https://doi.org/10.1590/1806-9282.20230848

Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation

Mauro Gobira[1], Luis Filipe Nakayama[2,3]*, Rodrigo Moreira[1], Eric Andrade[2], Caio Vinicius Saito Regatieri[2], Rubens Belfort Jr.[2]

SUMMARY

OBJECTIVE: The aim of this study was to evaluate the performance of ChatGPT-4.0 in answering the 2022 Brazilian National Examination for Medical Degree Revalidation (Revalida) and as a tool to provide feedback on the quality of the examination.

METHODS: A total of two independent physicians entered all examination questions into ChatGPT-4.0. After comparing the outputs with the test solutions, they classified the large language model answers as adequate, inadequate, or indeterminate. In cases of disagreement, they adjudicated and achieved a consensus decision on the ChatGPT accuracy. The performance across medical themes and nullified questions was compared using chi-square statistical analysis.

RESULTS: In the Revalida examination, ChatGPT-4.0 answered 71 (87.7%) questions correctly and 10 (12.3%) incorrectly. There was no statistically significant difference in the proportions of correct answers among different medical themes (p=0.4886). The artificial intelligence model had a lower accuracy of 71.4% in nullified questions, with no statistical difference (p=0.241) between non-nullified and nullified groups.

CONCLUSION: ChatGPT-4.0 showed satisfactory performance for the 2022 Brazilian National Examination for Medical Degree Revalidation. The large language model exhibited worse performance on subjective questions and public healthcare themes. The results of this study suggested that the overall quality of the Revalida examination questions is satisfactory and corroborates the nullified questions.

KEYWORDS: Artificial intelligence. Natural language processing. Education.

INTRODUCTION

Large language models (LLMs) are deep-learning algorithms capable of processing and generating text data[1]. ChatGPT (OpenAI, San Francisco, CA, USA) is one of the most recognized examples of a LLM and has gained attention for its advanced natural language processing and content-generating capabilities[2-4]. Within 2 months after its release, ChatGPT quickly became the fastest-growing consumer application, with 100 million users[5]. The ChatGPT system is user-friendly, partially free, and can interact with users on a wide range of topics[4]. As this disruptive technology is likely to have a significant and profound impact on various sectors, including healthcare, academic publishing, and medical education, it is crucial that we undertake a comprehensive assessment of its accuracy and reliability, particularly in the field of healthcare[6-8.]

The Brazilian government uses a specific examination, i.e., the National Examination for Revalidation of Medical Diplomas Issued by Foreign Higher Education Institutions ("Revalida"), to validate the training of foreign physicians seeking to practice medicine in Brazil. In the 2022/2 Revalida, a total of 7,006 candidates participated in the first stage, which consisted of objective and essay-based questions. Out of these, 893 candidates advanced to the second stage, which encompassed the practical component. Ultimately, only 263 candidates passed the examination[9].

This study assessed the performance of ChatGPT-4.0 in the Brazilian National Examination for Medical Degree Revalidation and as a tool to provide feedback on the quality of the examination.

METHODS

Language model and data source

ChatGPT powered by GPT version 4.0 (GPT-4) was used because it offers a larger training set that includes a wider variety

[1]Instituto Paulista de Estudos e Pesquisas em Oftalmologia, Vision Institute – São Paulo (SP), Brazil.
[2]Universidade Federal de São Paulo, Department of Ophthalmology – São Paulo (SP), Brazil.
[3]Massachusetts Institute of Technology, Institute for Medical Engineering and Science – Cambridge (MA), USA.
*Corresponding author: nakayama.luis@unifesp.br

Conflicts of interest: the authors declare there is no conflicts of interest. Funding: LFN was supported by the Lemann Foundation, Instituto da Visão-IPEPO. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoas de Nível Superior (CAPES), and the Vision Institute also provided funds for this research.
Received on July 06, 2023. Accepted on July 17, 2023.

of sources than the previous GPT model. Moreover, GPT-4 is expected to be more reliable, creative, and able to handle more nuanced instructions, which would enable more efficient and accurate information processing[3].

