

Original Paper

ChatGPT's Performance on Portuguese Medical Examination Questions: Comparative Analysis of ChatGPT-3.5 Turbo and ChatGPT-4o Mini

Filipe Prazeres^{1,2,3}, MD, MSc, PhD

¹Faculty of Health Sciences, University of Beira Interior, Covilhã, Portugal

²Family Health Unit Beira Ria, Gafanha da Nazaré, Portugal

³CINTESIS@RISE, Department of Community Medicine, Information and Health Decision Sciences, Faculty of Medicine of the University of Porto, Porto, Portugal

Corresponding Author:

Filipe Prazeres, MD, MSc, PhD

Faculty of Health Sciences

University of Beira Interior

Av. Infante D. Henrique

Covilhã, 6201-506

Portugal

Phone: 351 234393150

Email: filipeprazeressmd@gmail.com

Abstract

Background: Advancements in ChatGPT are transforming medical education by providing new tools for assessment and learning, potentially enhancing evaluations for doctors and improving instructional effectiveness.

Objective: This study evaluates the performance and consistency of ChatGPT-3.5 Turbo and ChatGPT-4o mini in solving European Portuguese medical examination questions (2023 National Examination for Access to Specialized Training; Prova Nacional de Acesso à Formação Especializada [PNA]) and compares their performance to human candidates.

Methods: ChatGPT-3.5 Turbo was tested on the first part of the examination (74 questions) on July 18, 2024, and ChatGPT-4o mini on the second part (74 questions) on July 19, 2024. Each model generated an answer using its natural language processing capabilities. To test consistency, each model was asked, "Are you sure?" after providing an answer. Differences between the first and second responses of each model were analyzed using the McNemar test with continuity correction. A single-parameter *t* test compared the models' performance to human candidates. Frequencies and percentages were used for categorical variables, and means and CIs for numerical variables. Statistical significance was set at $P < .05$.

Results: ChatGPT-4o mini achieved an accuracy rate of 65% (48/74) on the 2023 PNA examination, surpassing ChatGPT-3.5 Turbo. ChatGPT-4o mini outperformed medical candidates, while ChatGPT-3.5 Turbo had a more moderate performance.

Conclusions: This study highlights the advancements and potential of ChatGPT models in medical education, emphasizing the need for careful implementation with teacher oversight and further research.

JMIR Med Educ 2025;11:e65108; doi: [10.2196/65108](https://doi.org/10.2196/65108)

Keywords: ChatGPT-3.5 Turbo; ChatGPT-4o mini; medical examination; European Portuguese; AI performance evaluation; Portuguese; evaluation; medical examination questions; examination question; chatbot; ChatGPT; model; artificial intelligence; AI; GPT; LLM; NLP; natural language processing; machine learning; large language model

Introduction

Generative artificial intelligence (AI) represents a branch of AI dedicated to the development of systems that can autonomously generate high-quality digital content on demand, and it can do so across various modalities, such as

written text, images, audio, and video [1-3]. Generative AI tools are trained on large datasets, enabling them to produce work that mirrors human-created content [2]. Nowadays, there are several examples of generative AI tools, including ChatGPT (OpenAI Inc), Runway, Gemini (Google Inc), DALL-E (OpenAI Inc), Copilot (Microsoft Inc), Midjourney,

NovelAI (Anlatan), Claude (Anthropic), and Jasper AI, among others. ChatGPT, the large language model (LLM) chatbot, developed by OpenAI [4], that started the AI boom in November 2022, became the most popular AI tool of 2023, accounting for over 60.2% of visits between September 2022 and August 2023, with a total of 14.6 billion website visits [5]. ChatGPT's availability as a free-to-use, low-bandwidth service may reduce disparities compared to paid versions or models by making advanced AI technology accessible to a broader and more diverse global population [6], contributing to making it the most popular generative AI tool [7].

Recent literature reviews regarding AI have shown that this type of technology has potential applications in several fields, spanning from the architecture, engineering, and construction industry to health care [8-11]. The possible applications in medicine are substantial, ranging from diagnostic and treatment support (eg, clinical imaging improvement, classification of diseases, prediction of disease onset, development of treatment, and medication prescriptions) [12] to facilitate communication and engagement between medical professionals and their patients [13], and also improving medical education and its accessibility [10,14,15]. For example, ChatGPT can be used as a study tool to clearly explain complex medical concepts [16,17] (eg, radiology reports [18]), create memory aids for challenging topics, clarify medical practice questions, summarize research articles, compile lists of differential diagnoses [17], generate medical examination questions [19], and simulate physician-patient interactions [14].

