that higher performance expectancy and positive risk-reward appraisals were associated with improved perception of decision-making outcomes [17]. This improved perception in turn positively impacted participant intentions to use ChatGPT for self-diagnosis (78% of the 476 participants indicated that they were willing to do so) [17].

Our study suggests that participants are overall willing to receive health advice from a chatbot (especially for low-risk topics) and are only weakly able to distinguish between ChatGPT- versus human-generated responses. Based on our findings, identifying appropriate scenarios for deploying chatbots within health care systems is an important next step. Although chatbots are widely used in health care administrative tasks (eg, scheduling), optimal clinical use cases are still emerging [25]. Chatbots have been developed and deployed for highly specialized clinical scenarios such as symptom triage and postchemotherapy education [26]. More generalized chatbots that are similar to ChatGPT represent a new opportunity to use chatbots in support of more common chronic disease management for conditions such as hypertension, diabetes, and asthma. Health care providers' work may be transformed by using the products of generative AI (such as chatbots' output) as raw material to construct patient-provider interaction, including advice, the explanation of test results, the discussion of side effects, and many other types of interactions that currently require a human health care provider. For example, chatbots could be deployed with home blood pressure monitoring to support patient questions about treatment plans, medication titrations, and potential side effects [27].

Potential deployment models include chatbots that directly interact with patients (eg, through patient portals) or serve as clinician assistants, generating draft text or transforming clinician documentation into more patient-friendly versions. For health care providers' work, this would lead to a shift in focus from the *creation* of health care advice to the *curation* of advice in response to patient messages. Of note, it is critical that providers stay alert when curating rather than simply accepting the models' answers. ChatGPT and other LLMs have known limitations including producing incorrect or biased

XSL•FO
RenderX

answers [1,7,8], and automation bias (ie, humans favoring suggestions from automated decision-making systems over their own judgment) is a key concern to watch for [28]. Liability will also be a key concern that will necessitate careful curation of chatbot responses [29].

The appropriateness of each deployment model likely depends on the clinical complexity and severity of the condition. Higher-risk or -complexity clinical interactions could use chatbots to generate drafts for clinician editing or approval and lower-risk situations may allow for direct patient-chatbot interaction. Alternatively, it may be useful to have chatbots classify questions into administrative versus health questions, replying directly to administrative questions and drafting responses for provider approval to health questions. The role and impact of the disclosure of origination (human vs chatbot) also needs further exploration, especially with regards to ethics, effectiveness, and implications for the patient-provider relationship.

Although our study addressed new questions with state-of-the-art technology, it has some key limitations. First, ChatGPT was not trained on medical data and could be inferior to medically trained chatbots such as Med-PaLM [14]. Second, there was no specialized prompting of ChatGPT (eg, to be empathetic), which can help responses sound more human and could potentially increase patients' willingness to accept AI chatbot–generated responses [30]. Third, it is possible that individual style (of both the human provider and chatbot) can impact distinguishability, although the responses presented were for the most part short and impersonal. Fourth, it is possible that there were biases in the web-based survey since the participants were given the prior knowledge that 5 answers were human generated and 5 answers were chatbot generated. Fifth, this study was conducted using ChatGPT in January 2023 (based on GPT-3.5; OpenAI) [1]. Since then, more advanced underlying GPT models such as GPT-4 have been released, and further development has integrated GPT with EHRs and adapted it to medical tasks such as responding to patient portal messages [6]. Finally, this study used only 10 real-world questions with human responses from 1 provider. Further studies incorporating larger numbers of real-world questions and responses are warranted.

In addition, future research may explore how to prompt chatbots to provide an optimal patient experience [30], investigate if there are types of questions that chatbots are better at answering than others, and explore if patients feel more trusting if there is clinician review before chatbots respond. Continued studies investigating how model responses differ by patient demographics (eg, gender and race) [1,7,8] will be critical to ensure the recognition and mitigation of model biases and work toward equitable responses. Research to mitigate risks of AI chatbot–generated responses, including the potential for patient harm caused by incorrect answers; cybersecurity vulnerabilities [31]; and environmental, social, and financial risks [32] should also be further explored.

