RBEM Revista Brasileira de Educação Médica

ORIGINAL ARTICLE

DOI: https://doi.org/10.1590/1981-5271v49.4-2025-0026.ING

Medical students versus chatbots in solving a medical test: a comparative study

Estudantes de medicina versus chatbots *na resolução de um teste médico: um estudo comparativo*

Douglas Augusto Pasqual La Maison[1] douglaslamaison@gmail.com
Amanda Carolina Fonseca da Silva[1] amandacfsr1@gmail.com
Bruno Ferreira Glir[1] brunofglir@gmail.com
Ari Ojeda Ocampo Moré[1] darimore@hotmail.com

ABSTRACT

Introduction: Medical education is assessed globally by National Medical Licensing Examinations and Progress Tests. Recently, generative artificial intelligence (GAI), including Large Language Models (LLMs) and chatbots, have shown potential as a support tool in medical education, while studies have indicated that their performance is comparable to that of medical students in exams.

Objective: To evaluate the effectiveness of different LLM chatbots in solving questions from the NAPISUL-II Progress Test, comparing their performance with that of students in the last semester of medical school, with a view to the pedagogical applicability of LLMs in medical education.

Method: Data from the 2023 NAPISUL-II Progress Test, administered to medical students from southern Brazil were used. Four chatbots (ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard) answered the test three times. The responses were compared with those of students in the last semester of the course, according to the medical areas inherent to the PT and reasoning categories defined by the authors. Data analysis was descriptive, using accuracy rates to compare the groups.

Result: Chatbots had an average of 80.81% accuracy in 116 questions, while medical students obtained an average of 62%. Bing AI had the best performance, with 87.35% of accuracy, followed by ChatGPT 4.0 with 85.86%. In specific areas, ChatGPT 4.0 stood out in surgical clinic (88.33%) and internal medicine (86.67%), while Bing AI had the best performance among chatbots in basic sciences (94.44%), gynecology and obstetrics (94.74%), public health (83.33%) and pediatrics (82.46%). Bard, despite presenting the greatest range of variation between areas, and ChatGPT 3.5, inferior to the other chatbots, also outperformed students in all categories.

Conclusion: All chatbots assessed outperformed medical students in the twelfth semester of the NAPISUL-II block in the Progress Test. Bing stood out with the best performance among the chatbots, followed by Chat GPT-4. This indicates that chatbots have the potential to be an auxiliary tool in medical education, although more studies are needed to evaluate the possibilities of this tool in this context.

Keywords: Medical Education; Artificial Intelligence; Academic Performance.

RESUMO

Introdução: *A educação médica é avaliada globalmente por Exames Nacionais de Licenciamento Médico e Testes de Progresso. Recentemente, a inteligência artificial generativa (generative artificial intelligence – GAI), incluindo Large Language Models (LLM) e chatbots, mostrou potencial como ferramenta de suporte na educação médica, ao passo que estudos indicaram que seu desempenho é comparável ao de estudantes de Medicina em exames.*

Objetivo: *Este estudo teve como objetivo avaliar a efetividade de diferentes chatbots de LLM na resolução de questões do Teste de Progresso Napisul-II, comparando seu desempenho com o de estudantes do último semestre de Medicina, com vista à aplicabilidade pedagógica de LLM no ensino médico.*

Método: *Foram utilizados dados do Teste de Progresso Napisul-II de 2023, aplicados em estudantes de Medicina do Sul do Brasil. Quatro chatbots (ChatGPT 3.5, ChatGPT 4.0, Bing AI e Bard) responderam três vezes ao teste. As respostas foram comparadas com as dos estudantes do último semestre do curso, de acordo com as áreas médicas inerentes ao TP e as categorias de raciocínio definidas pelos autores. A análise dos dados foi descritiva, utilizando taxas de acerto para comparar os grupos.*