The examination data were collected from the 2,022 second-semester Revalida examinations, which are publicly available on the Brazilian government website[9]. The test is organized by the Educational Research and Studies National Institute (INEP) and is composed of two distinct sections: 100 theoretical multiple-choice questions (20 each in the areas of Internal Medicine, Surgery, Pediatrics, Preventive Medicine, and Gynecology and Obstetrics) and 15 discursive questions (4 in Internal Medicine, 2 in Surgery, 4 in Pediatrics, 3 in Preventive Medicine, and 2 in Gynecology and Obstetrics). Additionally, there is a clinical skills test carried out using the Objective Structured Clinical Examination model with 10 stations. After an appeal request, the wrong and dubious questions were nullified by the INEP.

In this analysis, we included all objective questions from the theoretical section and divided them into non-nullified and nullified questions. We excluded discursive and image-based questions, as well as the practical test.

Encoding

The questions were organized according to medical themes in the following subgroups: Preventive Medicine, Gynecology and Obstetrics, Surgery, Internal Medicine, and Pediatrics. The questions were converted to plain text, and the input in the GPT-4 was in Portuguese. Multiple-choice questions were entered in full without forced justification. To reduce the memory retention bias, a new chat session was started in ChatGPT for each entry.

Adjudication

A total of two physicians (MG and RM) independently submitted all the questions and scored them for accuracy. The accuracy of objective answers was evaluated by comparing them with the examination key and classifying the responses as adequate, inadequate, or indeterminate.

The responses were considered adequate when the final answer was aligned with the responses. Inadequate answers were defined as instances in which an incorrect answer was chosen. Responses were deemed indeterminate when the answer was not present in the set of available options or there were insufficient data to provide a confident answer.

After individual evaluations, the physicians performed a third assessment to reach a consensus on the questions with differing results. The accuracy of the responses that the ChatGPT-4 presented to questions that included mathematical concepts was also included. Figure 1 shows a diagrammatic summary of the study protocol.

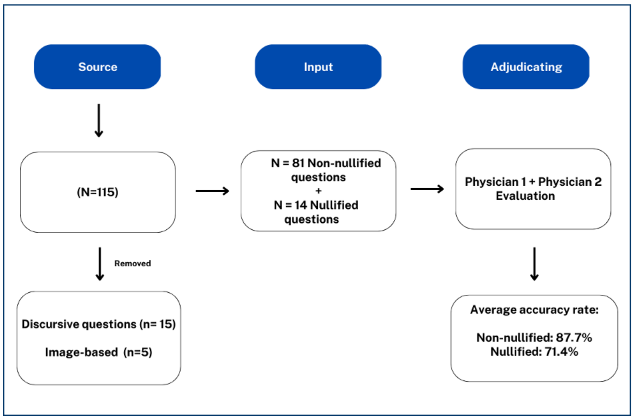
Source
Input
Adjudicating
(N=115)
N = 81 Non-nullified questions
+
N = 14 Nullified questions
Physician 1 + Physician 2 Evaluation
Removed
Discursive questions (n= 15)
Image-based (n=5)
Average accuracy rate:
Non-nullified: 87.7%
Nullified: 71.4%

Figure 1. Workflow for sourcing, encoding, and adjudicating results.

Statistical analysis

The statistical analysis performed using Fisher's exact test compared the LLM performance across the test themes due to small sample sizes and chi-square between non-nullified and nullified questions. A two-sided test was used with a statistical significance level of α=0.05, and statistical analysis was performed using Python 3.9 libraries.

Incorrect answers were described and analyzed to evaluate the model's performance, aiming to investigate possible biases, limitations, or algorithm hallucinations.

The responses generated by ChatGPT for non-nullified questions were assessed based on the official final answer key provided by INEP. For nullified questions, the evaluation was based on the preliminary answer key provided.

RESULTS

A total of 81 objective questions were included for evaluation. Notably, 14 nullified questions were evaluated separately to analyze the reasons for their exclusion from the examination, and five image-based questions were excluded because, as a language model, ChatGPT is not designed to analyze visual content. The evaluators agreed on 72 (88.89%) of the answers and disagreed on 9 (11.11%), with no answer classified as indeterminate by both evaluators. In the settled results, the ChatGPT answered 71 (87.7%) questions correctly, 10 (12.3%) incorrectly, and there was no indeterminate answer.

Test themes

A comparison of ChatGPT-4 results through different medical themes showed the following performance: Internal Medicine 100%, Gynecology and Obstetrics 88.9%, Surgery 85.7%, Pediatrics 83.3%, and Preventive Medicine 81.3%. There was no statistically significant difference in the proportions of correct answers among different medical themes (p=0.241). The comparison within groups is described in Figure 2.

Incorrect answers

ChatGPT answered two Gynecology and Obstetrics questions incorrectly, one about medical conduct regarding hemorrhagic shock secondary to spontaneous abortion and the other regarding the ethics of prescribing the morning-after pill. ChatGPT answered three Pediatrics questions incorrectly, which included gestational age neonates with respiratory discomfort, a neonatal dermatologic lesion, and an inguinal hernia. ChatGPT also answered three Preventive Medicine questions incorrectly, which included the approach for educating the population on the importance of COVID vaccines, the prescription of contraceptive methods, and violence against women, all of which required ethical considerations. Finally, ChatGPT answered two surgery questions incorrectly, one relating to a thyroid lesion and the other involving a clinical case.

Nullified questions

Of the 14 nullified questions, ChatGPT had an accuracy of 71.4% (n=10), with no indeterminate answers. There was no significant statistical difference in the performance of ChatGPT with non-nullified and nullified questions (p=0.240). The nullified questions comprised Gynecology and Obstetrics (n=2), Internal Medicine (n=4), Pediatrics (n=2), Surgery (n=3), and Preventive Medicine (n=3), including topics such as prescription of contraceptives, human papillomavirus screening, dyslipidemia, *Helicobacter pylori* infection and bulimia treatment, acute intoxication, neonatal syphilis, breastfeeding, cholelithiasis, trauma, recommendations against COVID-19, high-risk prenatal care, and hypertension treatment.

DISCUSSION

In this study, ChatGPT-4.0 had an overall accuracy of 87.7% in the non-nullified questions and 71.4% in the nullified questions, which corroborated the low quality of the excluded questions. Potential reasons for nullification may include ambiguity, insufficient clarity, or the presence of multiple valid answers. The evaluation comprises the objective questions of the test, excluding image-based and discursive questions.

Several articles have examined the performance of ChatGPT in the domain of general medical content[10,11]. Notably, GPT-4 demonstrated impressive performance by successfully passing Japanese medical licensing exams from 2018 to 2022[10]. Furthermore, Kung and collaborators conducted a separate study that highlighted ChatGPT's capability to pass the United States Medical Licensing Examination without any human intervention[11]. Regarding open-ended questions, ChatGPT

