ORIGINAL ARTICLE

FROM TEXT TO DIAGNOSE: CHATGPT'S EFFICACY IN MEDICAL DECISION-MAKING

DOI: 10.36740/WLek202311101

Yaroslav Mykhalko, Pavlo Kish, Yelyzaveta Rubtsova, Oleksandr Kutsyn, Valentyna Koval
UZHHOROD NATIONAL UNIVERSITY, UZHHOROD, UKRAINE

ABSTRACT

The aim: Evaluate the diagnostic capabilities of the ChatGPT in the field of medical diagnosis.

Materials and methods: We utilized 50 clinical cases, employing Large Language Model ChatGPT-3.5. The experiment had three phases, each with a new chat setup. In the initial phase, ChatGPT received detailed clinical case descriptions, guided by a "Persona Pattern" prompt. In the second phase, cases with diagnostic errors were addressed by providing potential diagnoses for ChatGPT to choose from. The final phase assessed artificial intelligence's ability to mimic a medical practitioner's diagnostic process, with prompts limiting initial information to symptoms and history.

Results: In the initial phase, ChatGPT showed a 66.00% diagnostic accuracy, surpassing physicians by nearly 50%. Notably, in 11 cases requiring image interpretation, ChatGPT struggled initially but achieved a correct diagnosis for four without added interpretations. In the second phase, ChatGPT demonstrated a remarkable 70.59% diagnostic accuracy, while physicians averaged 41.47%. Furthermore, the overall accuracy of Large Language Model in first and second phases together was 90.00%. In the third phase emulating real doctor decision-making, ChatGPT achieved a 46.00% success rate.

Conclusions: Our research underscores ChatGPT's strong potential in clinical medicine as a diagnostic tool, especially in structured scenarios. It emphasizes the need for supplementary data and the complexity of medical diagnosis. This contributes valuable insights to AI-driven clinical diagnostics, with a nod to the importance of prompt engineering techniques in ChatGPT's interaction with doctors.

KEY WORDS: artificial intelligence, large language models, ChatGPT, clinical decision support, diagnose

Wiad Lek. 2023;76(11):2345-2350

INTRODUCTION

Artificial intelligence (AI) is defined as "the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings" [1]. AI involves creating computer systems or software that can mimic human thinking processes, such as learning from experience, reasoning, problem-solving, and decision-making.

AI systems are designed to analyze and process large amounts of data, extract meaningful patterns, and make predictions or decisions based on that analysis. It enables tasks like speech recognition, natural language understanding, image and video analysis, recommendation systems, and autonomous decision-making.

Large language models (LLM) represent a specific category of AI models that emulate and comprehend human-like text. Built upon deep learning methodologies, these models employ neural networks characterized by multiple layers and a multitude of parameters. The primary objective underlying LLM is to attain an understanding of the structure, syntax, semantics, and contextual nuances of natural language. This comprehension equips them to generate responses that are not only coherent but also contextually relevant. Consequently, they acquire the ability to generate responses covering an expansive array of subjects [2]. The contemporary landscape witnesses an upsurge in the utilization and refinement of LLMs, attributed to the accessibility of extensive datasets and the evolution of AI technologies. This progression has resulted in the significant enhancement of these models' capabilities, enabling them to yield text that closely resembles human-generated content. Furthermore, these models have demonstrated exceptional performance across a spectrum of natural language processing tasks, underscoring their multifaceted potential [3].

In order to be useful for medical application LLMs undergo training on medical data through various methodologies, enhancing their applicability in the medical domain. One fundamental approach involves provisioning pertinent training data, encompassing electronic health records, medical literature, and clinical data. By immersing LLMs in this diverse dataset, they cultivate specialized knowledge tailored to distinct medical disciplines [4]. A pivotal step in the training process entails pre-training LLMs on expansive text datasets which equips LLMs with an understanding of language structures and patterns. Consequently, they can be adeptly repurposed across a multitude of domains and tasks, a versatility that extends to medical applications [5]. Following pre-training, LLMs undergo fine-tuning on medical data, facilitating adaptation to specific medical functions. For instance, LLMs can be fine-tuned for tasks like clinical decision support or scientific writing assistance. This involves refining the model's capabilities through exposure to a smaller dataset of medical text, thereby enhancing its performance on targeted medical tasks [6,7].

