Journal of Multidisciplinary Healthcare

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Are Different Versions of ChatGPT's Ability Comparable to the Clinical Diagnosis Presented in Case Reports? A Descriptive Study

Jingfang Chen[1–3], Linlin Liu[3], Shujin Ruan[3], Mengjun Li[3], Chengliang Yin[1]

[1]Faculty of Medicine, Macau University of Science and Technology, Macau, People's Republic of China; [2]Department of Research and Teaching, the Third People's Hospital of Shenzhen, Shenzhen, People's Republic of China; [3]Hengyang Medical School, School of Nursing, University of South China, Hengyang, People's Republic of China

Correspondence: Chengliang Yin, Faculty of Medicine, Macau University of Science and Technology, Macau, People's Republic of China, Email chengliangyin@163.com

Objective: ChatGPT, an advanced language model developed by OpenAI, holds the opportunity to bring about a transformation in the processing of clinical decision-making within the realm of medicine. Despite the growing popularity of research related on ChatGPT, there is a paucity of research assessing its appropriateness for clinical decision support. Our study delved into ChatGPT's ability to respond in accordance with the diagnoses found in case reports, with the intention of serving as a reference for clinical decision-making.

Methods: We included 147 case reports from the Chinese Medical Association Journal Database that generated primary and secondary diagnoses covering various diseases. Each question was independently posed three times to both GPT-3.5 and GPT-4.0, respectively. The results were analyzed regarding ChatGPT's mean scores and accuracy types.

Results: GPT-4.0 displayed moderate accuracy in primary diagnoses. With the increasing number of input, a corresponding enhancement in the accuracy of ChatGPT's outputs became evident. Notably, autoimmune diseases comprised the largest proportion of case reports, and the mean score for primary diagnosis exhibited statistically significant differences in autoimmune diseases.

Conclusion: Our finding suggested that the potential practicality in utilizing ChatGPT for clinical decision-making. To enhance the accuracy of ChatGPT, it is necessary to integrate it with the existing electronic health record system in the future.

Keywords: ChatGPT, artificial intelligence, clinical decision support systems, case reports

Introduction

ChatGPT leverages large language models (LLMs) developed by OpenAI to generate text that closely resembles human writing, advancing scientific inquiry. It is modeled on the Generative Pretrained Transformer (GPT), which undergo extensive training on large text corpora. This training involves creating question-and-answer tasks with a user-friendly interface. Since its introduction, researchers have explored the practical value of clinical medicine, such as drug development, image recognition, data analysis, improving medical reports, providing medical information, conducting literature reviews, and personalizing medicine.[1–4] A study evaluated the performance of GPT-4.0 in radiology, with a 54% overall accuracy, which showed its potential as a valuable tool in radiology.[5] Furthermore, ChatGPT facilitates healthcare professionals grasping the key insights in their respective domains,[6] and can be employed as a tool to assess clinical skills.[7] These applications hold groundbreaking importance, given that 60% of Americans actively seek medical information online.[8] The AI chatbot assistant generated high-quality and empathetic responses to online inquiries, which may decrease the demand for doctor visits and hospital consultations.[9] Additionally, the output of high-quality and empathetic responses may have a positive impact on patients' health-related behaviors and enhance clinical outcomes.

Clinical decision-making is often influenced by physicians' clinical thinking and patients' complex condition, which may lead to cognitive bias.[10] AI models could aid the clinical decision support systems (CDSS). AI-based CDSS

Received: 23 September 2023
Accepted: 26 November 2023
Published: 6 December 2023

provides medical information and recommendations for diagnosis and treatment to clinicians. These systems leverage extensive medical knowledge, employing algorithms to emulate the clinical diagnosis and treatment thought of clinicians. ChatGPT could assist with clinical decision support, which achieves optimization of clinical decision-making. Previous studies reported that GPT-3 have been used in clinical settings, such as ophthalmology and dementia prediction.[11–13] Rao[14] found that ChatGPT achieved an overall accuracy rate of 71.7% across all 36 clinical cases. Stokel[15] revealed that ChatGPT could answer some open-ended medical queries almost as well as the average physician. Furthermore, ChatGPT achieves moderate accuracy in radiologic decision making.[16] These findings underscore the potential for ChatGPT to assist clinical decision-making and optimize CDSS.