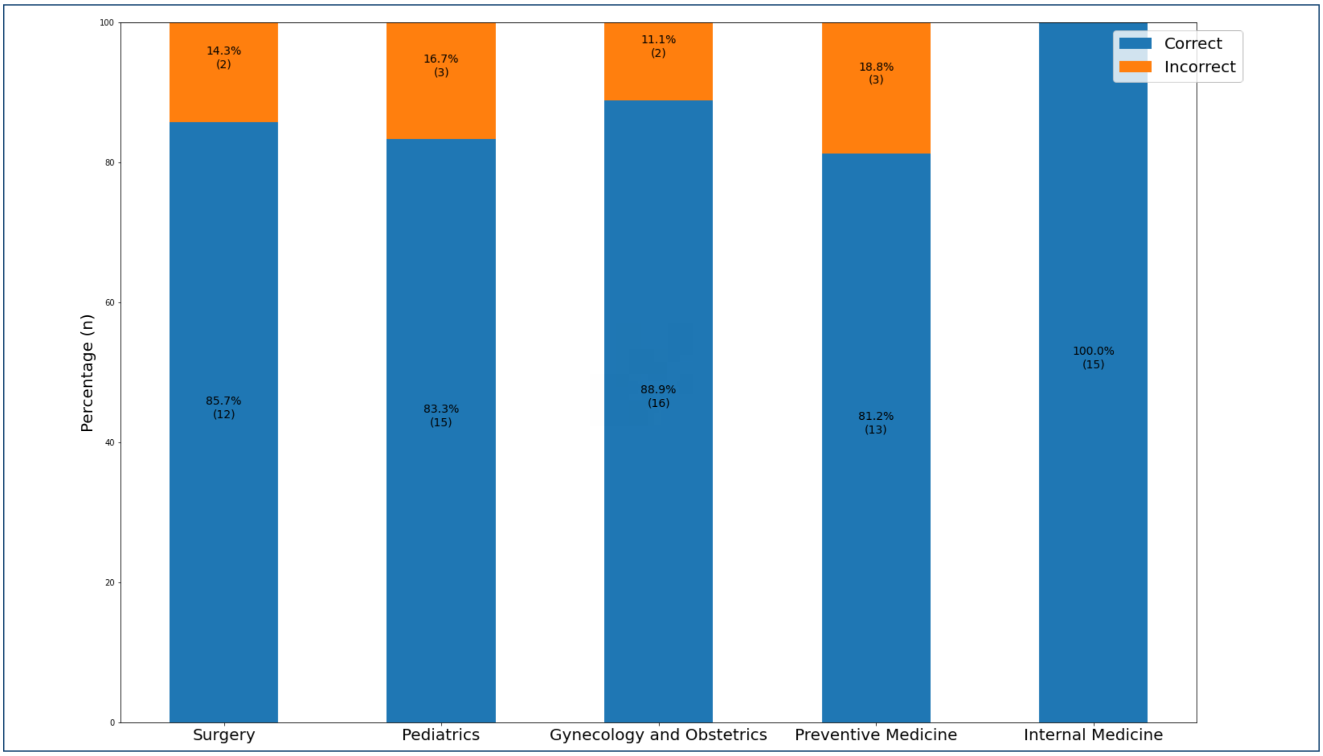

Figure 2. Performance across different examination topics.

achieved accuracies of 75.0%, 61.5%, and 68.8% for USMLE Steps 1, 2CK, and 3, respectively[11].

More recently, Belfort et al. and Mihalache evaluated the performance of ChatGPT-3.5 in answering specialist-level ophthalmology questions and found a low accuracy rate of 40.2% in the Brazilian Board exam and 46% in the American OphtoQuestions[12,13]. On the contrary, in a study conducted by the NUS Obgyn-AI Collaborative Group, ChatGPT outperformed human candidates in a Virtual Objective Structured Clinical Examination (OSCE) in Obstetrics and Gynecology. The average ChatGPT score was 77.2%, while the average historical human score was 73.7%[14].

It is important to emphasize that the evaluation of the ChatGPT-4.0's responses by two physicians allowed for a more rigorous and accurate analysis of the information provided, which contributed to a higher level of reliability when evaluating the results. The subjectivity of medical questions was also a factor to be considered, as there is often no single or correct answer for certain clinical cases, which can generate different interpretations and opinions among experts.

Comparing the medical themes, there was a discrepancy in the accuracy of the responses, with variations ranging from 100 to 81.3%, but with no statistical difference within groups due to the small sample sizes. Different question characteristics, such as clarify and less ambiguity in some themes, may have contributed to the varying performances. ChatGPT encountered challenges when responding to subjective questions within the Preventive Medicine section, which involved specific concepts related to the Brazilian public healthcare system and ethical decisions. In addition, ChatGPT did not cover information from 2022, which was related to recently issued guidelines, and included the COVID-19 pandemic. This may account for the lower accuracy in the Preventive Medicine subject area.