AI has made significant strides within the medical field, reshaping various aspects of healthcare and paving the way for improved patient outcomes and more efficient healthcare processes. One prominent application lies in Biomedical Information Extraction, where AI is harnessed to dissect biomedical texts and extract structured data such as named entities and semantic relationships [8]. Similarly, AI's transformative impact is evident in Drug Discovery and Development, where it plays a pivotal role in tasks like peptide synthesis, virtual screening, toxicity prediction, and drug monitoring. This AI-driven approach not only slashes time consumption and production costs but also introduces efficiencies that address the inefficiencies inherent in traditional drug design methods [9]. The integration of AI into Medical Diagnosis and Treatment is particularly noteworthy. Across various medical disciplines, machine learning algorithms are employed to evaluate radiological images, pathology slides, and electronic medical records [10]. Technologies based on AI and large language models in particular are increasingly penetrating into various areas of the medical industry such as anesthesiology, dentistry, radiation medicine, ophthalmology, cardiology and many others [11-15]. There were attempts to evaluate the performance of LLMs on medical examination.

Usage of AI in supporting clinical decision-making is of great interest as it holds the promise of revolutionizing the healthcare landscape by harnessing the capabilities of artificial intelligence to augment and inform the decision-making processes of medical professionals [16-18].

THE AIM

Evaluate the diagnostic capabilities of the ChatGPT in the field of medical diagnosis

MATERIALS AND METHODS

A selection of 50 clinical cases from the Medscape Case Challenge series, spanning the period between February and July 2023, was utilized for analysis. As the LLM ChatGPT-3.5 was used.

The experiment was structured into three phases. New chat was created for every phase of the study. In the initial phase, comprehensive clinical case descriptions were presented to ChatGPT, encompassing patient complaints, the history of the present illness, past medical records, data derived from physical examinations, outcomes of laboratory tests, and results from imaging studies. At this phase our methodology incorporated a prompt engineering technique referred to as "Persona Pattern". This approach was implemented to guide ChatGPT's responses and to encourage it to simulate the analytical process and decision-making of a proficient medical practitioner when diagnosing the presented clinical cases.

Subsequently, the cases in which the LLM made diagnostic errors were included in the second phase. For such cases, the LLM was proposed a roster of potential diagnoses extracted from the presented cases. We used the same prompt formulation as in the previous phase, ChatGPT was then prompted to choose the most suitable diagnosis from the provided list.

The third and final phase of the experiment focused on assessing ChatGPT's ability to emulate the decision-making process of a medical practitioner in a patient-doctor interaction. The structured prompt was designed to guide ChatGPT in adopting a professional medical perspective and mimicking the diagnostic reasoning of a skilled medical practitioner, while also allowing the model to access supplementary diagnostic information when required. In this context, each clinical case's initial information was restricted to encompass only the description of symptoms, patient complaints, medical history, and physical examination data. Within this constrained framework, ChatGPT was tasked with determining the optimal diagnosis and proposing a set of supplementary investigative methods necessary for precise elucidation. In cases when the required information was available in the clinical case, it was provided to the LLM. Conversely, if such information was absent, the LLM was apprised that the data was unavailable for consideration. This research methodology facilitated a comprehensive evaluation of ChatGPT's diagnostic capacities within a simulated clinical context.

During each phase, a comparative analysis was conducted to assess the accuracy of responses provided by ChatGPT in contrast to medical professionals who resolved the same clinical cases on the Medscape platform. This comparative evaluation allowed us to gauge the alignment between ChatGPT's diagnostic performance and the practicing physicians.

Statistical data are presented as M±SD.

RESULTS

During the initial phase, ChatGPT demonstrated the overall accurate diagnostic capability in 66.00% of cases, whereas the accuracy of physicians on the Medscape platform averaged at 44.82±18.54% which was almost 50% lower than LLM's result.

Of particular significance is a subgroup of 11 clinical cases where the results of supplementary diagnostic methods, including ECG, CT, MRI, and blood smear images, were presented without their corresponding interpretations. The assumption was that physicians have to analyze these images by themselves, upon which the accuracy of the ultimate diagnosis relied heavily. In these instances, the clinical case descriptions were initially provided to ChatGPT without including the interpretation of the aforementioned diagnostic findings. If ChatGPT produced an inaccurate diagnosis, the subsequent step involved offering it the interpretations of these diagnostic studies, performed by the authors of the present study. Notably, across this subset, ChatGPT erred in diagnosis for 3 cases. However, for 4 cases, the correct diagnosis was subsequently achieved after presenting ChatGPT with the additional interpretations of the diagnostic studies. Impressively, in the remaining 4 cases, ChatGPT accurately diagnosed the cases without requiring supplementary interpretations of the diagnostic findings. This subset underscores the intricate interplay between diagnostic expertise and the interpretation of visual diagnostic data, a domain in which ChatGPT demonstrated varying degrees of proficiency.