Resultado: *Os chatbots tiveram uma média de 80,81% de acertos em 116 questões, enquanto os estudantes de Medicina obtiveram uma média de 62%. O Bing AI teve a melhor performance, com 87,35% de acertos, seguido pelo ChatGPT 4.0 com 85,86%. Em áreas específicas, o ChatGPT 4.0 destacou-se em clínica cirúrgica (88,33%) e clínica médica (86,67%), enquanto o Bing AI teve o melhor desempenho entre os chatbots em ciências básicas (94,44%), ginecologia e obstetrícia (94,74%), saúde coletiva (83,33%) e pediatria (82,46%). O Bard, apesar de apresentar a maior amplitude de variação entre as áreas, e o ChatGPT 3.5, inferior aos outros chatbots, também superaram os estudantes em todas as categorias.*

Conclusão: *Todos os chatbots estudados tiveram um desempenho superior ao dos discentes de Medicina do 12º semestre do bloco Napisul-II no Teste de Progresso. O Bing destacou-se com o melhor desempenho entre os chatbots, seguido pelo ChatGPT 4.0. Isso indica que os chatbots têm potencial de ser uma ferramenta auxiliar na educação médica, embora sejam necessários mais estudos para avaliar as possibilidades dela nesse contexto.*

Palavras-chave: *Educação Médica; Inteligência Artificial; Desempenho Acadêmico.*

[1] Universidade Federal de Santa Catarina, Florianopolis, Santa Catarina, Brasil.

Chief Editor: Rosiane Viana Zuza Diniz. | Associate Editor: Jorge Guedes.

Received on 02/03/25; Accepted on 06/20/25. | Evaluated by double blind review process.

INTRODUCTION

Medical education is assessed through National Medical Licensing Examinations in a wide range of countries, such as the United States, Germany, and China. These exams seek to assess whether medical students' performance in medical skills is sufficient for a practice of excellence. In addition to being a tool for assessing competencies, the exams also end up serving to guide the evolution of medical education, since their results can represent the quality of universities and guide them to improve their teaching. However, these exams have board approval as their main objective, and not the evaluation of medical education[1–3]. For this purpose, a tool called the Progress Test was developed in 1970 at the University of Missouri-Kansas City School of Medicine, in the United States, and at the University of Limburg, in the Netherlands. The test consists of multiple-choice questions that assess cognitive aspects, and the same test is applied at the same time to students of all levels of medical school. The evaluation process aims to understand the correspondence between the performance of students and the year of study in which they are enrolled, in addition to analyzing the development of the same student's learning throughout the course. Thus, didactic methods are improved based on the identification of flaws in certain areas of education, visualized in the academic performance curves obtained through the results of the Progress Test[4–9].

In Brazil, the application of Progress Tests is more recent and its questions are based on the national curriculum guidelines for medical schools. Related to ABEM (Brazilian Association of Medical Education), there are 16 interinstitutional centers that bring together a certain number of medical schools to apply the test, enabling more robust analyses and comparisons. In the south of the country, there is the South II Interinstitutional Pedagogical Support Center (NAPISUL-II), consisting of six schools in Paraná and seven in Santa Catarina. In this nucleus, the progress test had its 13th edition in October 2023[10].

Recently, the use of Generative Artificial Intelligence (GAI) has become popular, which shows promise in the area of education[11–15], and there are studies in the literature involving its use to solve National Medical Licensing Exams[16,17]. GAI refers to a technology capable of producing content based on previous databases[18]. Within the universe of generative AI is the family of GPTs (Generative Pre-trained Transformers), which are Large Language Models (LLMs)[19]. LLMs, in turn, are Artificial Intelligence (AI) algorithms designed to determine the probability of certain sequences of words, taking into account the context of the words that precede them[20]. The improvement of these algorithms has been revolutionary in the field of Natural Language Processing (NLP), a type of machine learning that simulates human language[19].

Considering that language is a key factor in the field of medicine, LLMs show great promise in learning how to represent medical knowledge. However, it is evident that these tools denote generations of text that may not represent the veracity of medical science or its ethics[21]. More recently, the popularization of chatbots (ChatGPT, Bing, Bard) — interfaces for AI systems based on large, generative pre-trained language models (GPT), which produce human-like texts in response to messages sent by users, has been observed. This has made it possible for such tools, previously used only by professionals related to AI, to be used by the lay population, which brings into consideration the possibility and dangers of their use in the context of medical education[19,22,23].

Considering these possibilities, numerous studies have been developed exploring the capabilities of these tools in solving medical licensing exams, specific tests for medical specialties or other areas of health, and progress tests[16,17,24–26]. Complementary results were obtained, since ChatGPT 4.0 obtained a performance similar to that of a third-year medical student[17] in the United States Medical Licensing Examination (USMLE), in addition to having behaved like a first-year plastic surgery resident[25] and having passed the Japanese Medical Licensing Examination[16], which demonstrates its potential as an interactive support instrument for medical education.

There is a gap in the scientific literature regarding the comparison of performance between multiple LLM Chatbots and Brazilian medical students, since there are already similar studies involving students of other nationalities or limited to only one chatbot[27–30]. Although international research provides insights into the potential of LLM Chatbots in medical education, the cultural, linguistic, and especially curricular variability between different countries suggests that the results and implications of these technologies may be significantly different in Brazil.

In this context, the objective of this study is to fill this gap by evaluating the effectiveness of different LLM Chatbots in solving questions of the NAPISUL-II progress test. This involves a comparative analysis between the answers provided by these AI tools and those of students in the last semester of the medical course, to compare the level of accuracy in the answers. Such investigation will not only contribute to the understanding of the applicability of LLM Chatbots in Brazilian medical education, but will also provide valuable insights into the potential limitations and advantages of these emerging technologies in improving the quality of teaching and medical assessment.

METHOD

Context

The progress test is a model of objective test of specific knowledge of a given higher education course, being an individual evaluation strategy that is well consolidated today to demonstrate the student's learning process throughout their training. In Brazil, the Brazilian Association of Medical Education (ABEM) is responsible for the creation, expansion and application of the Progress Test (PT) in the context of undergraduate medical courses.

Data collection

This study used as a question database the PT prepared by the Southern Interinstitutional Pedagogical Support Center II (NAPISUL-II), one of the regional groups of ABEM. The edition of the test used was the 2023 edition, whose application date was October 18, 2023. The base document, from which the questions were transcribed, was sent to us by the coordination of the undergraduate medical course of one of the institutions that constitute the NAPISUL-II block. One researcher transcribed the questions into an editable document in Google Docs and, with the help of two other researchers, reviewed possible errors.

To conduct this study, proper authorization was obtained for the use of data from the results of the 2023 Progress Test for research purposes. Personal contact and subsequent communication by e-mail were made with the request for the use of the data with one of the members of the NAPISUL-II Coordination. On December 7, 2023, the email response was obtained with the authorization for the use of the data. It is important to highlight that all data used are anonymous and without individual identification of students or specific centers, being used exclusively for academic and research purposes. Additionally, it is noteworthy that these general data from NAPISUL-II used in this research, in addition to having public access, are widely disseminated in the university context.

Study Design

Initially, four chatbots were defined for analysis: ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard (Bing AI and Bard later renamed Copilot and Gemini, respectively)[31–33], because they were the most relevant at the time of collection[34]. The chatbots were accessed through their public web interfaces. The tests of all chatbots were carried out between November 14 and December 2, 2023. All series of questions were preceded by the following standard command: "From now on, multiple choice questions will be sent and you must select only one correct alternative. Be punctual, mark the alternative without explaining the reason", in order to follow the prompt engineering model for "direct one-shot"[35] medical questions. Answers that indicated more than one alternative as correct were considered wrong. Given the initial command, the questions were sent individually, being separated into groups of 20 questions, delimited by areas of medical knowledge pre-defined by the PT, with the objective of facilitating the organization and preventing the chatbot from neglecting the original command due to memory bias[20,36]. At the end of each block, a new chat was opened and the initial command sent again. All PT questions were forwarded, with the exception of canceled questions(2), with images(1) or tables(1) given that, at the time of collection, there were different limitations in the processing capacity and visual interpretation of chatbots[37]. After completing this stage, all questions and their respective answers were recorded in a document on the Google Docs platform, through screenshots and links that refer to the block of answered questions in the area of medical knowledge presented. This entire process was carried out three times, as in a previous similar study[38], so that the reliability of the obtained data was greater. This multiple-essay approach to each question is critical when testing LLMs, given the inherent variability in their responses, ensuring a more robust and representative assessment of their potential performance. After that, a table was created using Google Sheets, in which the answers given by the chatbots were evidenced next to the official PT template, for later evaluation of assertiveness.