Conclusion

Overall, our study shows that ChatGPT responses to patient questions were weakly distinguishable from provider responses. Furthermore, laypeople trusted chatbots to answer lower-risk health questions. It is important to continue studying how patients interact (objectively and emotionally) with chatbots as they become a commodity and move from administrative to more clinical roles in health care.

Acknowledgments

The authors receive financial support from the US National Science Foundation (awards 1928614 and 2129076) for the submitted work. The funding source had no further role in this study. We used the generative artificial intelligence tool ChatGPT (Chat Generative Pre-trained Transformer) by OpenAI [1] to draft the chatbot responses for the research survey.

Data Availability

The anonymized data generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ON, NS, and DM designed the study, selected the content for the experiment, and wrote the first draft of the manuscript. ON and NS implemented the experiment and performed the statistical analysis. All authors vouch for the data, analyses, and interpretations; critically reviewed and contributed to the preparation of the manuscript; and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Distribution of correct responses.
[PNG File , 93 KB-Multimedia Appendix 1]

References

1. Introducing ChatGPT. OpenAI. 2022. URL: https://openai.com/blog/chatgpt [accessed 2023-07-03]

XSL•FO
RenderX

2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239 [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
3. Biswas SS. Role of Chat GPT in public health. Ann Biomed Eng 2023 May 15;51(5):868-869 [doi: 10.1007/s10439-023-03172-7] [Medline: 36920578]
4. Bruni F. Will ChatGPT make me irrelevant? The New York Times. 2022 Dec 15. URL: https://www.nytimes.com/2022/12/15/opinion/chatgpt-artificial-intelligence.html [accessed 2023-07-03]
5. Stern J. ChatGPT wrote my AP English essay—and I passed. The Wall Street Journal. 2022 Dec 21. URL: https://www.wsj.com/articles/chatgpt-wrote-my-ap-english-essayand-i-passed-11671628256 [accessed 2023-07-03]
6. Turner BEW. Epic, Microsoft bring GPT-4 to EHRs. Modern Healthcare. 2023 Apr 17. URL: https://www.modernhealthcare.com/digital-health/himss-2023-epic-microsoft-bring-openais-gpt-4-ehrs [accessed 2023-07-03]
7. Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. Nat Mach Intell 2021 Jun 17;3(6):461-463 [doi: 10.1038/s42256-021-00359-2]
8. Bolukbasi T, Chang KW, Zou J, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. 2016 Dec 5 Presented at: NIPS'16: 30th International Conference on Neural Information Processing Systems; December 5-10, 2016; Barcelona, Spain p. 4356-4364 URL: https://dl.acm.org/doi/10.5555/3157382.3157584
9. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput. Surv 2023 Mar 03;55(12):1-38 [doi: 10.1145/3571730]
10. Huang J, Chang KCC. Towards reasoning in large language models: a survey. arXiv Preprint posted online on May 26, 2023. [doi: 10.48550/arXiv.2212.10403]
11. Holmgren AJ, Downing NL, Tang M, Sharp C, Longhurst C, Huckman RS. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. J Am Med Inform Assoc 2022 Jan 29;29(3):453-460 [FREE Full text] [doi: 10.1093/jamia/ocab268] [Medline: 34888680]
12. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
13. Marmor R, Clay B, Millen M, Savides T, Longhurst C. The impact of physician EHR usage on patient satisfaction. Appl Clin Inform 2018 Jan 03;9(1):11-14 [FREE Full text] [doi: 10.1055/s-0037-1620263] [Medline: 29298451]
14. Singhal K, Azizi S, Tu T, Mahdavi S, Wei J, Chung HW, et al. Large language models encode clinical knowledge. arXiv Preprint posted online on December 29, 2022. [doi: 10.48550/arXiv.2212.13138]
15. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844 [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
16. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
17. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
18. Fisher RA. Design of experiments. BMJ 1936 Mar 14;1(3923):554-554 [doi: 10.1136/bmj.1.3923.554-a]
19. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021 Sep;3(9):e599-e611 [FREE Full text] [doi: 10.1016/S2589-7500(21)00132-1] [Medline: 34446266]
20. Chang IC, Shih YS, Kuo KM. Why would you use medical chatbots? interview and survey. Int J Med Inform 2022 Sep;165:104827 [doi: 10.1016/j.ijmedinf.2022.104827] [Medline: 35797921]
21. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, Talks J, Denniston AK, Kelly CJ, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023 Jan 10;25:e39742 [FREE Full text] [doi: 10.2196/39742] [Medline: 36626192]
22. Attia ZI, Harmon DM, Dugan J, Manka L, Lopez-Jimenez F, Lerman A, et al. Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. Nat Med 2022 Dec 14;28(12):2497-2503 [FREE Full text] [doi: 10.1038/s41591-022-02053-1] [Medline: 36376461]
23. DocsGPT. Doximity. 2023. URL: https://www.doximity.com/docs-gpt [accessed 2023-07-03]
24. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019 Aug 21;5:2055207619871808 [FREE Full text] [doi: 10.1177/2055207619871808] [Medline: 31467682]
25. Montenegro JLZ, da Costa CA, da Rosa Righi R. Survey of conversational agents in health. Expert Syst Appl 2019 Sep;129:56-67 [doi: 10.1016/j.eswa.2019.03.054]
26. Winn AN, Somai M, Fergestrom N, Crotty BH. Association of use of online symptom checkers with patients' plans for seeking care. JAMA Netw Open 2019 Dec 02;2(12):e1918561 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.18561] [Medline: 31880791]