Among the subset of clinical cases in which ChatGPT accurately identified the diagnosis, real healthcare practitioners achieved an average diagnostic accuracy of 46.55±17.00%. For the clinical cases where ChatGPT did not return correct diagnosis, the average accuracy rate achieved by actual medical practitioners was lower than in the previous subset (41.47±21.38%) but the difference was not statistically significant ($p>0.05$).

In the second phase, a subset of 17 clinical cases was chosen. ChatGPT was presented with comprehensive clinical case descriptions, as in the preceding phase, along with an additional roster of potential diagnoses given in original descriptions on Medscape. Its objective was to choose the most suitable option from the provided variants. Remarkably, the LLM demonstrated a correct decision-making rate of 70.59% within this context. This achievement was nearly twice as high as the average percentage of accurate responses recorded among doctors, which stood at 41.47%.

Among the clinical cases in which ChatGPT correctly identified the diagnosis, the average accuracy rate for real doctors was 45.50±21.94%. However, in instances where ChatGPT's diagnosis was not accurate, the average accuracy of human doctors' responses was 31.80±18.42% ($p>0.05$).

The third phase, meticulously designed to emulate the decision-making process of a real healthcare worker, witnessed ChatGPT achieving accurate disease diagnoses in 23 out of 50 cases, yielding a success rate of 46.00%. In situations where ChatGPT provided accurate diagnoses, the average percentage of correct responses from doctors, who were presented with comprehensive patient data and potential diagnoses to choose from, was very close to ChatGPT's performance at 45.65±18.51%. Moreover, in cases where the LLM, ChatGPT, faced challenges in correctly diagnosing the disease, medical professionals attained an average accuracy rate of 44.11±18.90%.

Within this phase, it is pertinent to underscore distinct subsets of cases that yielded noteworthy observations. In 8 cases, ChatGPT demonstrated a high level of diagnostic acumen by accurately diagnosing diseases without requesting any supplementary information, achieving a correct diagnosis in 5 out of these 8 instances. In 12 cases, the initial diagnostic output of the LLM was erroneous. Although the diagnosis has been changed upon the provision of additional information, it remained inaccurate. In 11 cases, ChatGPT's initial diagnoses were incorrect. However, upon the acquisition of the requested supplemental information, the diagnoses were revised to correct ones. Notably, in 18 cases, the LLM's preliminary diagnostic decisions remained consistent even after the inclusion of the requested supplementary information. Lastly, it's worth noting that in a singular case, an initially accurate diagnosis was altered to an incorrect one following the incorporation of the requested supplementary information.

DISCUSSION

Artificial intelligence is increasingly penetrating into all spheres of human activity and there is no doubt that this process will only intensify over time. Until recently, access to this technology was the prerogative of a limited number of people who used it to solve specific scientific or industrial problems. However, opening access to this technology to the general public has become a real revolution. In this context, large language models deserve special attention, which allow generating texts based on queries entered by the user. Trained on a large amount of data, it can serve as a good assistant in daily activities. However, far from always, the obtained results coincide with expectations and correspond to reality. This is especially important to consider when trying to use this technology in professional activities, especially in the field of medicine. Nevertheless, this phenomenon should not be completely removed or ignored either. In this situation, it is critical to find a balance and a certain edge, which can only be done empirically. Therefore, it is important to conduct researches and evaluate the possibilities of AI in solving specific practical issues. Such an attempt was made in this study. To do this, clinical cases offered by Medscape platform in the Case Challenge series were used. These cases often cover scenarios that are not commonly encountered by most clinicians but nonetheless occur in actual clinical practice. In this way it gives medical professionals a unique opportunity to test their knowledge, diagnostic and treatment skills in a wide variety of medical fields. The clinical cases are well described and contain all the information necessary to answer the questions posed. In the majority of cases medical practitioners should diagnose the disease based on the given description and choose the correct option from the proposed 4 to seven variants. This approach facilitates the decision-making process for doctors, as they are not required to navigate through the entirety of possible conditions. Instead, their diagnostic considerations are limited to the predefined options, significantly simplifying the task of making an accurate diagnosis. Nevertheless, in this relatively simple situation the average percentage of correct diagnoses made by real doctors in 50 clinical cases chosen for this study was lower than the 50% (44.82±18.54%). This underscores the intricate nature of medical diagnosis and highlights the challenges that healthcare professionals encounter in arriving at correct diagnoses, even when provided with a set of diagnostic options. On the other hand, in the first phase of the current study ChatGPT, being in worse condition, compared to doctors as it was not provided with the variants of answers to choose from, showed almost 50% higher accuracy in making diagnosis. This outcome underscores the unique capabilities of ChatGPT in processing and interpreting clinical information, even in scenarios where it lacks the structured diagnostic choices available to human doctors and highlights the comparative performance of ChatGPT and human doctors across the various diagnostic scenarios.