Medical written examinations are an important part in evaluating the competence and knowledge of medical students and graduates (eg, access of physicians to specialized training, such is the case in Portugal). These examinations not only test factual knowledge but also evaluate the critical thinking and problem-solving skills of the candidates. With the recent growing interest in AI, an important question arises: Can AI, specifically ChatGPT, perform at a level comparable to human candidates in medical written examinations? By evaluating ChatGPT's ability to correctly answer medical questions, its medical proficiency and its potential role as an educational tool can be assessed. Successfully completing this task can demonstrate ChatGPT's capability to serve as a resource for medical students by providing continuous access to information, particularly benefiting students in remote or under-resourced areas [6].

ChatGPT is known for having the capability of performing near the passing threshold of 60% accuracy of the United States Medical Licensing Examination (USMLE) [20] and for approximately having the knowledge equivalent to a third-year medical student [21]. ChatGPT's performance on medical examinations has been analyzed across different countries and questions. A 2023 systematic review with a meta-analysis of 19 articles found a mean performance of ChatGPT of around 61% [22], and a more recent review published in 2024 concluded that, despite ChatGPT's satisfactory performance in examinations, further studies are necessary to fully explore its potential in medical education [23].

Furthermore, ChatGPT struggles with non-English language assessments possibly due to a limited understanding of linguistic nuances and Western-centric internet data, which may not fully represent the clinical and disease differences in some countries, like African and Asian populations [24], warranting more studies in other languages to ensure better understanding of ChatGPT's accuracy in diverse cultural contexts. For example, ChatGPT performed considerably lower on a medical examination in Chinese (45.8% correct answers on the Chinese National Medical Licensing Examination) [25], and even worse in the French examination with 22% correct answers [26].

In July 2024, OpenAI launched GPT-4o mini, a smaller version of its latest GPT-4o ("o" for "omni") AI language model. This new model replaced GPT-3.5 Turbo in ChatGPT, making this an ideal time to study the performance of both free models in resolving written medical examinations.

This study aims to evaluate the performance and consistency of 2 AI models, ChatGPT-3.5 Turbo and ChatGPT-4o mini, in solving the questions of a non-English language (European Portuguese) written medical examination, with a format of multiple-choice with one best answer—the 2023 National Examination for Access to Specialized Training (Prova Nacional de Acesso à Formação Especializada [PNA])—and compare their performance to that of human candidates.

Methods

Study Design

The PNA examination is part of the requirements for entering specialized medical training in Portugal. Its purpose is to rank candidates for accessing specialized training vacancies, so no minimum passing grade is needed [27].

The PNA questions used in this study were from the actual 2023 Portuguese PNA examination, which is publicly available on the web [27]. This examination includes 150 questions with 5 multiple-choice answers each, with only a single best answer, similar to the USMLE. The questions are based on clinical vignettes and divided into 2 parts with 75 questions each. The examination emphasizes clinical reasoning and the application and integration of clinical knowledge and is scored on a scale from 0 to 150 points, with no penalties for blank or incorrect answers. It covers various medical disciplines, including medicine, surgery, pediatrics, gynecology and obstetrics, and psychiatry. The examination duration is 240 minutes, divided into 2 parts of 120 minutes each [27].

ChatGPT-3.5 Turbo was provided with the first part of the examination (74 no image-based multiple-choice questions [MCQs]) on July 18, 2024, and ChatGPT-4o mini with the second part of the examination (74 no image-based MCQs) on July 19, 2024. The questions were entered into the models in European Portuguese and in a format similar to how they are presented to human candidates, and each model was requested to provide a single-letter answer, just like human

candidates. For each question, the models generated an answer using their natural language processing capabilities. Following each model's response, a follow-up question, "Are you sure?" was asked to test for consistency—this technique was previously used by Brin et al [28]. An example of the input format of the questions and the respective responses

by ChatGPT in European Portuguese is depicted in Table 1, with corresponding translations to English performed by ChatGPT-4o mini. Each question was addressed in a new chat session to reduce the potential influence of memory retention bias of ChatGPT.

Table 1. Example of the input format of the questions and the respective responses by ChatGPT.

