ACTA MÉDICA PORTUGUESA
A Revista Científica da Ordem dos Médicos

Performance of ChatGPT in the Portuguese National Residency Access Examination

Desempenho do ChatGPT na Prova Nacional de Acesso

Gonçalo FERRAZ-COSTA✉*[1,2,3], Mafalda GRINÉ*[1,3], Manuel OLIVEIRA-SANTOS[1,3,4], Rogério TEIXEIRA[1,2,4]
Acta Med Port (In Press) ▪ https://doi.org/10.20344/amp.22506

ABSTRACT

ChatGPT, a language model developed by OpenAI, has been tested in several medical board examinations. This study aims to evaluate the performance of ChatGPT on the Portuguese National Residency Access Examination, a mandatory test for medical residency in Portugal. The study specifically compares the capabilities of ChatGPT versions 3.5 and 4o across five examination editions from 2019 to 2023. A total of 750 multiple-choice questions were submitted to both versions, and their answers were evaluated against the official responses. The findings revealed that ChatGPT 4o significantly outperformed ChatGPT 3.5, with a median examination score of 127 compared to 106 (p = 0.048). Notably, ChatGPT 4o achieved scores within the top 1% in two examination editions and exceeded the median performance of human candidates in all editions. Additionally, ChatGPT 4o's scores were high enough to qualify for any specialty. In conclusion, ChatGPT 4o can be a valuable tool for medical education and decision-making, but human oversight remains essential to ensure safe and accurate clinical practice.

Keywords: Artificial Intelligence; Clinical Competence; Educational Measurement; Internship and Residency; Portugal

RESUMO

O ChatGPT, um modelo de linguagem desenvolvido pela OpenAI, foi testado em vários exames de acesso à profissão médica. Este estudo tem como objetivo avaliar o desempenho do ChatGPT na Prova Nacional de Acesso à Formação Especializada, um exame obrigatório para o início do internato médico em Portugal. O estudo compara especificamente as capacidades das versões 3.5 e 4o do ChatGPT em cinco edições do exame, de 2019 a 2023. Um total de 750 perguntas de escolha múltipla foram submetidas a ambas as versões, e as suas respostas foram avaliadas em comparação com as respostas oficiais. Os resultados revelam que o ChatGPT 4o superou significativamente o ChatGPT 3.5, com uma pontuação mediana de 127 em comparação com 106 (p = 0,048). Notavelmente, o ChatGPT 4o obteve pontuações dentro do *top* 1% em duas edições do exame e superou o desempenho mediano dos candidatos humanos em todas as edições. Além disso, as pontuações do ChatGPT 4o foram suficientemente elevadas para se qualificar para qualquer especialidade. Em conclusão, o ChatGPT 4o pode ser uma ferramenta valiosa para a educação médica e tomada de decisões, mas a supervisão humana continua a ser essencial para garantir uma prática clínica segura e precisa.

Palavras-chave: Avaliação Educacional; Competência Clínica; Inteligência Artificial; Internato e Residência; Portugal

INTRODUCTION

Artificial intelligence (AI) has seen rapid advancements, notably with the development of sophisticated large language models like ChatGPT, a tool that has demonstrated potential across diverse fields, including healthcare and education.[1] ChatGPT, developed by OpenAI, is a web-based language model that became widely accessible in 2022.[1] Since then, researchers have been assessing its capabilities in various settings, including its proficiency in answering medical licensing examination questions.[2,3] Multiple versions of ChatGPT have been released, each with improved capacities: ChatGPT-3.5, which operates solely on text-based inputs, and ChatGPT-4o, which includes the additional capability of processing images, enhancing its potential for answering complex questions.[4]

Previous studies have investigated ChatGPT's worldwide performance in medical licensing examinations, reporting mixed results. A recent systematic review and meta-analysis, for instance, examined 45 studies and found that ChatGPT-4o attained an average performance of 81% (95% CI: 78% - 84%), which was significantly higher than ChatGPT-3.5's average of 58% (95% CI: 53% - 63%).[2] This marked improvement suggests that each version of ChatGPT is becoming increasingly competent in specialized domains, such as medicine.