In the second phase the LLM was actually in the same situation as the doctors, as it was given all the information available to doctors in selected 17 clinical cases. This resulted in further increasement of its productivity. These findings underscore the improved performance of ChatGPT when aided by a list of potential diagnoses and provide valuable insights into its diagnostic capabilities. Furthermore, if to count the overall accuracy achieved by LLM in the first and second phases together it gives an overwhelming 90% of correct diagnoses.

In the third phase, by restricting initial information to encompass only patient complaints, medical history, and physical examination data, and by providing ChatGPT with a prompt to adopt a professional medical perspective, we created a controlled clinical context for evaluation. It's worth mentioning that during this part of the experiment ChatGPT frequently requested additional diagnostic methods that were not originally provided in the clinical case descriptions. As mentioned earlier, in these instances, ChatGPT was informed that the requested results are not available. Conversely, in certain cases, ChatGPT did not request any additional test results. In this challenging for LLM situation it demonstrated the correct diagnosing accuracy noninferior to doctors' one despite they were provided not only with a full range of information related to cases but also variants of answer. The results obtained in different case subgroups within this phase shed light on ChatGPT's diagnostic process, illustrating instances of inherent accuracy, adaptability, and steadfastness across various diagnostic scenarios, especially when presented with additional data. Overall these outcomes emphasize the proficiency of ChatGPT in emulating the diagnostic decision-making process of real doctors, even when faced with complex diagnostic scenarios.

Nonetheless, the integration of large language models (LLMs) within the medical field brings forth several challenges and limitations, necessitating meticulous attention. Among these challenges are potential biases that LLMs might inherit or magnify from the training data. Such biases could yield inequitable outcomes and impede scientific advancement [4,20]. Moreover, the application of LLMs in medicine introduces concerns regarding patient privacy and data security. Responsible usage of LLMs demands robust safeguards to ensure patient information remains protected [4]. While LLMs exhibit promise in medical contexts, their efficacy must be substantiated through rigorous validation and clinical trials within real-world healthcare settings [21,22]. Ethical considerations also surface with the deployment of LLMs, spanning issues like algorithmic bias and the imperative for transparent decision-making processes. Addressing these ethical dimensions is critical to their responsible application [4,20]. In addition, the generalizability of LLMs to novel or unseen data emerges as a limitation. Their competence might wane on tasks lying outside their training data domain [23]. The utilization of LLMs in medicine accompanies a set of challenges and constraints that warrant scrupulous examination. Acknowledging and mitigating these concerns is paramount to ensure the responsible and effective application of LLMs in healthcare. With appropriate attention, LLMs possess the potential to transform medicine by aiding clinical decision-making, thus fostering enhanced patient outcomes and streamlined healthcare workflows.

CONCLUSIONS

Our findings collectively highlight ChatGPT's high potential and performance as a diagnostic tool in clinical medicine and its value in decision-making process. While it exhibits impressive diagnostic capabilities, particularly in structured scenarios, it also demonstrates the importance of supplementary data and the complex nature of medical diagnosis. This research contributes valuable insights into the evolving landscape of AI-driven diagnostic assistance within clinical contexts. It's worth to note that prompt engineering technics usage plays an important role in the interaction between ChatGPT and a doctor.

