Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/54/20241225

Application, investigation and prediction of ChatGpt/GPT-4 for clinical cases in medical field

Xingyu Zhao

Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S10 2TN, England, UK

xzhao82sheffield@ldy.edu.rs

Abstract. The integration of Artificial Intelligence (AI) and medical treatment not only makes the clinical diagnosis more accurate, but also makes the patient's rehabilitation more systematic and professional, especially after the advent of the Large Language Model(LLM) in the past 2 years. This paper discusses 3 clinical cases of 2 kinds of LLMs: ChatGPT (GPT-3.5) and GPT-4 in Physical Medicine and Rehabilitation (PM&R), and shows their powerful analytical reasoning ability. In the first experiment, ChatGPT and the leading professional doctors in the industry were asked to classify the emergency records of ophthalmology during the 10-year period, infer the severity of each patient's illness and determine the nursing requirements. In the later experiment, GPT-4, an upgraded version of GPT-3.5, delayed the diagnosis of medical history data of patients aged 65 and over, to study the clinical diagnosis opinions and systematic treatment scheme of GPT-4 as a "professional doctor". ChatGPT and GPT-4 participated in the examination with 12 categories of neurosurgery medical fields, which was shown in the last experiment, aiming at studying their medical professional level and discussing their clinical reliability and effectiveness, as well as LLMs' ability of reasoning questions step by step. The experimental results show that these 2 kinds of large language models have professional and powerful ability to analyze actual projects, and their performance even far exceeds that of professional clinicians. At the same time, the existing defects of the models and their more applications in the medical field in the future are prospected.

Keywords: Large Language Model, Physical Medicine and Rehabilitation, ChatGPT, GPT-4, Medical Treatment.

1. Introduction

For a long time from the 1950s, the concept of Artificial Intelligence (AI) did not appear frequently in people's lives due to technical obstacles. However, with the emergence of current computing infrastructure, big data and deep learning algorithms, people began to explore and study in this field. With AI algorithms like machine learning and deep neural networks now outperforming clinicians in certain scenarios, the outlook is becoming increasingly clear. AI is poised to assume a significant role across various medical care domains, encompassing diagnosis, prognosis, and patient management [1, 2]. Currently, AI has outstanding performance in all walks of life. The same is true in the medical industry. They analyze complex medical data and make use of the potential of meaningful relationship with the data set, which can be used for diagnosis, treatment and prediction in many clinical scenarios. The application maturity of AI technology has been explored in almost every medical field [3].

The advent of OpenAI's ChatGPT has facilitated its use in an expanding array of fields, including the medical profession, where it is being employed to tackle challenging issues. Globally, there is a shortage of rehabilitation medical professionals. In 2009, China had 16,286 doctors, 12,062 nurses and 13,715 therapists, dedicated to PM&R rehabilitation. However, by 2018, the figures had only increased to 38,260 doctors and 15,514 nurses. Projections suggest that within the next decade, the medical demand for rehabilitation doctors, therapists and nurses is expected to surge to about 71,530, 155,365 and 58,721 respectively. Japan and the United States are also grappling with similar challenges, as the aging population is driving an increased demand for rehabilitation doctors in these countries [4]. The appearance of ChatGPT is undoubtedly a solution to the problem both the shortage of medical doctors and the clinical load. Isaac Kohane currently works in the Department of Biomedical Informatics of Harvard Medical School and is a doctor and chairman. By testing GPT, he thinks that reducing doctors' tedious written work and making time to study patients' pathology is the most obvious advantage of AI at present. At present, doctors spend more time dealing with hospital paperwork and have no time to study medical skills and stay with patients, which has an impact on doctors' psychological and academic ability.