The Portuguese National Residency Access Examination (*Prova Nacional de Acesso*, or PNA) is a high-stakes, competitive examination required for medical graduates in Portugal to access residency programs.[5] Since 2019, the PNA has consisted of 150 multiple-choice questions with a single best-answer format, covering clinical, diagnostic, therapeutic, and epidemiological knowledge, presented as clinical vignettes. Candidates have 240 minutes, divided into two 120-minute sessions, to complete the examination. Approximately 2200 to 2500 candidates take the PNA each year, with scores

* Shared co-authorship.
1. Cardiology Department. Unidade Local de Saúde de Coimbra. Coimbra. Portugal.
2. Faculdade de Medicina. Universidade de Coimbra. Coimbra. Portugal.
3. Coimbra Institute for Clinical and Biomedical Research (iCBR). Coimbra. Portugal
4. Digital Health Study Group, Sociedade Portuguesa de Cardiologia. Lisboa. Portugal.
✉ **Autor correspondente:** Gonçalo Costa. gnfcosta.93@gmail.com
Recebido/*Received*: 27/10/2024 - **Aceite/*Accepted*:** 21/11/2024 - **Publicado *Online*/*Published Online*:** 20/12/2024

determining eligibility for residency placements. To date, no study has evaluated ChatGPT's performance on the PNA.

This study aims to evaluate ChatGPT's performance on the PNA, specifically comparing ChatGPT-3.5 and ChatGPT-4o across five examination editions (2019 - 2023) and assessing their potential to match into different medical specialties alongside human candidates.

METHODS

Procedures

For this study, we used the ChatGPT Plus plan, which provides access to both models without the usage restrictions of the free version. While the plan allows up to 80 interactions per 3-hour session, the examinations were divided into two sessions of 75 questions each in the same chat window for each examination.

All questions from the publicly available 2019 to 2023 editions of the PNA were retrieved. ChatGPT-3.5 lacks the capability to process image inputs, whereas the newer version, ChatGPT-4o, includes this functionality. Nevertheless, every question was included in this analysis for both versions. For image-containing questions, only the text was submitted to ChatGPT-3.5, while both text and images were submitted to ChatGPT-4o. Each examination question was manually entered in sequence, following the order of the actual examination (version A). For each examination, a new chat window was created to prevent memory retention bias. Questions were submitted exactly as they appeared in the original examination, without any introductory prompts. The responses given by ChatGPT were marked according to the official final answer key.

Data analysis

We analysed the scores from each of the five PNA editions (2019 - 2023) for both ChatGPT-3.5 and ChatGPT-4o, treating each year as an independent case for comparison. The performances of ChatGPT-3.5 and ChatGPT-4o were then compared using the Mann-Whitney U test, given the small sample size. We also identified the lowest-scoring candidates who successfully matched into a residency program each year and compared their scores with ChatGPT's performance. A sensitivity analysis was conducted by excluding all questions containing images.

RESULTS

In total, 150 questions from each of the five PNA exam editions (2019 - 2023) were retrieved and evaluated, resulting in a total of 750 questions. The individual exam performance results of ChatGPT-3.5 and ChatGPT-4o across different years are summarized in Fig. 1. ChatGPT-4o showed significantly higher performance compared to ChatGPT-3.5, with median exam scores of 106 [Interquartile Range (IQR): 99 - 114.5] for ChatGPT-3.5 and 127 (IQR: 122.5 – 134) for ChatGPT-4o (p = 0.048). The maximum possible score for each exam is 150, as each question is worth 1 point.

In the total question cohort, there were nine questions that included figures. ChatGPT-3.5 answered seven of these questions correctly, while ChatGPT-4o correctly answered six. Excluding questions with figures, ChatGPT-4o maintained a higher performance, with a median score of 124 (IQR: 122.5 - 127.5), compared to a median score of 106 (IQR: 106 – 111) for ChatGPT-3.5 (p = 0.016).