XSL•FO
RenderX

27. Mann DM, Lawrence K. Reimagining connected care in the era of digital medicine. JMIR mHealth uHealth 2022 Apr 15;10(4):e34483 [FREE Full text] [doi: 10.2196/34483] [Medline: 35436238]
28. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 2023 May;307(4):e222176 [doi: 10.1148/radiol.222176] [Medline: 37129490]
29. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. JAMA Health Forum 2023 May 05;4(5):e231938 [FREE Full text] [doi: 10.1001/jamahealthforum.2023.1938] [Medline: 37200013]
30. Sebastian G, George A, Jackson G. Persuading patients using rhetoric to improve artificial intelligence adoption: experimental study. J Med Internet Res 2023 Mar 13;25:e41430 [FREE Full text] [doi: 10.2196/41430] [Medline: 36912869]
31. Sebastian G. Do ChatGPT and other AI chatbots pose a cybersecurity risk?: an exploratory study. International Journal of Security and Privacy in Pervasive Computing\ 2023;15(1):1-11 [doi: 10.4018/ijsppc.320225]
32. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? 2021 Mar Presented at: FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency; March 3-10, 2021; Virtual event, Canada p. 610-623 [doi: 10.1145/3442188.3445922]

Abbreviations

AI: artificial intelligence
ChatGPT: Chat Generative Pre-trained Transformer
EHR: electronic health record
GPT: Generative Pre-trained Transformer
LLM: large language model

Edited by K Venkatesh, MN Kamel Boulos; submitted 02.03.23; peer-reviewed by C Yan, J Ropero, G Sebastian, N Mungoli; comments to author 04.05.23; revised version received 26.05.23; accepted 14.06.23; published 10.07.23

Please cite as:
Nov O, Singh N, Mann D
Putting ChatGPT's Medical Advice to the (Turing) Test: Survey Study
JMIR Med Educ 2023;9:e46939
URL: https://mededu.jmir.org/2023/1/e46939
doi: 10.2196/46939
PMID: 37428540

©Oded Nov, Nina Singh, Devin Mann. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.07.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX