ORIGINAL ARTICLE

CONTENTS

From open-ended to multiple-choice: evaluating diagnostic performance and consistency of ChatGPT, Google Gemini and Claude AI

Yaroslav O. Mykhalko, Yaroslav F. Filak, Yuliia V. Dutkevych-Ivanska, Mariana V. Sabadosh, Yelyzaveta I. Rubtsova
UZHHOROD NATIONAL UNIVERSITY, UZHHOROD, UKRAINE

ABSTRACT

Aim: To determine the performance and response repeatability of freely available LLMs in diagnosing diseases based on clinical case descriptions.

Materials and Methods: 100 detailed clinical case descriptions were used to evaluate the diagnostic performance of ChatGPT 3.5, ChatGPT 4o, Google Gemini, and Claude AI 3.5 Sonnet large language models (LLMs). The analysis was conducted in two phases: Phase 1 with only case descriptions, and Phase 2 with descriptions and answer variants. Each phase used specific prompts and was repeated twice to assess agreement. Response consistency was determined using agreement percentage and Cohen's Kappa (k). 95% confidence intervals for proportions were calculated using Wilson's method. Statistical significance was set at $p<0.05$ using Fisher's exact test.

Results: In Phase 1 of the study, ChatGPT 3.5, ChatGPT 4o, Google Gemini, and Claude AI 3.5 Sonnet's efficacy was 69.00%, 64.00%, 44.00%, and 72.00% respectively. All models showed high consistency as agreement percentages ranged from 93.00% to 97.00%, and k ranged from 0.86 to 0.94. In Phase 2 all models' productivity increased significantly (90.00%, 95.00%, 65.00%, and 89.00% for ChatGPT 3.5, ChatGPT 4o, Google Gemini, and Claude AI 3.5 Sonnet respectively). The agreement percentages ranged from 97.00% to 99.00%, while k values were between 0.85 and 0.93.

Conclusions: Claude AI 3.5 Sonnet and both ChatGPT models can be used effectively for the differential diagnosis process, while using these models for diagnosing from scratch should be done with caution. As Google Gemini's efficacy was low, its feasibility in real clinical practice is currently questionable.

KEY WORDS: artificial intelligence, large language model, diagnosis, performance

Wiad Lek. 2024;77(9):1852-1856. doi: 10.36740/WLek/195125 DOI

INTRODUCTION

Artificial intelligence (AI) is gaining more and more application in various spheres of modern life. The medical field, which in recent years has undergone significant changes thanks to the introduction of these technologies, was no exception. AI-based solutions are successfully used in pharmaceutical research, development of new drugs, medical documentation management, treatment strategies improvement, interpretation of medical images (including X-rays, CT and MRI), as well as in solving many other tasks [1, 2]. Recently, special attention of the medical community, including scientists, has been attracted by the possibility of using AI systems for diagnosis, forecasting and classification of diseases [3].

An important milestone was the emergence of so-called large language models (LLMs), such as ChatGPT, Google Bard, LLaMA-2 and others. They have brought AI technologies closer not only to scientists, but also to ordinary users. These complex neural network models, trained on huge amounts of data, demonstrate impressive abilities in solving a variety of tasks. The potential of the LLMs is huge and needs to be fully explored.

The use of the LLMs in routine clinical practice opens up broad perspectives from facilitating clinical decision-making to improving medical education and analyzing scientific research. Their ability to process and generate human-like text based on contextual understanding creates new opportunities for improving diagnosis, developing treatment plans and communicating with patients. However, the implementation of such powerful tools in the healthcare system requires a thorough and comprehensive analysis of their effectiveness and reliability.

Numerous studies are being conducted to assess the effectiveness and reliability of available LLMs in various medical fields. In particular, there are many publications devoted to the use of these models for the diagnosis of diseases within narrow medical specialties. A significant part of the research concerned particular diseases [4-11]. Some studies have been conducted to assess the diagnostic performance of LLMs in the context of mul-

tiple medical specialties simultaneously. These works represent a more comprehensive approach and are aimed at evaluating the effectiveness of such models for general clinical use [12-16]. In addition, specialized tests are being developed for a comprehensive assessment of the potential of LLMs in clinical practice [17].