Regarding the overall medical graduate cohort, ChatGPT 4o surpassed the median score for each exam edition, while ChatGPT 3.5 performed below the 50th percentile in one examination version. Additionally, ChatGPT 4o ranked within the top 1% in two exam editions (2023 and 2019), achieving the highest score. Table 1 illustrates the comparison of ChatGPT's scores with the lowest-scoring candidates into various medical specialties across the different examination editions. Competitive specialties, such as dermatology, ophthalmology, and plastic surgery, required higher minimum scores, ranging from 110 to 129. ChatGPT-4o's scores consistently exceeded these thresholds in all years evaluated. ChatGPT-3.5, in contrast, fell below the minimum matching score in some instances. Less competitive specialties, including family medicine, internal medicine, and clinical pathology, presented the lowest scores. In these cases, both ChatGPT-3.5 and ChatGPT-4o performed above the required score, and many of these specialties had unfilled positions in certain years.

DISCUSSION

The major findings of this study are: (a) ChatGPT-4o demonstrated exceptional performance on the PNA, surpassing all medical graduate candidates in two exam editions. For highly competitive specialties, such as dermatology, ophthalmology, and plastic surgery – those with the highest minimum score – ChatGPT-4o's performance would have secured a match into any specialty; (b) This latest version of the AI program outperformed its predecessor, ChatGPT-3.5. The substantial improvement in ChatGPT-4o's score is attributed to the mitigation of previously identified weaknesses, resulting in a more proficient model in the medical domain.[3,6] A systematic review and meta-analysis of 45 studies assessing different versions

of ChatGPT in medical licensing examinations found that ChatGPT-4o achieved an overall performance of 81% (95% CI: 78% - 84%), significantly surpassing the 58% (95% CI: 53% - 63%) performance of ChatGPT-3.5, supporting our findings.[2]

However, ChatGPT has documented areas of limited performance. First, there is a documented inconsistency in test-retest results, which raise concerns about its reliability.[7] Second, an analysis of ChatGPT-4o's performance on the United States Medical Licensing Examination revealed a tendency to make errors on questions requiring knowledge transfer skills, indicating a potential deficit in abstract thinking. Nonetheless, the PNA is a clinical vignette-based exam, and ChatGPT's exceptional performance in our study contrasts with previous data. Third, ChatGPT has shown better performance in English-speaking countries compared to non-English-speaking ones, which did not appear to affect the AI's results in PNA.[2] Fourth, recent findings on ChatGPT-4o's performance in specialized examinations like the Adult Clinical Cardiology Self-Assessment Program (namely when replying to a question bank that includes imaging and is used by the American Board of Internal Medicine on their general cardiology board exam), showed a 73.9% accuracy rate for text-only questions but a lower 55.3% for image-based questions, particularly electrocardiograms.[4] In our study, only nine questions contained images, with ChatGPT-3.5 performing marginally better than ChatGPT-4o on these (seven *versus* six questions answered correctly). Although this small sample limits the extent to which conclusions can be drawn about ChatGPT-4o's image-analysis capacity, it underscores an important consideration: the structured nature of certain questions may allow accurate answers based solely on the clinical vignette information, potentially bypassing the need for image interpretation.

Several hypotheses may explain why neither human candidates nor ChatGPT achieved the maximum score of 150 on the PNA. The PNA is intentionally rigorous, designed to assess a broad spectrum of medical knowledge and clinical reasoning skills. Its complexity and high standards may naturally prevent both AI and human candidates from achieving perfect scores. Additionally, the nuances of clinical scenarios presented in the PNA may challenge both groups; certain questions require advanced clinical inference and contextual judgment, which can pose difficulties for human candidates and AI alike. Moreover, some questions may contain inherent ambiguities or complex phrasing, adding another layer of difficulty. For human candidates, cognitive load and fatigue throughout the examination may further impact performance, an aspect not affecting the AI.