The case report is a common genre in medical articles, involving the documentation and description of individual case with the aim of providing firsthand medical information on aspects, such as disease presentation, mechanisms, diagnosis, and treatment. Case reports are widely recognized and valued by clinicians and encompass a variety of topics, including rare diseases, adverse drug reactions, and disease-specific clinical manifestations. Patients exhibit diverse manifestations and etiologies, complicating the diagnostic process. Clinicians must make decisions based on intricate information.[17] Among downloaded case reports, autoimmune diseases being much more common. Approximately 5% of the worldwide population is affected by autoimmune diseases.[18] Autoimmune diseases represent a family of at least 80 conditions that share common pathogenesis.[19,20] It has been reported that autoimmune diseases are a leading cause of death among young and middle-aged women.[21] Due to their complex pathogenic mechanisms, the clinical treatment of autoimmune diseases is challenging and poses a heavy burden on patients. In the case of rare diseases, clinicians may also face difficulty in making timely and accurate diagnosis.

In the context of rare diseases or specific disease subtypes, clinicians frequently encounter diagnostic challenges stemming from their limited experience and the scarcity of pertinent reference. ChatGPT can offer a potential list of disease diagnoses based on the clinical data provided by physicians, aiding in preliminary screening and differential diagnosis, thereby reducing the risk of misdiagnosis or underdiagnosis. Moreover, it could be linked to medical databases and guideline websites, providing authoritative guidance and treatment recommendations. This comprehensive information equips healthcare professionals to gain a deeper understanding of the condition and manage disease more effectively. As of now, there is no official guideline outlining the standards for utilizing LLMs like ChatGPT in academic medicine.[22] Consequently, the objective of this study is to assess the accuracy of ChatGPT's responses aligned with evidence-based case reports.

Methods

Study Design

In this descriptive study, we aimed to assess the capacity of ChatGPT to deliver accurate responses based on evidence drawn from case reports. Furthermore, we conducted an evaluation of the accuracy of GPT-3.5 and GPT-4.0 in diagnostic types. All case reports were gathered from the Chinese Medical Association Journal Database (https://www.yiigle.com/Paper/Search?type=Case&q=%E7%96%BE%E7%97%85andsearchType=pt), which serves as a prominent position as the foremost repository, including numerous journals in medicine.

The inclusion of case reports were as follows: (1) publication between January 1, 2013 and June 1, 2023; (2) a definitive diagnosis of the illness; (3) main text include: summary of medical history, signs and symptoms, diagnostic approach, treatment and clinical regression. The exclusion were: (1) duplicate publications; (2) unable to access full text; (3) Chinese medicine therapy, nursing or disease management-related case reports. We used the following search terms: Title="disease" AND "Article type"=Case report AND Publication time=[2013-01-01 TO 2023-06-01]. Around 285 case reports showed up. Finally, our analysis included 147 case reports. All included case reports were published in Chinese. As this study did not involve human subjects, it did not necessitate approval from an institutional review board or the procurement of informed consent.

Study Data

Disease-related data from 147 case reports were independently input into GPT-3.5 and GPT-4.0 on three times, and the outputs were recorded to assess the reproducibility of ChatGPT. Two senior clinical experts independently evaluated

these outputs and assigned scores to each response. The responses were categorized as either "consistent", "partially consistent", or "inconsistent". The overall score for each case was assessed on a scale ranging from 0 to 3, based on the number of "consistent" responses. Disparate viewpoints were harmonized through a panel discourse.

Statistical Analysis

Data analysis was conducted using SPSS 25.0 software. Quantitative variables were reported as means±SD (standard deviations), while categorical variables were described by indicating the absolute number of cases within each distinct group. The paired *t*-test was employed to assess normally distributed data. Differences between groups were evaluated for statistical significance using the chi-square test, with significant differences of $P < 0.05$.

Results

Baseline Characteristics

A list of scores from 147 case reports is provided in Supplementary Table 1. The downloaded case reports pertained to various systems, including the immune system, central nervous system, lymphatic system, respiratory system, endocrine system, cardiovascular system, skeletal muscle system, and others. Among these, autoimmune disorders accounted for 37% of the case reports, followed by central nervous system (20%), respiratory system (11%), and lymphatic system (10%). When considering individual diseases, IgG4-related diseases comprised 33% of the case reports, with lymphoproliferative disorders (10%), neuromyelitis optica-spectrum disorders (8%), and chronic obstructive pulmonary disease (5%).

Comparison of Mean Scores in Diagnostic Types

Table 1 displayed a comparison of mean scores between GPT-3.5 and GPT-4.0. A significant difference was observed in the mean scores of the primary diagnosis between the GPT-3.5 and GPT-4.0 (P=0.013). GPT-4.0 displayed a higher score in the diagnoses than GPT-3.5.

Comparison of Mean Scores in Different Systems

Figure 1 showed the mean scores of different systems between GPT-3.5 and GPT-4.0. ChatGPT generally exhibited higher accuracy in the prevalent diseases than rare diseases. In terms of the primary diagnoses, a significant difference was observed in the mean score of the autoimmune system between the GPT-3.5 (0.84±1.11) and GPT-4.0 (1.07±1.29). There were no significant differences in central nervous system, lymphatic system, and respiratory system. In central nervous system, the mean score for GPT-3.5 and GPT-4.0 were (0.87±1.20) and (1.03±1.28), respectively. In lymphatic system, the mean score for GPT-3.5 and GPT-4.0 were (0.50±0.91) and (0.86±1.25), respectively. In respiratory system, the mean score for GPT-3.5 and GPT-4.0 were (1.69±1.40) and (1.75±1.35), respectively.

Regarding secondary diagnoses, there were no significant difference in autoimmune system, central nervous system, lymphatic system, and respiratory system. In autoimmune system, the mean score for GPT-3.5 and GPT-4.0 were (0.50 ±0.39) and (0.54±0.36), respectively. In central nervous system, the mean score for GPT-3.5 and GPT-4.0 were (0.44 ±0.46) and (0.50±0.50), respectively. In lymphatic system, the mean score for GPT-3.5 and GPT-4.0 were (0.59±0.34) and (0.63±0.40), respectively. In respiratory system, the mean score for GPT-3.5 and GPT-4.0 were (0.71±0.32) and (0.60 ±0.37), respectively.

Table 1 Comparison of Mean Scores in Diagnostic Types

Characteristic	GPT-3.5 (Means ± SD)	GPT-4.0 (Means ± SD)	P-value
Primary diagnosis	1.01±1.26	1.26±1.33	0.013
Secondary diagnosis	0.55±0.38	0.61±0.41	0.285