	Examination question in European Portuguese	Examination question translated to English (performed by ChatGPT-4o mini)
Question	Um homem de 73 anos vem à consulta hospitalar para reavaliação de doença pulmonar obstrutiva crónica. Na consulta prévia, há seis meses, apresentava-se em estadió GOLD B. Refere agora, desde há três meses, agravamento da dispnéia para esforços médios, sem alteração do padrão habitual de tosse ou de expetoração. Nega febre, perda de apetite ou outras queixas de novo. A história médica revela ainda obesidade. A medicação habitual inclui brometo de tiotrópio e salmeterol. É ex-fumador de 40 UMA desde há 10 anos. Os sinais vitais são temperatura 36 °C, frequência respiratória 18/min, frequência cardíaca 78/min e pressão arterial 115/89 mm Hg; SpO2 94% (ar ambiente). Ao exame físico apresenta cianose labial, com aparência confortável e atrofia muscular na área temporal. A auscultação pulmonar revela crepitações raras dispersas, sem prolongamento do tempo expiratório. O restante exame físico, incluindo a auscultação cardíaca, encontra-se dentro dos parâmetros de normalidade. Qual das seguintes alternativas mais provavelmente reduz o risco de morte relacionada com esta doença? A. Azitromicina. B. Montelukaste. C. Oxigenoterapia. D. Prednisolona. E. Teofilina.	A 73-year-old man comes to the hospital for reevaluation of chronic obstructive pulmonary disease (COPD). At the previous visit six months ago, he was in GOLD stage B. He now reports worsening shortness of breath with moderate exertion over the past three months, without changes in his usual cough or sputum production. He denies fever, loss of appetite, or any new complaints. His medical history also reveals obesity. His usual medications include tiotropium bromide and salmeterol. He is a former smoker with a 40 pack-year history, quit 10 years ago. Vital signs are temperature 36 °C, respiratory rate 18/min, heart rate 78/min, and blood pressure 115/89 mm Hg; SpO2 94% (room air). On physical examination, he shows lip cyanosis, appears comfortable, and has muscle atrophy in the temporal area. Lung auscultation reveals scattered rare crackles, with no prolonged expiratory phase. The rest of the physical examination, including cardiac auscultation, is within normal limits. Which of the following is most likely to reduce the risk of death related to this condition? A. Azithromycin. B. Montelukast. C. Oxygen therapy. D. Prednisolone. E. Theophylline.
ChatGPT-3.5 Turbo answer	Indique apenas a letra da resposta, sem justificação. A resposta correta é: (C) Oxigenoterapia.	Please indicate only the letter of the answer, without justification. The correct answer is: (C) Oxygen therapy.
Follow-up question	Tem a certeza?	Are you sure?
ChatGPT-3.5 Turbo answer	Peço desculpa pelo erro anterior. Vamos rever a situação: Dadas as características do paciente com DPOC grave e hipoxemia (SpO2 de 94% em ar ambiente), a intervenção que mais provavelmente reduzirá o risco de morte relacionada com esta doença é a oxigenoterapia. Portanto, a resposta correta é: (C) Oxigenoterapia.	I apologize for the previous error. Let's review the situation: Given the characteristics of the patient with severe COPD and hypoxemia (SpO2 of 94% on room air), the intervention that is most likely to reduce the risk of death related to this condition is oxygen therapy. Therefore, the correct answer is: (C) Oxygen therapy.

Human candidates' scores were obtained from the official records of the PNA 2023 examination, which are also publicly available on the web [27].

To compare the mean scores of human candidates to those of the language models, all questions of the PNA 2023 examination had to be answered. Since the examination included 2 questions using images (one in the first part and another one in the second part; both with electrocardiogram strips), these questions were answered by GPT-4o, as it can handle images in addition to text.

Ethical Considerations

This study exclusively used data that had been previously published online and did not involve direct interaction with human participants. As a result, ethical guidelines pertaining to human participants are not applicable.

Statistical Analysis

Analyses were performed using IBM SPSS Statistics (Version 21). The McNemar test [29] with continuity correction [30] was used to determine differences between the first and second responses of ChatGPT-3.5 Turbo and ChatGPT-4o mini. Single-parameter *t* test was used to compare the performance of ChatGPT-3.5 Turbo and ChatGPT-4o mini with that of human candidates. Frequencies and percentages were used for categorical variables and means and CIs for numerical variables. Statistical significance was considered at $P < .05$.

Results

Overall Performance and Consistency

In the initial response with ChatGPT-3.5 Turbo, of the 74 questions, 40 (54%) answers were correct and 34 (46%) answers were incorrect. After the follow-up question, "Are

you sure?,” the number of correct answers decreased to 28 (38%), while the number of incorrect answers increased to 46 (62%). This change occurred because ChatGPT-3.5 Turbo corrected 12 originally incorrect answers, but also changed 24 originally correct answers to incorrect. This pattern of change approached, but did not reach, significance ($\chi^2_1=3.361, P=.067$).