Data analysis

For data analysis, the descriptive epidemiology approach was used, suitable for describing and comparing the success rates between different groups (students and chatbots) in different contexts. This approach allows a detailed evaluation of the distribution of correct answers, facilitating the visualization of patterns and trends, such as the differences in performance between students in the twelfth phase of NAPISUL-II and chatbots. The analytical epidemiology approach was not used because the main objective was to describe and compare the rates of correct answers between groups, without investigating causal associations between variables. In addition, it was not the purpose of this research to demonstrate statistically significant differences, as the study focused on the presentation and interpretation of patterns and data distributions.

The correct answers of each chatbot were counted, dividing the number of correct answers by the number of possible correct answers, thus obtaining the success rate of each chatbot. Since we chose to solve the PT completely three times, this count is based on the average obtained from the total number of correct answers in the three attempts, divided by the total number of chances (348). In the case of information about medical students in the NAPISUL-II block, the database was the official result of the PT. Thus, by comparing the average

percentage of correct answers between the medical students of NAPISUL-II in general, the medical students of the NAPISUL-II block who were in the twelfth semester of the course and the average percentage of correct answers of each chatbot, it was possible to have an overview of the performance of these three groups in solving the PT. After that, greater focus was given to the specific comparison between students of the twelfth phase of NAPISUL-II and chatbots, also through the correct averages, but with a distinction between the six blocks defined by ABEM, that is: Basic sciences; Collective health; Gynecology and obstetrics; Internal Medicine; Surgical clinic; Pediatrics. Subsequently, to have a better idea of how the chatbots and medical students of the twelfth semester are distributed, the intervals in which the average of each chatbot was were indicated. In addition, a breakdown of the assertiveness was made through a table, in which the percentages of correct answers of the students and each chatbot in relation to the areas of the PT itself were arranged: Basic sciences; Gynecology and obstetrics; Collective health; Surgical clinic; Internal Medicine; Pediatrics. To test the reliability of the information, a graph was set up that facilitated the visualization of the constancy of the chatbots by elucidating the number of times each chatbot got 0, 1, 2 or 3 times the right answers. Finally, two researchers listed the questions according to specific skills of medical practice. The groups were defined to include more than 10 questions each, ensuring at least 30 answers per chatbot, since all questions were answered 3 times by the same chatbot. Thus, the following categories were defined: Diagnostic management; therapeutic conduct; therapeutic drug conduct; therapeutic conduct of procedure; diagnosis. The objective of each of the categories was, respectively: Diagnostic decision-making, which depends on different areas of knowledge because they relate different aspects of medical knowledge; Complex therapeutic decision-making, which involves the possibility of using drugs, performing maneuvers, procedures or even deciding not to intervene; Selecting drugs appropriate to the situation, knowledge that is also multifactorial, as it depends not only on the patient's conditions, but also on knowledge related to pharmacology; Choosing correctly in situations that require physical management, which includes surgical maneuvers and techniques; Analyze all the signs and symptoms that the patient has, the context in which they are, and all other information presented to relate to a comorbidity. With that, a second table was constructed, so that each type of knowledge had its success rate per chatbot in relation to the students of the twelfth semester of NAPISUL-II demonstrated separately.

Ethical considerations

The research model used in the present study does not require prior evaluation by the Research Ethics Committee (REC), since the data regarding the academic performance of the NAPISUL-II block in the PT are anonymous and accessible to the general public[39].