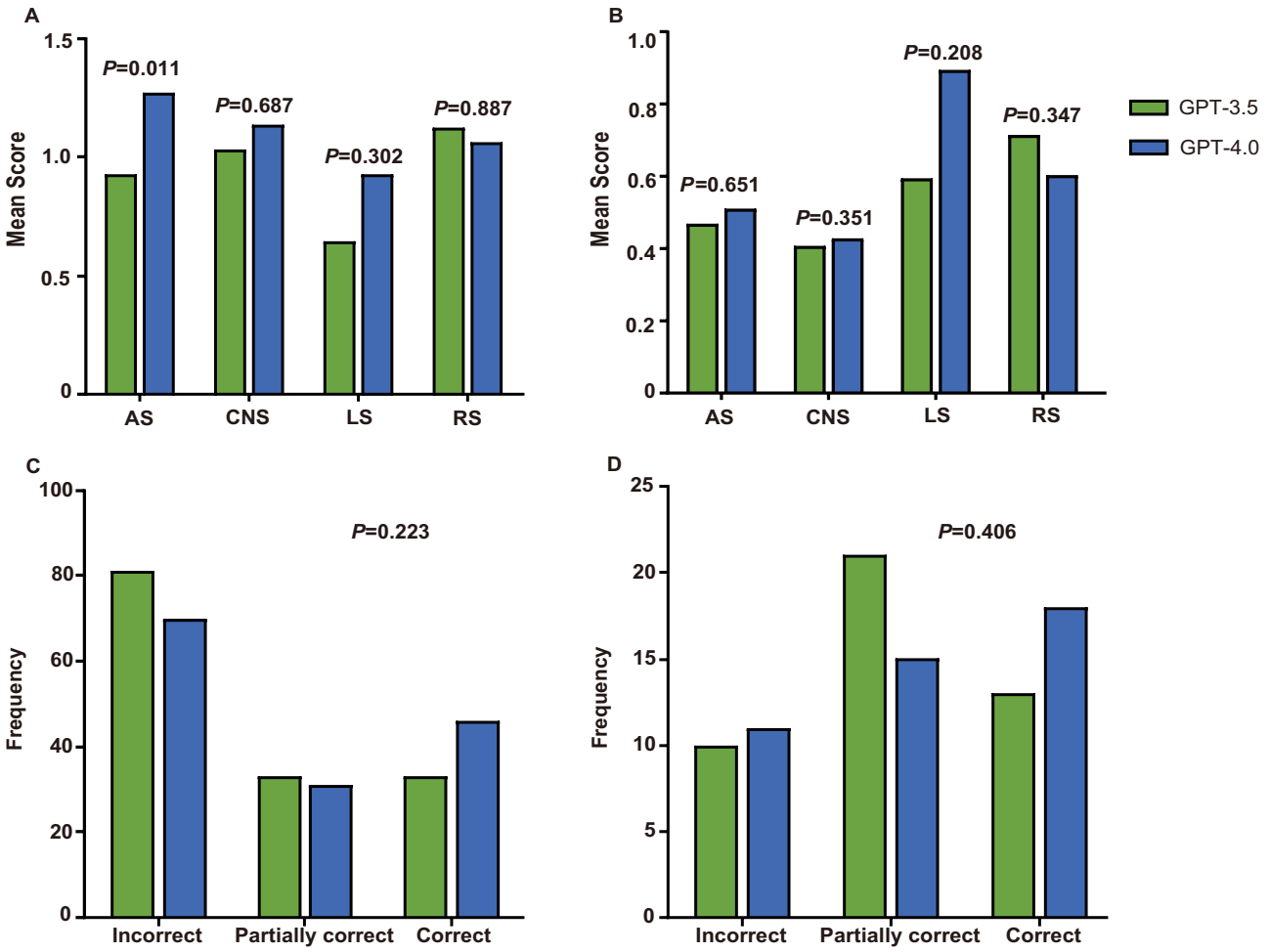

Figure 1 Comparison of mean scores and diagnostic types according to GPT-3.5 and GPT-4.0. (**A**) The mean scores of primary diagnoses in different systems, (**B**) The mean scores of secondary diagnoses in different systems, (**C**) The accuracy of primary diagnoses, (**D**) The accuracy of secondary diagnoses.
Abbreviations: AS, autoimmune system; CNS, central nervous system; LS, lymphatic system; RS, respiratory system.

Comparison of Accuracy in Diagnostic Types

Despite no statistically difference in the comparison of diagnostic types between GPT-3.5 and GPT-4.0, GPT-4.0 generally excelled than GPT-3.5. In the primary diagnosis, GPT-3.5 showed 22.45% of outputs completely consistent, 22.45% partially consistent, and 55.10% inconsistent. GPT-4.0 demonstrated 31.29% of outputs completely consistent, 21.09% partially consistent, and 47.62% inconsistent. In the secondary diagnosis, GPT-3.5 had 29.55% of outputs completely consistent, 47.73% partially consistent, and 22.72% inconsistent. GPT-4.0 displayed 40.91% of outputs completely consistent, 34.09% partially consistent, and 25.00% inconsistent (Figure 1).

On three separate input times, we submitted identical query to both GPT-3.5 and GPT-4.0, respectively. As illustrated in Tables 2 and 3, the accuracy of the output in the third time slightly surpassed that of the first output.