Our study suggests that ChatGPT exhibits strong medical knowledge. As an interactive resource, ChatGPT consistently provides correct answers and effectively clarifies why alternatives are incorrect, supporting deeper understanding and active learning. It is important to note that the PNA's structured format, featuring straightforward questions and answers, avoids the complexity found in real-world clinical scenarios.[3] This design choice enhances clarity and minimizes potential disputes over correct responses, thus streamlining the ranking process. Consequently, our findings do not assess ChatGPT's effectiveness in clinical decision-making within practical settings. While our results highlight ChatGPT's capabilities in a controlled testing environment, its performance may not seamlessly transfer to dynamic clinical contexts, where adaptive reasoning and contextual judgment are essential. This limitation underscores the critical role of human oversight in real-world applications and the need for further studies to assess ChatGPT's reliability and adaptability in actual clinical scenarios. Ultimately, the responsibility for clinical management must reside with qualified healthcare professionals, as exclusive reliance on ChatGPT's responses for patient care remains, at this stage, ethically unsound. Future studies could also focus on integrating ChatGPT into clinical practice, assessing how AI can collaborate with physicians to enhance decision-making without replacing human judgment. Additionally, examining ChatGPT's performance across different cultural and linguistic contexts would provide a better understanding of its applicability in medical examinations from various regions and languages.

Another limitation of this study is the potential for ChatGPT to have been exposed to publicly available PNA questions from earlier editions (2019 - 2020), as both models have been trained with data up to September 2021. While neither ChatGPT-3.5 nor ChatGPT-4o have real-time browsing capabilities and therefore could not access more recent examination questions online, we acknowledge this potential bias. Additionally, the browsing-enabled ChatGPT-4 Turbo (released in November 2023) was not used in this study, further minimizing the likelihood of direct access to PNA questions. Evaluating the performance of ChatGPT-4 Turbo on the PNA would be an interesting focus for future studies.

Finally, another area for future research would be the development of a ChatGPT-enabled mock examination platform to evaluate the AI's capacity for generating diverse, high-quality medical questions, particularly multiple choice questions.[8] Such a tool could enhance educators' productivity by quickly creating question banks and support self-directed learning for students by providing accessible, examination-style practice. Future studies could assess the accuracy, relevance, and educational impact of these AI-generated questions to ensure alignment with clinical and educational standards in medical training.

CONCLUSION

ChatGPT-4o demonstrated excellent performance on the PNA, consistently outperforming the average examination participant and achieving high enough scores to match into any specialty.

AUTHOR CONTRIBUTIONS

GFC, MG: Study design, data acquisition and analysis, drafting of the manuscript.

MOS, RPT: Critical review of the manuscript.

All authors approved the final version to be published.

PROTECTION OF HUMANS AND ANIMALS

The authors declare that the procedures were followed according to the regulations established by the Clinical Research and Ethics Committee and to the Helsinki Declaration of the World Medical Association updated in October 2024.

DATA CONFIDENTIALITY

The authors declare having followed the protocols in use at their working center regarding patients' data publication.

COMPETING INTERESTS

MOS received payment or honoraria from Novartis, Bial, Biotronik and Boston Scientific for lectures, presentations, speakers' bureaus, manuscript writing or educational events; received support for attending meetings and/or travel from Viatris, Terumo, Medinfar, Medtronic and Abbot.

All other authors have declared that no competing interests exist.