The rapid development of this field necessitates the urgent need to conduct empirical studies that would compare various LLM-based solutions, evaluate their effectiveness according to accepted criteria, and study the possibilities of their interaction with other AI technologies in medicine. Such research is extremely important and plays a key role in shaping approaches to the responsible development and implementation of LLM in medical practice. They are designed to improve the quality and safety of patient treatment and minimize the risks associated with the implementation of AI in routine medical practice.

AIM

The aim of this study was to determine the performance of freely available LLMs in diagnosing diseases based on clinical case descriptions as well as their response repeatability.

MATERIALS AND METHODS

In our study, we evaluated the performance of ChatGPT 3.5, ChatGPT 4o (OpenAI Inc, San Francisco, CA), Google Gemini, and Claude AI 3.5 Sonnet (Anthropic, California, U.S.) in diagnosing diseases based on clinical case descriptions. For that reason 100 clinical cases were used. Each clinical case consisted of detailed information about a patient's complaints, history of present illness, past medical and family histories, results of physical, laboratory and instrumental methods of examination. An average length of clinical case description was 527±74 words. The analysis was conducted in two phases. In Phase 1 models were given only clinical case descriptions while in Phase 2 models were provided with clinical case descriptions along with variants of answer to choose from. In both phases we used an initial prompt to instruct every model. In the first phase the prompt was "In the next prompt I'll give you a description of a clinical case. Act as a professional doctor and diagnose the most suitable disease based on the description. Write the diagnosis only without any explanations". In the second phase this prompt was partially simplified to "In the next prompt I'll give you a description of a clinical case. Act as a professional doctor. Write the diagnosis only without any explanations". This change was done because the sentence "On the basis of these findings only, what is the most likely diagnosis?" was added to the end of each clinical case description before the list of answer variants. Each phase involved presenting the same set of clinical cases twice using the new chat to each LLM. The diagnostic accuracy of the models was estimated as the percentage of correct answers when given a set of clinical cases for the first time to each model. Response consistency and repeatability was determined using the agreement percentage and Cohen's Kappa coefficient (k) along with 95% confidence intervals (CI). k values were interpreted as <0.0 – poor; 0.0-0.2 – slight; 0.2-0.4 – fair; 0.4-0.6 – moderate; 0.6-0.8 – substantial; and 0.8-1.0 – almost perfect agreement [18]. 95% CI for proportions was calculated using Wilson's method. Two-tailed Fisher's exact test was used for comparative analysis of frequency tables. The difference was considered to be statistically significant if $p<0.05$.

RESULTS

Analysis of the diagnostic accuracy in the first phase of the study, which was based only on the description of clinical cases, revealed significant differences between the studied LLMs (Table 1). In particular, the result was the highest in Claude AI 3.5 Sonnet, which established the correct diagnosis in 72 cases out of 100 offered. ChatGPT 3.5 also demonstrated a strong ability to interpret clinical data. The performance of ChatGPT 4o was almost 10% lower compared to Claude AI 3.5 Sonnet. At the same time, no statistically significant difference was found between the results of these LLMs ($p>0.05$). In this phase of the study, the lowest performance was shown by Google Gemini, which correctly diagnosed less than 50% of the given cases and its performance was statistically lower compared to other models ($p<0.05$).

Providing variants of possible answers to the clinical cases description in the second phase of this study significantly increased the diagnostic accuracy of all models. In this phase, ChatGPT 4o correctly identified the largest number of diagnoses (91 cases out of 100). Both ChatGPT 3.5 and Claude AI 3.5 Sonnet showed almost the same performance (90.00 % and 89.00 % respectively, $p>0.05$). Google Gemini, despite a significant improvement compared to the results in Phase 1, showed the lowest efficiency. At the same time, its performance in this phase of the study was statistically lower compared to other LLMs ($p<0.05$).