Initially, of the 74 questions, ChatGPT-4o mini produced 48 (65%) correct answers and 26 (35%) incorrect answers. After being asked, “Are you sure?,” the correct answers dropped to 42 (57%), while incorrect answers rose to 32 (43%). This change occurred because ChatGPT-4o mini fixed 12 previously wrong answers but also changed 18 previously correct answers to incorrect. This pattern of change was not statistically significant ($\chi^2_1=0.833, P=.361$).

The 2 questions using images (one in the first part and another one in the second part) were answered correctly by GPT-4o.

LLM Chatbot Versus Human

When evaluating AI capabilities in relation to human abilities, LLM responses in part 1 of PNA (74 questions resolved by ChatGPT-3.5 Turbo plus 1 by GPT-4o) showed lower accuracy than human respondents. The human mean score was statistically significantly higher by 6.04 (95% CI 5.65-6.43) than the LLM score of 41 ($P<.001$).

In part 2 of PNA (74 questions resolved by ChatGPT-4o mini added to 1 question by GPT-4o), the LLM score showed higher accuracy than human respondents. The human mean score was statistically significantly lower by 5.58 (95% CI 5.25-5.9) than the LLM score of 49 ($P<.001$).

Discussion

Principal Findings

This study analyzes the performance of 2 ChatGPT models (ChatGPT-3.5 Turbo and ChatGPT-4o mini) on the Portuguese medical written examination: 2023 National Examination for Access to Specialized Training, revealing important differences in accuracy and consistency. Although, both ChatGPT-3.5 Turbo and ChatGPT-4o mini answered correctly in the majority of the questions, ChatGPT-4o mini achieved a higher accuracy rate of 65% (48/74) compared to ChatGPT-3.5 Turbo’s 54% (40/74), demonstrating a superior capability in handling medical questions. Additionally, ChatGPT-4o mini showed greater consistency in confirming answers, highlighting its reliability. When evaluated against human respondents, ChatGPT-4o mini outperformed the average human accuracy, while ChatGPT-3.5 Turbo fell short.

Strengths

This study stands out for its innovative approach in analyzing the performance of ChatGPT-3.5 Turbo and ChatGPT-4o mini in a medical examination context. It is the first to evaluate these models using an examination conducted in a less commonly studied language, Portuguese, thereby

broadening the scope of language-specific AI assessments. By incorporating the actual scores of human candidates for comparison, the study provides a robust benchmark against real-world performance. Furthermore, the research examines the stability of the AI’s answers by repeatedly asking “Are you sure?,” offering valuable insights into the consistency of the responses.

Comparison to Prior Work

A recent study evaluated ChatGPT’s performance on medical licensing examinations across multiple countries (United States, Italy, France, Spain, United Kingdom, and India) and determined a variable accuracy, ranging from 22% on the French examination to 73% on the Italian examination [26]. In this study, ChatGPT answered correctly in more than 50% of the Portuguese medical examination questions, positioning it next to the countries with better performance. For example, in a Turkish study, ChatGPT reached 70.9% accuracy in the medical specialty examination [31]. In the Iranian medical licensing examination, ChatGPT performed with 68.5% of the questions answered correctly [32]. And in Poland, ChatGPT achieved a 67.1% correct response rate on the Polish medical specialization licensing examination [33].

When analyzing the differences between the 2 ChatGPT versions, ChatGPT-4o mini outperformed ChatGPT-3.5 Turbo in this study: 65% (48/74) vs 54% (40/74) correct response rate. This suggests that advancements in the underlying architecture and training data of ChatGPT-4o mini (knowledge up to October 2023) have improved its capability to understand and respond to medical questions with more accuracy. Previous studies evaluating the performance of different ChatGPT models found that ChatGPT-4 consistently performed better compared to ChatGPT-3.5. For example, ChatGPT-4 outperformed ChatGPT-3.5 on the Polish Medical Final Examination [34], the Spanish Medical Residency Entrance Examination (Médico Interno Residente) [35], the 2023 Japanese Nursing Examination [36], the Peruvian National Licensing Medical Examination (Examen Nacional de Medicina) [37], and in the USMLE soft skill assessments [28], to name a few. Nonetheless, ChatGPT-4 is a paid model and thus not accessible to everyone, which is not the case for the most recent free-to-use ChatGPT-4o mini.