FUNDING SOURCES

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Berşe S, Akça K, Dirgar E, Kaplan Serin E. The role and potential contributions of the artificial intelligence language model ChatGPT. Ann Biomed Eng 2024;52:130-3.
2. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res. 2024;26:e60807.
3. Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, et al. Pure wisdom or potemkin villages? A comparison of chatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. JMIR Med Educ. 2024;10:e51148.
4. Malik A, Madias C, Wessler BS. Performance of ChatGPT-4o in the adult clinical cardiology self-assessment program. Eur Hear J - Digit Heal. 2024:ztae077.
5. Ribeiro JC, Villanueva T. The new medical licensing examination in Portugal. Acta Med Port. 2018;31:293-4.
6. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młynczak. Evaluation of the performance of GPT-3.5 and GPT-4 on the polish medical final examination. Sci Reports. 2023;13:1-13.
7. Alexandrou M, Mahtani AU, Rempakos A, Mutlu D, Ogaili AA, Gill GS, et al. Performance of ChatGPT on ACC/SCAI interventional cardiology certification simulation exam. JACC Cardiovasc Interv. 2024;17:1292-3.
8. Indran IR, Paranthaman P, Gupta N, Mustafa N. Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using Chat GPT. Med Teach. 2024;46:1021-6.

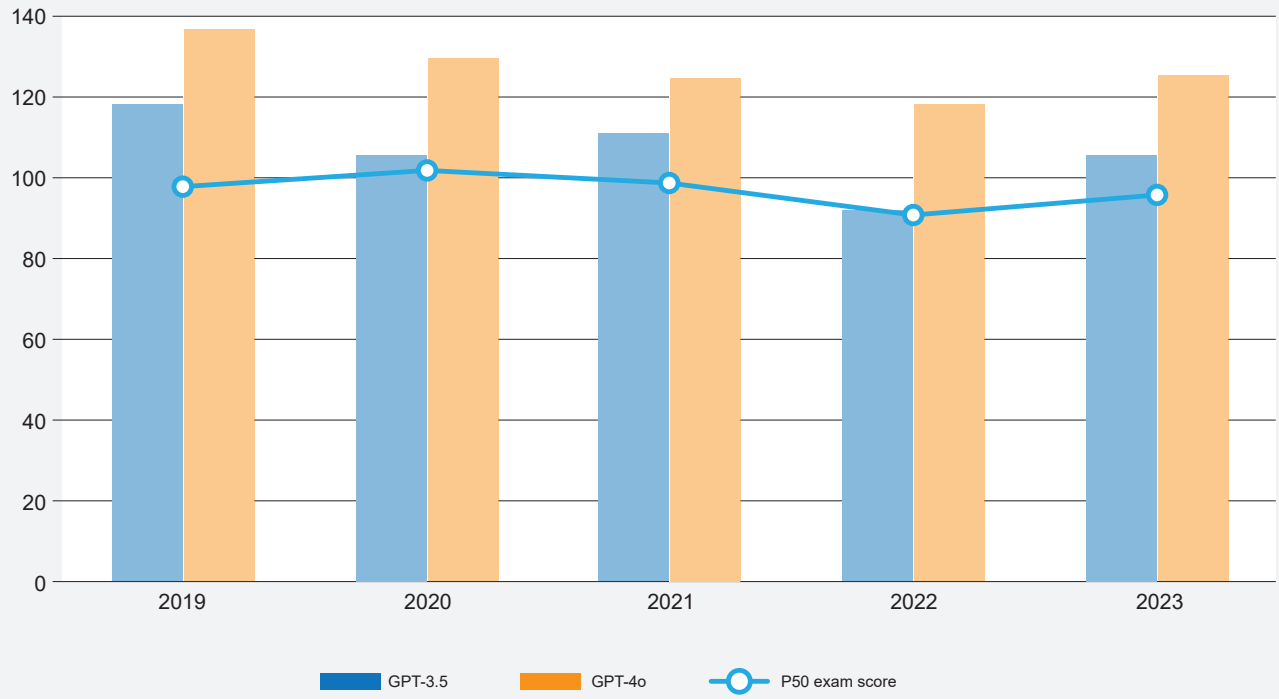

Figure 1 – Performance of ChatGPT on the Portuguese National Residency Access Examination. The figure shows the scores for ChatGPT-3.5 and ChatGPT-4o across five editions of the exam, with 150 questions per exam, totalling 750 questions evaluated.
PNA: Portuguese National Residency Access Exam (*Prova Nacional de Acesso*); P50: median score for each exam edition.