ChatGPT/GPT-4 have a human-computer interactive chat interface, which can collect patients' complaints and symptoms information with high-efficiency, and at the same time, quickly organize and construct the trapped information to ensure detailed and precise records. This function is very helpful for identifying correlative information and possible hazard factors, and provides useful suggestions for other similar research [5-7]. By comprehensively analyzing and summarizing patient data, ChatGPT and GPT-4 can support health care professionals to collect patients' medical history more comprehensively, simplify the process of collecting patients' information, and better understand patients' medical background. In addition, ChatGPT and GPT-4 give assistance to the diagnosis process by analyzing patients' symptoms, bodily signs and other medical data. They can make a preliminary assessment of urgency and severity, help health care professionals prioritize patients according to the information provided, and provide initial guidance to decide the suitable tier of rehabilitation or necessary intervention. These abilities help medical workers to allocate health care resources efficiently [8].

2. Method

2.1. The Classification and Diagnosis of Ophthalmic Emergency Patients' Tiers by ChatGPT

This experiment determined the medical records of all adults in the ophthalmic Emergency Department (ED) during the 10 years (2012 to 2023) at the University of California, San Francisco (UCSF) [9]. The urgency degree of the Emergency Severity Index (ESI) was recorded, from the highest level to the lowest level: Immediate, Emergent, Urgent, Less Urgent, Non-Urgent, and the homologous ED doctor records generated during the visit. Regular expressions in deep learning are used to extract the features of patients' clinical history from the patient data set of clinical text corpus. 10,000 pairs of medical samples were extracted. Through the application programming interface of Microsoft, ChatGPT was used to infer the sample results of each pair of ED, so as to judge which patient has higher sensitivity (the determination of clinical sensitivity is a measure of the severity of the patient's disease and the tier of rehabilitation care required, and it is one of the basic factors of medical analyzing in emergency medicine). In contrast, doctors manually classified 500 pairs of balanced samples to compare the classification performance of ChatGPT [4].

2.2. The Use of Gpt-4 to Extensive Investigationsand Delayed Diagnosis

The data of this study is based on the diagnosis delay of patients aged 65 and over for more than 1 month, and the diagnosis data from the medical department of Queen Mary Hospital in 2022. This time, GPT-4 will be compared with the clinical inference of the clinical history. The medical history data covers all the complete records from the patient's admission to a week before the final diagnosis. The study has been approved by the Institutional Review Committee of the University of Hong Kong in China and the Kowloon West Cluster Hospital Authority. All clinical case guidelines will be observed and all patients will agree before starting the experiment. During the experiment, all the data will be input into GPT-4 in chronological order. The system does not contain clear diagnostic information, and any judgment will be diagnosed by doctors and GPT-4. The purpose of this study is to collect and compare the reaction and diagnosis opinions of GPT-4 and clinicians on these medical histories [1, 3].

2.3. Achievements of 2 LLMs in Neurosurgery Written Test

The reliability and effectiveness of ChatGPT/GPT-4 have not been evaluated in detail by developers so far, and no related data sets or evaluation reports have been published. Therefore, Self-Assessment Neurosurgery Exams (SANS) and American Board of Neurological Surgery (ABNS)are adopted to evaluate the performance of these 2 large language models. A question bank containing 500 questions was established, and multiple-choice questions were designed in strict accordance with the examination format, all of which appeared in the form of multiple-choice questions. In this process, the answer options of each question are copied and the answer options are retained. It should also be noted that this study still uses questions with images. Although ChatGPT only supports the input of texts, the multifunctional function of GPT-4 has not been made public so far. However, the input of data is still text data instead of any image data [7, 10].

12 categories have been classified by the author according to the problems, all collected through the performance of neurosurgery interns, but the performance of these problems at the personal level has not been reported. RA and OYT [1, 8], two authors, also classify problems as 2 kinds of problem solving methods based on manual evaluation as follows.

First-order question: it can be defined by some questions based on simple facts, such as the dependence of a patient on a drug, or choosing a most likely clinical diagnosis illustration as a patient's diagnosis book.