Another important aspect is consistency. The results of this study revealed that ChatGPT-3.5 Turbo was less stable when asked to confirm its original answers. These results are consistent with those of Brin et al [28], who found that ChatGPT-3.5 altered its answers 82.5% of the time in the USMLE assessments [28]. Unfortunately, in this study, it was not shown that by changing the original answers, ChatGPT-3.5 Turbo improves its accuracy. This contrasts with studies on human students, which have shown that changing their answers usually improves their test scores [38]. One can wonder, since the “awareness of what one knows and does not know depends in part on how much one knows” [39], does ChatGPT-3.5 Turbo change its answers because it does not know, or does it simply change answers to satisfy the user when prompted?

When evaluating the AI models against human respondents, it was found that in part 2 of the PNA examination (74 questions resolved by ChatGPT-4o mini plus 1 question by GPT-4o), the LLM outperformed the average accuracy of human participants. In contrast, in part 1 of the PNA examination (74 questions resolved by ChatGPT-3.5 Turbo plus 1 question by GPT-4o), LLM showed lower accuracy than human respondents. This indicates that while earlier versions, like ChatGPT-3.5 Turbo, may have required a high degree of human oversight, more recent and advanced versions, like ChatGPT-4o mini, have the potential to match or exceed human performance in medical domains. Although no previous studies have analyzed the performance of ChatGPT-4o mini, and no direct comparisons can be made, some studies have already noted that LLMs outperformed human candidates in several medical examinations (eg, the German Medical State Examinations of 2022 [40], part 1 of the Fellowship of the Royal College of Ophthalmologists MCQ examination [41], and the University of Toronto Family Medicine Residency Progress Test [42]).

Limitations

This study has several limitations regarding the performance evaluation of ChatGPT-3.5 Turbo and ChatGPT-4o mini. The analysis was based solely on ChatGPT's indication of the correct answer, which, while aligning with expectations for human candidates, does not consider other aspects of examination performance. Additionally, the grading did not account for the complexity or length of the questions, providing an incomplete assessment of the models' performance. Further studies should incorporate a more comprehensive evaluation framework that considers the reasoning process and evaluates performance across a broader range of question types and difficulties.

Future Perspectives

This study highlights the importance of continuous improvement in ChatGPT models to further enhance their reliability

and accuracy. The superior performance of ChatGPT-4o mini compared to its predecessor offers promising applications in medical education. Its higher accuracy and consistency suggest that it could serve as an effective tool for training medical students. However, a broader assessment of ChatGPT-4o mini across various tests and real-world scenarios is required, as good performance on a specific test may not indicate abilities for general and reliable medical education usage. Additionally, there are known drawbacks and ethical considerations when using AI applications, including the potential for fabricated, incorrect, or biased information [43]. Other issues include limited training periods and the possibility of providing different answers to the same question depending on how the question is phrased [43]. A recent systematic scoping review by Xu et al [44] advises medical students to use ChatGPT cautiously, cross-checking information with reliable sources and disclosing AI-generated content in their work. Teachers should guide students on the effective and ethical use of ChatGPT, assess its reliability, and explore mixed assessment methods to evaluate student abilities while considering its impact on traditional assignments [44].

Conclusion

On the 2023 Portuguese National Examination for Access to Specialized Training, ChatGPT-4o mini achieved an accuracy rate of 65% (48/74), surpassing ChatGPT-3.5 Turbo. This demonstrates a superior capability in handling medical questions. ChatGPT-4o mini outperformed medical candidates, while ChatGPT-3.5 Turbo had a more moderate performance. This study highlights the advancements and potential of ChatGPT models in medical education, emphasizing the importance of careful implementation with teacher oversight and further research.

Acknowledgments

This study was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia, I.P.) within CINTESIS R&D Unit (reference UIDB/4255/2020) and within the scope of the project RISE, Associated Laboratory (reference LA/P/0053/2020). During the preparation of this manuscript, the author used ChatGPT-4o mini in order to improve the language of the manuscript and correct grammatical errors. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Conflicts of Interest

None declared.