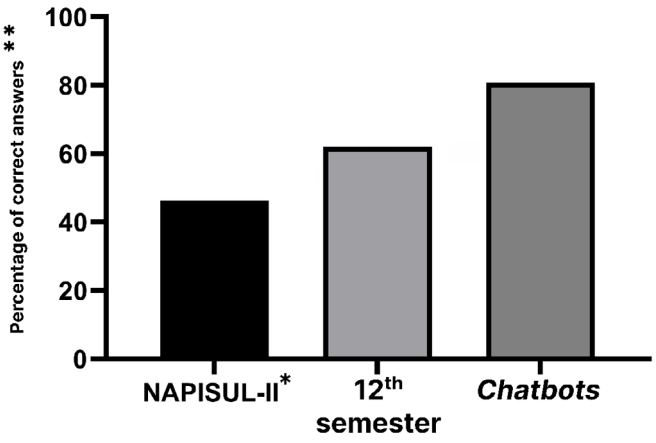
RESULTS

The chatbots had an average score of 93.73 (80.81%) of correct answers in 116 questions, surpassing medical students in the twelfth semester of the NAPISUL-II block, who obtained an average of 74.4 (62%) correct answers in 120 questions, followed by medical students in general in the NAPISUL-II block, with 55.58 (46.31%) correct answers in 120 questions.

In a more detailed analysis, it was observed that Bing outperformed all others, with an average score of 101.33 out of 116 (87.35%), surpassing Chat GPT-4, which achieved 99.6 out of 116 (85.86%). They were followed by, respectively: Bard, with an average of 89 out of 116 (76.72%) and Chat GPT-3.5, with an average of 85 out of 116 (73.27%). This information was compared to the frequency of students in each correct answer range with the use of geometric figures representing the range of the average of correct answers by each chatbot.

In the division by areas, i.e. basic sciences, gynecology and obstetrics, public health, surgical clinic, internal medicine and pediatrics, the students had only one area with a deviation >5% in relation to the general average: Gynecology and obstetrics, with 67.2% of correct answers, being the highest rate of correct answers achieved by them. Chat GPT-3.5, despite being the chatbot with the worst average of correct answers,

Figure 1. Percentage of correct answers in the progress test of all NAPISUL-II students, NAPISUL-II 12th semester students and chatbots.



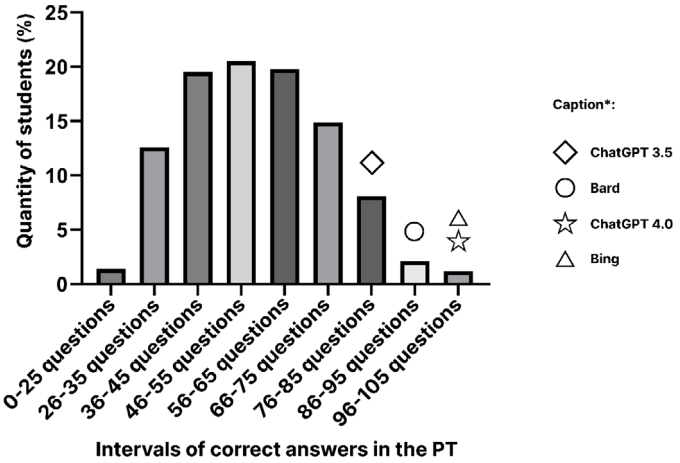
*The average number of correct answers in NAPISUL-II corresponds to all students in the NAPISUL-II block.
**The average of NAPISUL-II and the 12th semester were collected from the official results of the PT, which contained 120 questions, while the average of the chatbots disregarded two canceled questions and two questions with table or image, being analyzed regarding the remaining 116 questions.
Source: Prepared by the authors.

maintained a higher average than students in all areas. In the case of Chat GPT-4, there were highlights in surgical clinic (88.33%) and internal medicine (86.67%), areas in which it scored higher than any other chatbot. Bing, as the chatbot with the highest overall average, achieved the highest average in basic sciences (94.44%), gynecology and obstetrics (94.74%), public health (83.33%) and pediatrics (82.46%). Bard was the chatbot with the highest average inconstancy, reaching 92.59% in basic sciences and 66.67% in pediatrics. This information is depicted on a table.