Discussion

In this study, we revealed that GPT-4.0 displayed moderate accuracy in the primary diagnosis, indicating its potential usefulness as an adjunct for clinical diagnosis. ChatGPT could generate potential diagnoses based on patient symptoms, medical history, and laboratory test results.[23] Despite no statistically significant difference between the GPT-3.5 and GPT-4.0 groups, the results suggested a potential refinement in GPT-4.0's comprehension and processing of medical information. GPT-4.0 manifests profound grasp of input context, consequently enhancing the accuracy of the generated text.

Table 2 Comparison of the Diagnostic Types in Primary Diagnoses

Categories	Diagnostic Types (%)			
	Incorrect	Partially Correct	Correct	Total
GPT-3.5 Q1	99 (67.3%)	1 (0.7%)	47 (32.0%)	147
GPT-3.5 Q2	96 (65.3%)	1 (0.7%)	50 (34.0%)	147
GPT-3.5 Q3	94 (63.9%)	1 (0.7%)	52 (35.4%)	147
GPT-4.0 Q1	89 (60.5%)	0 (0.0%)	58 (39.5%)	147
GPT-4.0 Q2	87 (59.2%)	0 (0.0%)	60 (40.8%)	147
GPT-4.0 Q3	80 (54.4%)	0 (0.0%)	67 (45.6%)	147
Total	545 (61.8%)	3 (0.3%)	334 (37.9%)	882

Table 3 Comparison of the Diagnostic Types in Secondary Diagnoses

Categories	Diagnostic Types (%)			
	Incorrect	Partially Correct	Correct	Total
GPT-3.5 Q1	17 (38.6%)	16 (36.4%)	11 (25.0%)	44
GPT-3.5 Q2	19 (43.2%)	14 (31.8%)	11 (25.0%)	44
GPT-3.5 Q3	16 (36.4%)	16 (36.4%)	12 (27.3%)	44
GPT-4.0 Q1	10 (22.7%)	16 (36.4%)	18 (40.9%)	44
GPT-4.0 Q2	10 (22.7%)	16 (36.4%)	18 (40.9%)	44
GPT-4.0 Q3	10 (22.7%)	16 (36.4%)	18 (40.9%)	44
Total	82 (31.1%)	94 (35.6%)	88 (33.3%)	264

In terms of the output process, GPT-4.0 provided disease diagnoses accompanied by comprehensive explanations. And GPT-4.0 generally answered quicker than GPT-3.5, possibly due to the hardware enhancements and algorithmic improvements. Duey[24] found that GPT-3.5 tend to cite nonexistent references, while GPT-4.0 was more conservative in its responses. However, ChatGPT occasionally provided illogical or incorrect output. The references generated by ChatGPT have not undergone validation.[15] Clinicians should be cautious and verify the safety and efficacy of information provided by ChatGPT.

Among the downloaded case reports, a significant proportion pertained to rare diseases. The low prevalence of rare diseases presents formidable challenges in accurately diagnosing and providing care for affected patients.[25] Rare diseases are frequently responsible for chronic illness, disability, and premature death. Despite enduring extensive and costly evaluations at different hospitals, patients frequently experience underdiagnoses or misdiagnoses,[26,27] exacerbating the burden on their quality of life and their families, as well as imposing a substantial societal burden. ChatGPT could offer a potential diagnostic list when a patient presents with ambiguous and complex symptoms, which may reduce unnecessary costs.

The diagnostic accuracy for prevalent diseases was higher than rare diseases, a discrepancy attributed to potential bias within the training dataset. It may be associated with the clinical data and resources available for common diseases compared to rare diseases. Elevating the model's efficacy in handling a diverse spectrum of rare cases relies on providing more the patient-specific content in future updates.