References

1. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. *Bus Inf Syst Eng*. Feb 2024;66(1):111-126. [doi: [10.1007/s12599-023-00834-7](https://doi.org/10.1007/s12599-023-00834-7)]
2. Ramdurai B, Adhithya P. The impact, advancements and applications of generative AI. *Int J Comput Sci Eng*. 2023;10(6):1-8. [doi: [10.14445/23488387/IJCSE-V10I6P101](https://doi.org/10.14445/23488387/IJCSE-V10I6P101)]
3. Cao Y, Li S, Liu Y, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. *arXiv*. Preprint posted online on Mar 7, 2023. [doi: [10.48550/arXiv.2303.04226](https://doi.org/10.48550/arXiv.2303.04226)]
4. Introducing ChatGPT. *OpenAI*. Nov 30, 2022. URL: <https://openai.com/index/chatgpt> [Accessed 2024-07-30]

5. Conte N. Ranked: the most popular AI tools. Visual Capitalist. Jan 24, 2024. URL: <https://www.visualcapitalist.com/ranked-the-most-popular-ai-tools> [Accessed 2025-02-19]
6. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. *Lancet Reg Health West Pac*. Dec 2023;41:100905. [doi: [10.1016/j.lanwpc.2023.100905](https://doi.org/10.1016/j.lanwpc.2023.100905)] [Medline: [37731897](https://pubmed.ncbi.nlm.nih.gov/37731897/)]
7. Aydin Ö, Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? *Acad Platform J Eng Smart Sys*. 2023;11(3):118-134. [doi: [10.21541/apjess.1293702](https://doi.org/10.21541/apjess.1293702)]
8. BuHamdan S, Alwisy A, Bouferguene A. Generative systems in the architecture, engineering and construction industry: a systematic review and analysis. *Int J Archit Comput*. Sep 2021;19(3):226-249. [doi: [10.1177/1478077120934126](https://doi.org/10.1177/1478077120934126)]
9. Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MAA, Dwivedi YK. A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. *J Innov Knowl*. Jan 2023;8(1):100333. [doi: [10.1016/j.jik.2023.100333](https://doi.org/10.1016/j.jik.2023.100333)]
10. Younis HA, Eisa TAE, Nasser M, et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. *Diagnostics (Basel)*. Jan 4, 2024;14(1):109. [doi: [10.3390/diagnostics14010109](https://doi.org/10.3390/diagnostics14010109)] [Medline: [38201418](https://pubmed.ncbi.nlm.nih.gov/38201418/)]
11. Ruksakulpiwat S, Thornthip S, Niyomyart A, et al. A systematic review of the application of artificial intelligence in nursing care: where are we, and what's next? *J Multidiscip Healthc*. 2024;17:1603-1616. [doi: [10.2147/JMDH.S459946](https://doi.org/10.2147/JMDH.S459946)] [Medline: [38628616](https://pubmed.ncbi.nlm.nih.gov/38628616/)]
12. Bitkina OV, Park J, Kim HK. Application of artificial intelligence in medical technologies: a systematic review of main trends. *Digit Health*. 2023;9. [doi: [10.1177/20552076231189331](https://doi.org/10.1177/20552076231189331)] [Medline: [37485326](https://pubmed.ncbi.nlm.nih.gov/37485326/)]
13. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. *Clin Med (Lond)*. May 2023;23(3):278-279. [doi: [10.7861/clinmed.2023-0078](https://doi.org/10.7861/clinmed.2023-0078)] [Medline: [37085182](https://pubmed.ncbi.nlm.nih.gov/37085182/)]