Higher-order questions: questions that involve additional intermediate steps, such as the identification of diagnosis, but will increase the evaluation or analysis tasks compared with the first-order questions, making the answers more professional and three-dimensional. After conducting a brief clinical diagnosis, it is essential to guarantee the accuracy of the outcome. Moreover, it's crucial to analyze potential optimal management steps or identify other clinical characteristics that are likely to contribute to the diagnosis.

The standardized examination of medicine currently abides by this set of classification rules. The classification of questions is random, and the answer to any question is unknown before the exam. This experiment uses linear regression method to evaluate the accuracy score of classification [2, 3, 7]

3. Results and Discussion

3.1. The Analysis of Results

3.1.1. The Classification Results of ChatGPT for Ophthalmic Emergency Patients

In this paper, 10,000 balanced samples of patients were selected from 251,401 adult emergency cases, in which patients' visual acuity scores with different ESI were recorded. This study only used the clinical record information recorded by the first ED doctor. ChatGPT can correctly infer 8354/10000 patients with high vision (accuracy = 0.84, F1 score = 0.83). The accuracy of the model is as high as 98% when distinguishing "Immediate" and "Less-urgent" or "Non-urgent" patients. As anticipated, the extent of disparity in patients' visual acuity scores directly correlates with the proficiency of the large language model's judgment [4, 9]. Among 500 sub-samples manually classified, the F1 score of ChatGPT is 0.85, which is similar to that before. This study is ChatGPT's evaluation of real clinical cases, and its ability to stratify patients is very powerful. ChatGPT can accurately identify patients

DOI: 10.54254/2755-2721/54/20241225

with high vision. After extracting information from patient records, by comparing the evaluation level of ChatGPT and doctors, it is found that they are comparable and feasible for medical assistance diagnosis.

3.1.2. Case Analysis of GPT-4 for Patients Over 65 Years Old

Through the case analysis of 6 patients aged 65 years and older (4 males and 2 females), the judgment accuracy of Isabel DDx Companion, clinician and GPT-4 were 0, 2 (33.3%) and 4 (66.7%) patients. If the differential diagnosis is included, the accuracy rate of GPT-4 was 83.3%, clinicians 50.0% and Isabel DDx Companio 33.3% [1, 6]. As follows, GPT-4 has been developed to have great medical application in elderly patients without definite clinical diagnosis, with data supported by demographic and clinical information. GPT-4 can provide expert advice and alert clinicians to standardize their diagnostic capacity, enhance their diagnostic confidence and reduce clinician stress, which is of potential value in poor countries lacking specialist care.

3.1.3. Comparison of Medical Examination Results Between ChatGPT and GPT-4

By comparing the results of SANS and ABNS self-assessment test, the correct rates of ChatGPT (ChatGPT) and GPT-4 were 73.4% and 83.4%. It can be found that both the users of the question bank and the 2 large language models passed the exam (the passing threshold is 69%). Additionally, GPT-4 outperformed its predecessors, surpassing their performance levels. In this experiment encompassing 12 distinct categories, GPT-4 emerged as the most formidable performer, consistently outshining the user's performance across all categories [2, 10]. In addition, in neuroradiology questions, the accuracy of users' answers is higher than ChatGPT and GPT-4, and their accuracy is 73.2%. On the tumor issue, the accuracy of GPT-4 is higher than the other two, both of which are 92.5%.

3.2. Limitation and Future Prospect

Through the analysis results of the above real clinical cases, it is concluded that GPT's analytical ability in the medical field can be at the expert level. Although this technology has shown strong diagnostic and analytical ability at present, it is still in the preliminary exploratory stage, and it may take time to reach the technical landing of the project. Because GPT still has some limitations, it will be misunderstood and ambiguous in the real-world environment. The focus of disease pathogen cannot be determined by GPT. Moreover, GPT does not recommend using gallium scanning to find tumors [4]. Secondly, some investigations may be inappropriate, for example, a biopsy of the temporal artery without the typical symptoms of giant cell arteritis. Therefore, although GPT is a powerful big prediction model, the problems in specific fields still need to be analyzed in detail and cannot be mechanically copied. The text case analysis in the medical field still needs to be updated and optimized, and needs to be strengthened in the future [7].