The degree of improvement in disease diagnosis efficiency of the studied LLMs when comparing phases 2 and 1 varied depending on the model. The highest increase in productivity was demonstrated by Google Gemini (47.73%). ChatGPT 4o and ChatGPT 3.5 had

Table 1. Diagnostic Accuracy and Response Consistency of Large Language Models in Clinical Case Analysis, % (CI)

LLM	Phase 1	Agreements	Phase 2	Agreements
ChatGPT 3.5	69.00 (59.35-77.25)	96.00	90.00 (82.39 - 94.65)#	97.00
ChatGPT 4o	64.00 (54.22 - 72.74)	97.00	91.00 (83.58 - 95.38)#	99.00
Google Gemini	44.00 (34.67 - 53.77)*	93.00	65.00 (55.24 - 73.65)*#	97.00
Claude AI 3.5 Sonnet	72.00 (62.48 - 79.90)	97.00	89.00 (91.17 - 99.35)#	98.00

Note. * - the difference is statistically significant compared to ChatGPT 3.5, ChatGPT 4o and Claude AI 3.5 Sonnet ($p<0.05$), # - the difference is statistically significant compared to the results of Phase 1 ($p<0.05$)

a slightly lower degree of improvement (42.19% and 30.43%, respectively). The lowest increase in efficiency (23.61%) was demonstrated by Claude AI 3.5 Sonnet. These differences in improving disease diagnosis highlight the different ability of each model to use the diagnostic options provided to improve accuracy.

In addition to the effectiveness of LLMs in the diagnosis of various diseases, the repeatability and reproducibility of the results are also important. In Phase 1 of the study, the models showed a high consistency of responses. ChatGPT 4o and Claude AI 3.5 Sonnet showed the highest retest agreement of 97.00%. ChatGPT 3.5 and Google Gemini demonstrated slightly lower, but also high consistency. The obtained results were also confirmed when calculating k coefficients, which were 0.91 (95% CI 0.82 - 1.00), 0.94 (95% CI 0.86 - 1.00), 0.86 (95% CI 0.76 - 0.96) and 0.93 (95% CI 0.84 - 1.00) for ChatGPT 3.5, ChatGPT 4o, Google Gemini and Claude AI 3.5 Sonnet, respectively.

The consistency of LLMs' responses increased in Phase 2 of the study. The agreement percentage of ChatGPT 4o reached almost 100% (k = 0.93, 95% CI 0.862-1.000). Claude AI 3.5 Sonnet had slightly lower but still high agreement rates (98%, k = 0.91, 95% CI 0.86-1.00), as well as ChatGPT 3.5 and Google Gemini (97% each; k = 0.85, 95% CI 0.69-1.00 and 0.93, 95% CI 0.86-1.00, respectively).

DISCUSSION

The ability of LLMs to establish diagnoses based on clinical case descriptions alone is essential for their use in routine medical practice to support clinical decision making. In different studies the efficacy of different LLMs varied greatly [4-16]. Based on the results obtained in our study, three of the four models we tested showed an efficiency of more than 60%. The high results of Claude AI 3.5 Sonnet and ChatGPT 3.5 in diagnosis indicate their particular suitability for real clinical situations where there are no options for differential diagnosis. Interestingly, ChatGPT 3.5 and ChatGPT 4o had comparable performance. Despite the improvements inherent in version 4o, the lack of a significant improvement in performance in the Phase 1 study may indicate that the updates were not specifically aimed at improving medical diagnostic capabilities. At the same time, the relatively low performance of Google Gemini is an example of how important it is to carefully evaluate and validate AI models before they are put into clinical use. The identified differences in the effectiveness of the studied models are determined by many factors such as the type of training data, model architecture, fine-tuning features, understanding of the context and handling of uncertainty. A detailed study of the influence of these factors is critical for improving LLMs with the aim of their further use for medical purposes, as well as the selection of appropriate models for specific healthcare tasks. Future research should focus on identifying and studying these factors to improve the diagnostic capabilities of AI in open-ended scenarios.