14. Gandomani HS. ChatGPT in medical education: how we can use in medical education: challenges and opportunities. *J Multidiscip Care*. 2023;12(1):1-2. [doi: [10.34172/jmdc.1232](https://doi.org/10.34172/jmdc.1232)]
15. Sani I. Enhancing medical education with ChatGPT: a promising tool for the future. *Can J Med*. Apr 1, 2024;6(1):1-4. [doi: [10.33844/cjm.2024.6032](https://doi.org/10.33844/cjm.2024.6032)]
16. Hosseini M, Gao CA, Liebovitz DM, et al. An exploratory survey about using ChatGPT in education, healthcare, and research. *PLoS One*. 2023;18(10):e0292216. [doi: [10.1371/journal.pone.0292216](https://doi.org/10.1371/journal.pone.0292216)] [Medline: [37796786](https://pubmed.ncbi.nlm.nih.gov/37796786/)]
17. Guo AA, Li J. Harnessing the power of ChatGPT in medical education. *Med Teach*. Sep 2023;45(9):1063. [doi: [10.1080/0142159X.2023.2198094](https://doi.org/10.1080/0142159X.2023.2198094)] [Medline: [37036161](https://pubmed.ncbi.nlm.nih.gov/37036161/)]
18. Jeblick K, Schachtner B, Dextl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*. May 2024;34(5):2817-2825. [doi: [10.1007/s00330-023-10213-1](https://doi.org/10.1007/s00330-023-10213-1)] [Medline: [37794249](https://pubmed.ncbi.nlm.nih.gov/37794249/)]
19. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLoS One*. 2023;18(8):e0290691. [doi: [10.1371/journal.pone.0290691](https://doi.org/10.1371/journal.pone.0290691)] [Medline: [37643186](https://pubmed.ncbi.nlm.nih.gov/37643186/)]
20. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. Feb 2023;2(2):e0000198. [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
21. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. Feb 8, 2023;9:e45312. [doi: [10.2196/45312](https://doi.org/10.2196/45312)] [Medline: [36753318](https://pubmed.ncbi.nlm.nih.gov/36753318/)]
22. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. *BJOG*. Feb 2024;131(3):378-380. [doi: [10.1111/1471-0528.17641](https://doi.org/10.1111/1471-0528.17641)] [Medline: [37604703](https://pubmed.ncbi.nlm.nih.gov/37604703/)]
23. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. *J Med Educ Curric Dev*. 2024;11. [doi: [10.1177/23821205241238641](https://doi.org/10.1177/23821205241238641)] [Medline: [38487300](https://pubmed.ncbi.nlm.nih.gov/38487300/)]
24. Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B. Appraisal of ChatGPT's aptitude for medical education: comparative analysis with third-year medical students in a pulmonology examination. *JMIR Med Educ*. Jul 23, 2024;10:e52818. [doi: [10.2196/52818](https://doi.org/10.2196/52818)] [Medline: [39042876](https://pubmed.ncbi.nlm.nih.gov/39042876/)]
25. Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. *J Med Syst*. Aug 15, 2023;47(1):86. [doi: [10.1007/s10916-023-01961-0](https://doi.org/10.1007/s10916-023-01961-0)] [Medline: [37581690](https://pubmed.ncbi.nlm.nih.gov/37581690/)]
26. Alfertshofer M, Hoch CC, Funk PF, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. *Ann Biomed Eng*. Jun 2024;52(6):1542-1545. [doi: [10.1007/s10439-023-03338-3](https://doi.org/10.1007/s10439-023-03338-3)] [Medline: [37553555](https://pubmed.ncbi.nlm.nih.gov/37553555/)]