In addition, there is another problem that needs to be solved urgently, which is not a technical problem, but a GPT moral standard. Taking medical information as the training data of a new AI model, the privacy and security of a large number of users have not been guaranteed and verified, and the legislation has not been improved [1, 3, 7]. To address this challenge, the manner in which OpenAI gathers and utilizes data sources to train its expansive language model is currently under examination by data regulators worldwide. Simultaneously, efforts are underway to establish clearer legal frameworks for addressing issues stemming from the utilization of its text corpus system. When companies and organizations need to process people's personal information, the European General Data Protection (GDPR) laws requires them to provide legitimate reasons for these transactions, so as to understand the purpose of collecting and processing these information and reduce the risk of information disclosure and information abuse of residents.

4. Conclusion

This paper expounds the performance of 2 kinds of LLM, ChatGPT/GPT-4, in actual medical cases, both the correct rate of clinical diagnosis of GPT and their passing rate in complex professional medical knowledge have shown extraordinary and amazing performance. Between them, the performance of GPT-4 will be superior, because of its larger problem words length and the ability to solve higher-order problems. These rapidly developing large language models are the great breakthrough and wisdom crystallization of AI system, and they keep abreast of the potential applications in clinical medicine and the latest information. ChatGPT/GPT-4 has begun to be of great help to clinicians and has made outstanding contributions to PM&R. They not only collect the patient's medical history, but also evaluate the patient's recovery. They can undoubtedly promote the development and innovation of AI + Medicine. In the future, a large number of AI-related research will be involved, and the LLM based on AI research will make a contribution in this field.

References

[1] Shea Y F Lee C M Y Ip W C T Luk D W A & Wong S S W 2023 Use of GPT-4 to Analyze Medical Records of Patients With Extensive Investigations and Delayed Diagnosis JAMA Network Open 6(8) e2325000-e2325000
[2] Magnasco A Bacchini G Cappello A La Milia V Brezzi B Messa P & Locatelli F 2004 Clinical validation of glucose pump test (GPT) compared with ultrasound dilution technology in arteriovenous graft surveillance Nephrology Dialysis Transplantation 19(7) 1835-1841
[3] Wang C Ong J Wang C Ong H Cheng R & Ong D 2023 Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation Annals of Biomedical Engineering, 1-4
[4] Liu Z Yu X Zhang L Wu Z Cao C Dai H ... & Li X 2023 Deid-gpt: Zero-shot medical text de-identification by gpt-4 arXiv preprint arXiv:2303 11032
[5] Lee P Bubeck S & Petro J 2023 Benefits limits and risks of GPT-4 as an AI chatbot for medicine New England Journal of Medicine 388(13) 1233-1239
[6] Egli A 2023 ChatGPT GPT-4 and other large language models - the next revolution for clinical microbiology? Clinical Infectious Diseases, ciad407
[7] Takagi S Watari T Erabi, A & Sakaguchi K 2023 Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study JMIR Medical Education 9(1) e48002
[8] Nori H King N McKinney S M Carignan, D., & Horvitz E 2023 Capabilities of gpt-4 on medical challenge problems arXiv preprint arXiv:2303 13375
[9] Williams C Y Zack T Miao B Y Sushil M Wang M & Butte A J 2023 Assessing clinical acuity in the Emergency Department using the GPT-3.5 Artificial Intelligence Model medRxiv 2023-08
[10] Ali R Tang O Y Connolly I D Sullivan P L Z Shin J H Fridley J S ... & Telfeian A E 2022 Performance of ChatGPT and GPT-4 on neurosurgery written board examinations Neurosurgery 10-1227