DISCUSSION

This study advances significantly in relation to the previously published literature on the use of LLMs in the context of the Brazilian Progress Test. While the study by Rodrigues Alessi et al. (2024) exclusively evaluated the performance of ChatGPT 3.5 against PT, the present work expands the scope by including multiple AI models (ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard), introducing a more robust comparative approach. This expansion allows not only to confirm previous findings, but also to identify performance variations between different generative AI tools, better contextualizing their possible pedagogical applications[28].

Figure 2. Distribution of NAPISUL-II Students and Chatbots by Correct Answer Interval in the Progress Test



*Symbols represent what correct answer range each ChatBot's average performance was in
Source: Prepared by the authors

Table 1. Comparison of correct answer percentages between students in the twelfth semester of the undergraduate medical course at the institutions that comprise NAPISUL-II and chatbots in the Progress Test regarding divisions by inherent area

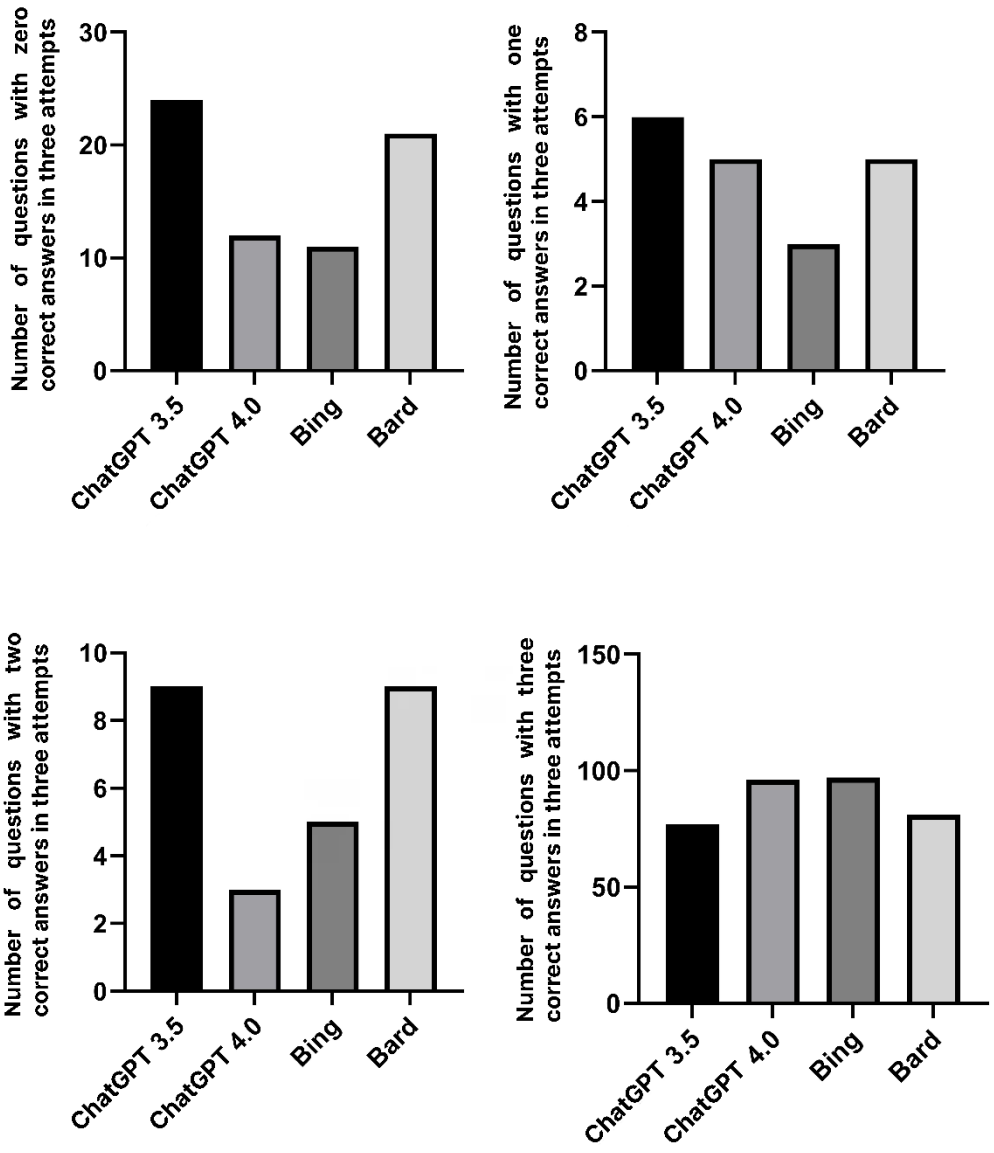
PT Area	12th semester	CHAT GPT 3.5	CHAT GPT 4.0	Bing	Bard
Basic Sciences	60.50%	81.48%	88.89%	94.44%	92.59%
Gynecology and Obstetrics	67.20%	75.44%	92.98%	94.74%	71.93%
Collective Health	59.40%	61.67%	80.00%	83.33%	75.00%
Surgical Clinic	62.10%	75.00%	88.33%	85.00%	80.00%
Internal Medicine	62.50%	76.67%	86.67%	85.00%	73.33%
Pediatrics	60.40%	66.67%	78.95%	82.46%	66.67%
Total	62.00%	73.27%	85.91%	87.36%	76.72%