REFERENCES

1. Copeland B. Artificial intelligence. Encyclopedia Britannica. 2023. https://www.britannica.com/technology/artificial-intelligence [date access 29.07.2023]
2. Dilmegani C. The Future of Large Language Models. 2023. https://research.aimultiple.com/future-of-large-language-models/ [date access 29.07.2023].
3. Introduction to Large language models; https://attri.ai/blog/introduction-to-large-language-models [date access 29.07.2023].
4. Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. Cureus. 2023;15(5):e39305. doi: 10.7759/cureus.39305.
5. Singhal K, Azizi S, Tu T et al. Large language models encode clinical knowledge. Nature. 2023; 620:172–180. doi:10.1038/s41586-023-06291-2.
6. 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.
7. Egli A. ChatGPT, GPT-4, and other large language models - the next revolution for clinical microbiology? Clin Infect Dis. 2023; ciad407. doi: 10.1093/cid/ciad407.
8. Fei H, Ren Y, Zhang Y et al. Enriching contextualized language model from knowledge graph for biomedical information extraction. Brief Bioinform. 2021;22(3):bbaa110. doi: 10.1093/bib/bbaa110.
9. Gupta R, Srivastava D, Sahu M et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers. 2021;25(3):1315-1360. doi: 10.1007/s11030-021-10217-3.
10. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol. 2019;28(2):73-81. doi:10.1080/13645706.2019.1575882.
11. Hashimoto DA, Witkowski E, Gao L et al. Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. Anesthesiology. 2020;132(2):379-394. doi:10.1097/ALN.0000000000002960.
12. Ossowska A, Kusiak A, Świetlik D. Artificial Intelligence in Dentistry-Narrative Review. Int J Environ Res Public Health. 2022;19(6):3449. doi:10.3390/ijerph19063449.
13. Nensa F, Demircioglu A, Rischpler C. Artificial Intelligence in Nuclear Medicine. J Nucl Med. 2019;60 (2):29S-37S. doi:10.2967/jnumed.118.220590.
14. Keskinbora K, Güven F. Artificial Intelligence and Ophthalmology. Turk J Ophthalmol. 2020;50(1):37-43. doi:10.4274/tjo.galenos.2020.78989.
15. Itchhaporia D. Artificial intelligence in cardiology. Trends Cardiovasc Med. 2022;32(1):34-41. doi:10.1016/j.tcm.2020.11.007.
16. Gilson A, Safranek CW, Huang T et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023; 9:e45312. doi:10.2196/45312.
17. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof. 2023; 20:1. doi:10.3352/jeehp.2023.20.1.
18. Wang X, Gong Z, Wang G et al. ChatGPT Performs on the Chinese National Medical Licensing Examination. J Med Syst. 2023; 47(1):86. doi:10.1007/s10916-023-01961-0.
19. Medscape Case Challenges https://reference.medscape.com/features/casechallenges [date access 29.07.2023]
20. Reddy S. Evaluating large language models for use in healthcare: A framework for translational value assessment. Informatics in Medicine Unlocked. 2023;41:101304. doi: 10.1016/j.imu.2023.101304.
21. 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.
22. Singhal K, Azizi S, Tu T et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2.
23. Safranek C, Sidamon-Eristoff A, Gilson A et al. The Role of Large Language Models in Medical Education: Applications and Implications. JMIR Med Educ. 2023; 1(9):639-650. doi: 10.2196/50945.

ORCID and contributionship:
Yaroslav Mykhalko: 0000-0002-9890-6665 [A,B,D,F]
Pavlo Kish: 0000-0002-9674-0657 [B,D]
Yelyzaveta Rubtsova: 0000-0001-9395-1822 [E]
Oleksandr Kutsyn: 0000-0001-7902-4598 [B,C]
Valentyna Koval: 0000-0001-8423-9534 [C,E]

Conflict of interest:
The Authors declare no conflict of interest.

CORRESPONDING AUTHOR
Yaroslav Mykhalko
Uzhhorod National University
3 Narodna Square, 88000 Uzhhorod , Ukraine
tel: +380660822769
e-mail: yaroslav.myhalko@uzhnu.edu.ua

Received: 30.07.2023
Accepted: 29.10.2023

A - Work concept and design, **B** – Data collection and analysis, **C** – Responsibility for statistical analysis, **D** – Writing the article, **E** – Critical review, **F** – Final approval of the article

CONTENTS

ORIGINAL ARTICLES

Wiadomości Lekarskie
Medical Advances

VOLUME LXXVI, ISSUE 11, NOVEMBER 2023
Official journal of Polish Medical Association has been published since 1928

ISSN 0043-5147
E-ISSN 2719-342X

INDEXED IN PUBMED/MEDLINE, SCOPUS, EMBASE, EBSCO, INDEX COPERNICUS,
POLISH MINISTRY OF EDUCATION AND SCIENCE, POLISH MEDICAL BIBLIOGRAPHY