Table 1 – ChatGPT residency matching results for the 2019 - 2022 Portuguese National Residency Access Examination editions. The results for ChatGPT-3.5 are shown on the left, while those for ChatGPT-4o are on the right. A green cell indicates that the score achieved by ChatGPT in that year was sufficient to match into the listed specialty, based on the score of the last candidate who successfully matched into that specialty. A red cell indicates that the score was not high enough to qualify for a match.

		ChatGPT-3.5				ChatGPT-4o			
	Exam year	2019	2020	2021	2022	2019	2020	2021	2022
	Chat GPT exam score	118	106	111	92	137	130	125	118
Acceptance score threshold by speciality	Pathology	76	78	84	NF	76	78	84	NF
	Anesthesiology	110	114	109	97	110	114	109	97
	Angiology and vascular surgery	115	114	110	106	115	114	110	106
	Cardiology	112	115	107	97	112	115	107	97
	Pediatric cardiology	99	107	101	92	99	107	101	92
	Cardiac surgery	104	110	105	91	104	110	105	91
	General surgery	93	103	90	74	93	103	90	74
	Maxillo-facial surgery	107	112	102	96	107	112	102	96
	Pediatric surgery	103	107	108	92	103	107	108	92
	Plastic surgery	122	123	119	111	122	123	119	111
	Thoracic surgery	107	111	103	94	107	111	103	94
	Dermatology	123	129	121	110	123	129	121	110
	Infeccious disease	85	80	63	NF	85	80	63	NF
	Endocrinology	109	110	108	103	109	110	108	103
	Stomatology	77	59	55	NF	77	59	55	NF
	Clinical pharmacology	59	72	NF	NF	59	72	NF	NF
	Gastroenterology	115	121	112	104	115	121	112	104
	Medical genetics and genomics	78	85	89	NF	78	85	89	NF
	Obstetrics and gynecology	105	114	106	94	105	114	106	94
	Hematology	76	85	61	NF	76	85	61	NF
	Allergy and immunology	100	106	96	84	100	106	96	84
	Transfusion medicine	63	NF	NF	NF	63	NF	NF	NF
	Sports medicine	110	NA	109	101	110	NA	109	101
	Physical medicine and rehabilitation	99	109	94	92	99	109	94	92
	Family medicine	63	NF	NF	NF	63	NF	NF	NF
	Critical care medicine	89	88	49	NF	89	88	49	NF
	Internal medicine	57	NF	NF	NF	57	NF	NF	NF
	Forensic medicine	84	78	86	NF	84	78	86	NF
	Nuclear medicine	107	104	105	93	107	104	105	93
	Occupational medicine	87	102	95	82	87	102	95	82
	Nephrology	100	105	94	90	100	105	94	90
	Neurosurgery	98	111	103	95	98	111	103	95
	Neurology	105	108	97	89	105	108	97	89
	Neuroradiology	108	115	109	95	108	115	109	95
	Ophthalmology	122	125	120	111	122	125	120	111
	Medical oncology	86	87	50	NF	86	87	50	NF
	Orthopedic surgery	100	110	102	95	100	110	102	95
	Otolaryngology	114	117	111	102	114	117	111	102
	Clinical pathology	57	NF	NF	NF	57	NF	NF	NF
	Pediatrics	99	104	99	84	99	104	99	84
	Pulmonology	101	104	97	89	101	104	97	89
	Psychiatric	93	95	98	87	93	95	98	87
	Child and adolescent pshychiatry	98	103	98	88	98	103	98	88
	Radiology	111	113	112	97	111	113	112	97
	Radiation oncology	88	82	69	NF	88	82	69	NF
	Rheumatology	106	110	99	95	106	110	99	95
	Public health	65	53	NF	NF	65	53	NF	NF
	Urology	111	113	110	99	111	113	110	99

NA: not available; U: unfilled positions.