The significant increase in diagnostic performance of all models in Phase 2 of this study is a key finding with important practical implications. The performance of Claude AI 3.5 Sonnet, ChatGPT 3.5 and ChatGPT 4o at 89.00-91.00 % suggests that these LLMs can be extremely useful in clinical decision support when used for differential diagnosis. Also of particular interest is the varying degree of improvement among models (from 23.61% to 47.73%) when diagnostic options are added. These differences reflect fundamental variations in how each model handles and uses additional context or constraints. The significant improvements in ChatGPT models indicate their good adaptability to multiple-choice tasks, which in turn may be related to their training methodology or architecture. The increased efficiency in the presence of choice options has several important implications for the practical application of LLMs in healthcare, such as improving human-AI collaboration, reducing diagnostic errors, training physicians and improving the efficiency of the differential diagnosis process.

When evaluating the feasibility of using LLMs in real clinical scenarios, it is important to ensure that the responses provided by these models are not random in nature. To determine the reproducibility of the result, the percentage of response repeatability and the k

coefficient are usually determined. According to the literature, these indicators for the models under study vary widely from average to significant levels [19, 20]. To a large extent, the reproducibility of the results depends on the study design, the prompt structure, and the amount of information provided for analysis.

High percentages of repeatability as well as k observed at both stages of our study indicate high reliability and reproducibility of the obtained results. Such a sequence indicates that the LLMs' responses were not random in nature, but were the result of an analysis of the provided clinical case descriptions. Furthermore, such high levels of consistency indicate that LLMs can maintain consistent performance across multiple trials, reducing the risk of random or unpredictable results. It also determines the possibility of using these models as assistants in clinical decision-making processes.

When the answer options were provided in the second phase of this study, the agreement percentage increased slightly. Although the improvement in repeatability was statistically insignificant ($p>0.05$), it demonstrates the high reliability of using LLMs in a multiple-choice format. Constraining the choice conditions helps these models produce more consistent results because they can better distinguish between the given options than generating responses from scratch, resulting in more consistent performance. The ability to generate consistent results is critical in building trust in AI systems among health care professionals. In addition, checking the repeatability of AI responses can be used in real clinical practice to determine the "confidence" of a particular model in the generated information. Although these findings are encouraging, it is important to note that consistency alone does not guarantee accuracy of results. The high levels of concordance should be considered together with the accuracy results to fully understand the potential and limitations of these models in clinical practice.

CONCLUSIONS

Our study revealed important aspects of LLMs' effectiveness when using them in diagnosis of diseases. Claude AI 3.5 Sonnet and both ChatGPT models showed moderate performance in open-ended scenarios. In the multiple-choice scenarios, their effectiveness was around 90%, which makes them particularly useful in the process of differential diagnosis. Google Gemini's efficacy was significantly lower compared to other models in both study phases, so its feasibility in real clinical practice is currently questionable.