ChatGPT holds potential for the progression of medicine, yet concurrently, it presents the potential limitations. The accuracy of the generated text is contingent upon the model's training, potentially resulting in misinformation or misleading interpretations. Besides, ChatGPT's training data may not comprehensively encompass the more recent advancements, which typically on data up to September 2021. Hallucinations, omissions, and errors have been documented through the utilization of ChatGPT.[28–30] These factors may contribute to contradictory or erroneous responses. When considering the utilization of ChatGPT-based research, it is essential to account for ethical concerns

and data governance, especially for privacy regulations and data security. Compliance with relevant legal standards, the patient data utilized for training ChatGPT is essential to undergo anonymization to maintain privacy. Researchers must ensure secure collection, storage, and usage of biomedical data. Alex[31] concluded that the largest barriers to the implementation of ChatGPT in clinical practice are deficits in situational awareness, inference, and consistency. Patients have very complex medical, social, and psychiatric backgrounds, which often requires a rigorous logical reasoning process, personal experience and even intuition. Only real doctors can play the role of comprehensive judgment and clinical treatment, ChatGPT is used as a tool to support clinical practice. Maintaining the accuracy and reliability of the content generated by ChatGPT is paramount, the repercussions of inaccuracy in medicine can be devastating, especially for patients or trainees lacking the prerequisite knowledge or experiential foundation. Hence, researchers should verify the information provided by LLMs with current and reputable medical sources,[32] ensuring the safety and efficacy of clinical diagnoses.

The limitation of our study is the sample size, we only included Chinese case reports, contributing to a limitation in the breadth of results. Another limitation is the absence of comparisons with other search engines like Google Bard and Bing. Previous studies have suggested that ChatGPT outperformed Google Bard.[33–35]

Conclusion

ChatGPT serves as an important ally in the medical realm, offering healthcare professionals a valuable opportunity for collaboration. Subsequent research could focus on integrating ChatGPT with existing electronic health record. The integrated model holds the capability to combine the specialized knowledge of medical experts with the capabilities of ChatGPT, thereby elevating the quality of responses and facilitating personalized clinical decision-making. To enhance the credibility in clinical decision-making, it is necessary to improve the transparency of ChatGPT. This can be achieved by advancing AI algorithms, enabling healthcare professionals to comprehend the rationale underpinning the model's recommendations. Additionally, the incorporation of medical experts group's feedback mechanisms emerges as a pivotal facet of the ongoing optimization process, empowering healthcare professionals to rectify and fine-tune the model's recommendations. We hope that there will be more refined medical scenarios of the application of the product to come out, really make doctors and patients benefit.

Funding

This work was supported by Shenzhen High-level Hospital Construction Fund (G2022006).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in Medical Research: current Status and Future Directions. *J Multidiscip Healthc*. 2023;16:1513–1520. doi:10.2147/JMDH.S413470
2. Mann DL. Artificial Intelligence Discusses the Role of Artificial Intelligence in Translational Medicine: a JACC: basic to Translational Science Interview With ChatGPT. *JACC Basic Transl Sci*. 2023;8(2):221–223. doi:10.1016/j.jacbts.2023.01.001
3. Blanco-González A, Cabezón A, Seco-González A, et al. The Role of AI in Drug Discovery: challenges, Opportunities, and Strategies. *Pharmaceuticals*. 2023;16(6):891. doi:10.3390/ph16060891
4. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care*. 2023;27(1):75. doi:10.1186/s13054-023-04380-2
5. Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. *Radiology*. 2023;308(1):e231040. doi:10.1148/radiol.231040
6. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595. doi:10.3389/frai.2023.1169595
7. Han JW, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. *BMC Med Educ*. 2022;22(1):830. doi:10.1186/s12909-022-03898-3
8. Gulati R, Nawaz M, Pyrsopoulos NT. Health literacy and liver disease. *Clin Liver Dis*. 2018;11(2):48–51. doi:10.1002/cld.690
9. 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. doi:10.1001/jamainternmed.2023.1838
10. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. *J Med Internet Res*. 2023;25:e48568. doi:10.2196/48568

11. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. *JMIR Med Inform*. 2022;10(2):e32875. doi:10.2196/32875
12. Nath S, Marie A, Ellershaw S, Korot E, Keane PA. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. *Br J Ophthalmol*. 2022;106(7):889–892. doi:10.1136/bjophthalmol-2022-321141
13. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. *PLOS Digit Health*. 2022;1(12):e0000168. doi:10.1371/journal.pdig.0000168
14. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: development and Usability Study. *J Med Internet Res*. 2023;25:e48659. doi:10.2196/48659
15. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature*. 2023;614(7947):214–216. doi:10.1038/d41586-023-00340-6
16. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J Am Coll Radiol*. 2023;20(10):990–997. doi:10.1016/j.jacr.2023.05.003
17. Roosan D, Samore M, Jones M, Livnat Y, Clutter J Big-Data Based Decision-Support Systems to Improve Clinicians' Cognition. In: 2016 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2016:285–288. doi:10.1109/ICHI.2016.39.
18. Bieber K, Hundt JE, Yu X, et al. Autoimmune pre-disease. *Autoimmun Rev*. 2023;22(2):103236. doi:10.1016/j.autrev.2022.103236
19. Rose NR. Prediction and Prevention of Autoimmune Disease in the 21st Century: a Review and Preview. *Am J Epidemiol*. 2016;183(5):403–406. doi:10.1093/aje/kwv292
20. Mané-Damas M, Hoffmann C, Zong S, et al. Autoimmunity in psychotic disorders. Where we stand, challenges and opportunities. *Autoimmun Rev*. 2019;18(9):102348. doi:10.1016/j.autrev.2019.102348
21. Walsh SJ, Rau LM. Autoimmune diseases: a leading cause of death among young and middle-aged women in the United States. *Am J Public Health*. 2000;90(9):1463–1466. doi:10.2105/ajph.90.9.1463
22. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. *J Pediatr Urol*. 2023. doi:10.1016/j.jpurol.2023.05.018
23. Ferdush J, Begum M, Hossain ST. ChatGPT and Clinical Decision Support: scope, Application, and Limitations. *Ann Biomed Eng*. 2023. doi:10.1007/s10439-023-03329-4
24. Duey AH, Nietsch KS, Zaidat B, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. *Spine J*. 2023;23(11):1684–1691. doi:10.1016/j.spinee.2023.07.015
25. Lee J, Liu C, Kim J, et al. Deep learning for rare disease: a scoping review. *J Biomed Inform*. 2022;135:104227. doi:10.1016/j.jbi.2022.104227
26. Molster C, Urwin D, Di Pietro L, et al. Survey of healthcare experiences of Australian adults living with rare diseases. *Orphanet J Rare Dis*. 2016;11(1):30. doi:10.1186/s13023-016-0409-z
27. Marwaha S, Knowles JW, Ashley EA. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med*. 2022;14(1):23. doi:10.1186/s13073-022-01026-w
28. Fernandes AC, Souto MEVC. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med*. 2023;388(25):2399–2400. doi:10.1056/NEJMc2305286
29. Wagner MW, Ertl-Wagner BB. Accuracy of Information and References Using ChatGPT-3 for Retrieval of Clinical Radiological Information. *Can Assoc Radiol J*. 2023;8465371231171125. doi:10.1177/08465371231171125
30. McGowan A, Gui Y, Dobbs M, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res*. 2023;326:115334. doi:10.1016/j.psychres.2023.115334
31. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? *Lancet Infect Dis*. 2023;23(4):405–406. doi:10.1016/S1473-3099(23)00113-5
32. Thapa S, Adhikari S. ChatGPT, Bard, and Large Language Models for Biomedical Research: opportunities and Pitfalls. *Ann Biomed Eng*. 2023;51(12):2647–2651. doi:10.1007/s10439-023-03284-0
33. Patil NS, Huang RS, Der Pol CB V, Larocque N. Comparative Performance of ChatGPT and Bard in a Text-Based Radiology Knowledge Assessment. *Can Assoc Radiol J*. 2023;08465371231193716. doi:10.1177/08465371231193716
34. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: chatGPT vs Google Bard. *Radiology*. 2023;307(5):e230922. doi:10.1148/radiol.230922
35. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery*. 2023. doi:10.1227/neu.0000000000002551