Table 2. Comparison of correct answer percentages among students of the twelfth semester of the undergraduate medical course of the institution that comprise NAPISUL-II and Chatbots regarding the main types of medical knowledge covered in the Progress Test

Type of knowledge	Number of questions*	12th semester	Chat GPT 3.5	Chat GPT 4.0	Bing	Bard
Diagnostic Management	16	54.53%	72.92%	71.88%	66.67%	60%
Therapeutic Conduct	36	56.81%	70.37%	85.19%	83.33%	67.59%
Therapeutic Conduct (Medication)	14	51.91%	73.81%	92.86%	100%	61.90%
Therapeutic Conduct (Procedure)	11	61.33%	69.70%	84.85%	75.76%	63.64%
Diagnosis	24	60.66%	78.26%	88.41%	91.30%	76.81%

*The number of attempts under review is three times the number of questions, as each chatbot performed the test three times.

Figure 3. Number of Questions with 0, 1, 2 and 3 Correct Answers in 3 Attempts by Chatbot in the NAPISUL-II Progress Test



Source: Prepared by the authors.

The results of this study highlight the remarkable effectiveness of Large Language Models (LLMs) chatbots in resolving medical issues, with Bing AI prevailing over the rest with an accuracy rate of 87.35%, followed by ChatGPT-4 with 85.86%. This performance was superior compared to students in the 12th semester of the South II Interinstitutional Pedagogical Support Center (NAPISUL-II), who had a success rate of 62%. These results were in line with previous studies, which demonstrated superiority of Bing AI and ChatGPT-4 over healthcare students of different levels and other chatbots, even those not explored in this study, such as Claude Instant and Claude. Compared to the results obtained by Rodrigues Alessi et al. (2024), the present study observes a general increase in the accuracy rates of language models. While ChatGPT 3.5 averaged 67.2% to 69.7% of accuracy in the previously analyzed years, in the present study it obtained 73.27% in 2023. ChatGPT 4.0 and Bing AI, not included in the previous analysis, vastly outperformed this performance, with averages of 85.86% and 87.35%, respectively. These data suggest not only technological evolution of the models, but also greater stability of response and capacity for clinical generalization, especially when evaluated in multiple trials. In addition, by adopting a methodology that includes the triple repetition of the questions by each model and the analysis by areas of knowledge and types of clinical reasoning, this study offers a higher analytical granularity than that of the previous study. This approach allows more accurate inferences about the type of medical competence that can be enhanced with the use of chatbots — for example, the superiority of models in matters of drug therapeutic management and clinical diagnosis, key areas for medical practice. Thus, this study contributes with a more solid basis for the development of educational