27. Prova Nacional de Acesso à Formação Especializada 2023 Perguntas Frequentes. ACSS. 2023. URL: https://www.acss.min-saude.pt/wp-content/uploads/2018/09/FAQ_PNA2023.pdf [Accessed 2024-08-01]
28. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. Oct 1, 2023;13(1):16492. [doi: [10.1038/s41598-023-43436-9](https://doi.org/10.1038/s41598-023-43436-9)] [Medline: [37779171](https://pubmed.ncbi.nlm.nih.gov/37779171/)]
29. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. Jun 1947;12(2):153-157. [doi: [10.1007/BF02295996](https://doi.org/10.1007/BF02295996)] [Medline: [20254758](https://pubmed.ncbi.nlm.nih.gov/20254758/)]
30. Edwards AL. Note on the correction for continuity in testing the significance of the difference between correlated proportions. *Psychometrika*. Sep 1948;13(3):185-187. [doi: [10.1007/BF02289261](https://doi.org/10.1007/BF02289261)] [Medline: [18885738](https://pubmed.ncbi.nlm.nih.gov/18885738/)]
31. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. *Medicine (Baltimore)*. Aug 11, 2023;102(32):e34673. [doi: [10.1097/MD.00000000000034673](https://doi.org/10.1097/MD.00000000000034673)] [Medline: [37565917](https://pubmed.ncbi.nlm.nih.gov/37565917/)]
32. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. *BMJ Health Care Inform*. Dec 11, 2023;30(1):e100815. [doi: [10.1136/bmjhci-2023-100815](https://doi.org/10.1136/bmjhci-2023-100815)] [Medline: [38081765](https://pubmed.ncbi.nlm.nih.gov/38081765/)]
33. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: performance of ChatGPT on a PES medical examination. *Cardiol J*. 2024;31(3):442-450. [doi: [10.5603/cj.97517](https://doi.org/10.5603/cj.97517)] [Medline: [37830257](https://pubmed.ncbi.nlm.nih.gov/37830257/)]
34. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Sci Rep*. Nov 22, 2023;13(1):20512. [doi: [10.1038/s41598-023-46995-z](https://doi.org/10.1038/s41598-023-46995-z)] [Medline: [37993519](https://pubmed.ncbi.nlm.nih.gov/37993519/)]
35. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): promising horizons for AI in clinical medicine. *Clin Pract*. Nov 20, 2023;13(6):1460-1487. [doi: [10.3390/clinpract13060130](https://doi.org/10.3390/clinpract13060130)] [Medline: [37987431](https://pubmed.ncbi.nlm.nih.gov/37987431/)]
36. Kaneda Y, Takahashi R, Kaneda U, et al. Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese nursing examination. *Cureus*. Aug 2023;15(8):e42924. [doi: [10.7759/cureus.42924](https://doi.org/10.7759/cureus.42924)] [Medline: [37667724](https://pubmed.ncbi.nlm.nih.gov/37667724/)]
37. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. *JMIR Med Educ*. Sep 28, 2023;9:e48039. [doi: [10.2196/48039](https://doi.org/10.2196/48039)] [Medline: [37768724](https://pubmed.ncbi.nlm.nih.gov/37768724/)]
38. Kruger J, Wirtz D, Miller DT. Counterfactual thinking and the first instinct fallacy. *J Pers Soc Psychol*. May 2005;88(5):725-735. [doi: [10.1037/0022-3514.88.5.725](https://doi.org/10.1037/0022-3514.88.5.725)] [Medline: [15988871](https://pubmed.ncbi.nlm.nih.gov/15988871/)]
39. Coutinho MVC, Thomas J, Fredricks-Lowman I, Alkaabi S, Couchman JJ. Unskilled and unaware: second-order judgments increase with miscalibration for low performers. *Front Psychol*. 2024;15:1252520. [doi: [10.3389/fpsyg.2024.1252520](https://doi.org/10.3389/fpsyg.2024.1252520)] [Medline: [38952836](https://pubmed.ncbi.nlm.nih.gov/38952836/)]
40. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. *JMIR Med Educ*. Sep 4, 2023;9:e46482. [doi: [10.2196/46482](https://doi.org/10.2196/46482)] [Medline: [37665620](https://pubmed.ncbi.nlm.nih.gov/37665620/)]
41. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. *Br J Ophthalmol*. Oct 2024;108(10):1379-1383. [doi: [10.1136/bjo-2023-324091](https://doi.org/10.1136/bjo-2023-324091)]
42. Huang RS, Lu KJQ, Meaney C, Kempainen J, Punnett A, Leung FH. Assessment of resident and AI chatbot performance on the University of Toronto Family Medicine Residency Progress Test: comparative study. *JMIR Med Educ*. Sep 19, 2023;9:e50514. [doi: [10.2196/50514](https://doi.org/10.2196/50514)] [Medline: [37725411](https://pubmed.ncbi.nlm.nih.gov/37725411/)]
43. Wong RSY, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. *JMIR Med Educ*. Nov 21, 2023;9:e47274. [doi: [10.2196/47274](https://doi.org/10.2196/47274)] [Medline: [37988149](https://pubmed.ncbi.nlm.nih.gov/37988149/)]
44. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. *J Educ Eval Health Prof*. 2024;21(6):6. [doi: [10.3352/jeehp.2024.21.6](https://doi.org/10.3352/jeehp.2024.21.6)] [Medline: [38486402](https://pubmed.ncbi.nlm.nih.gov/38486402/)]

Abbreviations

- AI:** artificial intelligence
- LLM:** large language model
- MCQ:** multiple-choice question
- PNA:** Prova Nacional de Acesso à Formação Especializada
- USMLE:** United States Medical Licensing Examination

Edited by Blake Lesselroth; peer-reviewed by Bill Thies, Luciano Antonio Digiampietri, Raffaele Pellegrino; submitted 05.08.2024; final revised version received 30.11.2024; accepted 12.12.2024; published 05.03.2025

Please cite as:

Prazeres F

ChatGPT's Performance on Portuguese Medical Examination Questions: Comparative Analysis of ChatGPT-3.5 Turbo and ChatGPT-4o Mini

JMIR Med Educ 2025;11:e65108

URL: <https://mededu.jmir.org/2025/1/e65108>

doi: [10.2196/65108](https://doi.org/10.2196/65108)

© Filipe Prazeres. Originally published in JMIR Medical Education (<https://mededu.jmir.org>), 05.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on <https://mededu.jmir.org/>, as well as this copyright and license information must be included.