The results of this study show acceptable ChatGPT performance compared to results from other tests. However, the published studies used different methodologies and evaluated different examinations, making it difficult to compare the results[10-14]. The significant discrepancy between the results may be associated with the fact that the Revalida examination involves general medical topics. Another consideration is that we used the ChatGPT 4.0 version, which may have a positive impact on the results.

The Revalida examination does not include mathematical questions for evaluation. It is noteworthy that ChatGPT typically does not perform well on numerical problem-solving tasks, as previous studies reported, and may present different performances according to the language used[15].

It is essential to conduct additional and comprehensive research to investigate the underlying factors behind the low approval rate in the Revalida examination and to consider aspects such as the quality of medical education, candidate preparation, effectiveness of evaluation methods, and potential socioeconomic and cultural barriers that may affect the performance of participants[9].

LLMs are straightforward in the decision-making process and struggle with dubious and uncertain questions. The results of this study suggest that the overall quality of the Revalida examination questions is satisfactory and corroborates the annulled questions. In light of these findings, it is crucial for the medical community to recognize the potential of artificial intelligence tools such as ChatGPT-4.0 while also acknowledging their limitations. Further research is necessary to enhance the accuracy and reliability of artificial intelligence in medical education.

In conclusion, ChatGPT-4.0 performed satisfactorily at the 2022 Brazilian National Examination for Medical Degree Revalidation. The LLM exhibited a worse performance in subjective questions, public healthcare themes, and nullified questions.

AUTHORS' CONTRIBUTIONS

MG: Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing. **LFN:** Conceptualization, Data curation, Formal Analysis, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. **RM:** Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing. **EA:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **PA:** Validation. **CVSR:** Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **RBJ:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing.

REFERENCES

1. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl. 2023;82(3):3713-744. https://doi.org/10.1007/s11042-022-13428-4
2. Introducing ChatGPT. [cited on Mar 29, 2023] Available from: https://openai.com/blog/chatgpt/
3. GPT-4. [cited on Mar 29, 2023] Available from: https://openai.com/research/gpt-4

4. Sallam M. ChatGPT Utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11(6):887. https://doi.org/10.3390/healthcare11060887

5. Bartz D, Bartz D. As ChatGPT's popularity explodes, US lawmakers take an interest; 2022.

6. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J. 2023;64(1):1-3. https://doi.org/10.3325/cmj.2023.64.1

7. Névéol A, Zweigenbaum P. Clinical natural language processing in 2014: foundational methods supporting efficient healthcare. Yearb Med Inform. 2015;10(1):194-8. https://doi.org/10.15265/IY-2015-035

8. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond). 2023;23(3):278-9. https://doi.org/10.7861/clinmed.2023-0078

9. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198

10. Gobira MC, Moreira RC, Nakayama LF, Regatieri CVS, Andrade E, Belfort Jr. R. Performance of chatGPT-3.5 answering questions from the Brazilian Council of Ophthalmology Board Examination. Pan Am J Ophthalmol. 2023;5(1):17. https://doi.org/10.4103/pajo.pajo_21_23.

11. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence Chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589-97. https://doi.org/10.1001/jamaophthalmol.2023.1144

12. Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol. 2023;229(2):172.e1-12. https://doi.org/10.1016/j.ajog.2023.04.020

13. Revalida. INEP - Revalida. [cited on April 01, 2023] Available from: http://revalida.inep.gov.br/revalida/

14. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ. 2023;9:e48002. https://doi.org/10.2196/48002

15. Frieder S, Pinchetti L, Griffiths RR, Salvatori T, Lukasiewicz T, Petersen PC, et al. Mathematical capabilities of ChatGPT. arXiv [cs.LG]. 2023.http://arxiv.org/abs/2301.13867.