REFERENCES

1. Bohr A, Memarzadeh K. Artificial intelligence in healthcare. Elsevier, Amsterdam. 2020, pp.25-60.
2. Hemamalini MR. The growing role of AI in health care. Journal of Management. 2024;14(6):68-71.
3. Ahsan MM, Luna SA, Siddique Z. Machine-learning-based disease diagnosis: a comprehensive review. Healthcare. 2022;10(3):541-571. doi: 10.3390/healthcare10030541. DOI
4. Hager P, Jungmann F, Holland R et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024. doi: 10.1038/s41591-024-03097-1. DOI
5. Braga AVNM, Nunes NC, Santos EN et al. Use of ChatGPT in Urology and its relevance in clinical practice: is it useful?. Int Braz J Urol. 2024;50(2):192-198. doi:10.1590/S1677-5538.IBJU.2023.0570. DOI
6. 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. DOI
7. Chee J, Kwa ED, Goh X. "Vertigo, likely peripheral": the dizzying rise of ChatGPT. Eur Arch Otorhinolaryngol. 2023;280(10):4687–9. doi: 10.1016/j.psychres.2023.115351. DOI
8. Wei Q, Cui Y, Wei B et al. Evaluating the performance of ChatGPT in differential diagnosis of neurodevelopmental disorders: A pediatricians-machine comparison. Psychiatry Res. 2023;327:115351. doi: 10.1016/j.psychres.2023.115351. DOI
9. Pillai J, Pillai K. Accuracy of generative artificial intelligence models in differential diagnoses of familial Mediterranean fever and deficiency of Interleukin-1 receptor antagonist. J Transl Autoimmun. 2023;7:100213. doi: 10.1016/j.jtauto.2023.100213. DOI
10. Levartovsky A, Ben-Horin S, Kopylov U et al. Towards ai-augmented clinical decision-making: an examination of ChatGPT's Utility in Acute Ulcerative Colitis Presentations. Am J Gastroenterol. 2023;118(12):2283-2289. doi:10.14309/ajg.0000000000002483. DOI
11. Sorin V, Kapelushnik N, Hecht I et al. GPT-4 multimodal analysis on ophthalmology clinical cases including text and images. bioRxiv. 2023. doi: 10.1101/2023.11.24.23298953. DOI
12. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78-80. doi:10.1001/jama.2023.8288. DOI
13. Hirosawa T, Kawamura R, Harada Y et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform. 2023;11:e48808. doi:10.2196/4880. DOI

14. Hirosawa T, Harada Y, Yokose M, Sakamoto T et al. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health. 2023;20(4): 3378. doi: 10.3390/ijerph20043378. DOI
15. Reese JT, Danis D, Caulfied JH et al. On the limitations of large language models in clinical diagnosis. Preprint. medRxiv. 2024;2023.07.13.23292613. doi:10.1101/2023.07.13.23292613. DOI
16. Han T, Adams LC, Bressem K et al. Comparative analysis of GPT-4Vision, GPT-4 and open source LLMs in clinical diagnostic accuracy: a benchmark against human expertise. Preprint. medRxiv. 2023. doi: 10.1101/2023.12.21.23300146. DOI
17. Derek MM, Ye C, Yan Y et al. CliBench: multifaceted evaluation of large language models in clinical decisions on diagnoses, procedures, lab tests orders and prescriptions. 2024. doi:10.48550/arXiv.2406.09923. DOI
18. Gwet K. Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Oxford: Advanced Analytics, LLC, Gaithersburg. 2014, pp.57-62.
19. Freire Y, Santamaría Laorden A, Orejas Pérez J et al. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. The Journal of prosthetic dentistry. 2024;131(4):659.e1-659.e6. doi:10.1016/j.prosdent.2024.01.018. DOI
20. Kochanek K, Skarzynski H, Jedrzejczak WW. Accuracy and Repeatability of ChatGPT Based on a Set of Multiple-Choice Questions on Objective Tests of Hearing. Cureus. 2004;16(5):e59857. doi:10.7759/cureus.59857. DOI

CONFLICT OF INTEREST

The Authors declare no conflict of interest

CORRESPONDING AUTHOR

Yaroslav O. Mykhalko
Uzhhorod National University
3 Narodna Square, 88000 Uzhhorod, Ukraine
e-mail: yaroslav.myhalko@uzhnu.edu.ua

ORCID AND CONTRIBUTIONSHIP

Yaroslav O. Mykhalko: 0000-0002-9890-6665 A B D F
Yaroslav F. Filak: 0000-0002-7510-263X A C D
Yuliia V. Dutkevych-Ivanska: 0000-0003-4306-4234 B D
Mariana V. Sabadosh: 0000-0001-9755-9107 B C
Yelyzaveta I. Rubtsova: 0000-0001-9395-1822 B E

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

RECEIVED: 09.06.2024
ACCEPTED: 20.09.2024

CREATIVE COMMONS 4.0