strategies based on AI[28]. In the study by Morreel, Verhoeven and Mathysen (2024)[40], Bing AI and GPT-4 both scored 76% on the Antwerp University Multiple-Choice Medical Licensing Exam, while the other chatbots evaluated scored 62-67% and students 61%, while Torres-Zegarra et al. found a score of 82.2% on the Peruvian Multiple-Choice Medical Licensing Exam by Bing, 86.7% by ChatGPT-4, with Bard and GPT 3.5 showing lower performance (both 68.8%), but still higher than students (55%)[38].

In addition, the analysis by specific areas revealed that chatbots maintained a higher performance than students in several medical disciplines. For example, ChatGPT-4 obtained in Basic Sciences, Gynecology and Obstetrics, Collective Health, Surgical Clinic, Internal Medicine and Pediatrics the performances of 88.89%, 92.98%, 80%, 88.33%, 86.67% and 78.95%, respectively, while students in the 12th semester obtained 60.50%, 67.20%, 59.40%, 62.10%, 62.50% and 60.40%. These results indicate that chatbots not only outperform students on overall average but also demonstrate superiority in specific areas of medical knowledge. In addition, BING demonstrated superiority over other chatbots in all medical areas evaluated in the progress test, except for Internal Medicine and Surgical Clinic, in which ChatGPT4.0 was superior. These data differed from Torres-Zegarra et al., in which BING was behind ChatGPT4.0 also in the area of Pediatrics and Public Health, but with the same number of correct answers in Internal Medicine[38].

Furthermore, it is worth mentioning that the demonstration of high success rates by chatbots regardless of the medical area or type of knowledge evaluated is present in studies carried out in several languages, countries and types of content assessment, as can be seen in previous studies by Meo et al., Tong et al., Alijindan et al., Ebrahimian et al., Friederichs et al. and Lai et al.[29:41–45].

One of the main limitations of this study lies in the fact that the sample used, consisting of medical students in the 12th semester of a single interinstitutional nucleus (NAPISUL-II), may not be representative of all medical students in Brazil or in other educational contexts. Variability in medical curricula and pedagogical practices can significantly influence student performance on standardized tests, suggesting the need for larger and more diverse samples to validate findings.

One specific issue related to the performance of chatbots is the possible memory bias between different attempts of the same question. Although the questions were presented in groups of 20 and preceded by a standard command to avoid memory bias, the chatbots' ability to store information from previous sessions may have influenced their results. More advanced chatbots, such as ChatGPT-4, may have a more sophisticated memory mechanism that allows them to remember previous responses, potentially influencing the accuracy of responses on subsequent attempts.

In addition, another significant bias in this study is the difference in internet access capacity between the tested chatbots. Some chatbots, such as Bing AI, have the ability to perform real-time internet searches to provide up-to-date answers, while others, such as ChatGPT-3.5 and ChatGPT-4, rely solely on their pre-existing training data without internet access. This discrepancy may have given Bing AI an unfair advantage, allowing it to access newer and potentially more accurate information, while the other models could have been limited by outdated or incomplete data.

Another important limitation to be considered in this study is that the Progress Test, used as an assessment tool, is applied to medical students without offering direct benefits, which can negatively influence student performance. Unlike medical licensing exams, where obtaining a professional license depends on the result and therefore motivates students to push themselves to the maximum, the Progress Test may not engage participants in the same way, resulting in performance that does not reflect their actual capabilities and knowledge. This can lead to an underestimation of students' competence, affecting the comparability of results with studies involving licensing exams, where motivation is significantly higher and performance tends to be more representative of candidates' abilities.

On the other hand, this study shows a solid basis for future investigations and for the development of guidelines for the integration of AI in medical education, since it makes comparative analyses not only with regard to the totality of correct answers, but also the performance related to the different medical areas and also types of reasoning to be developed, which can be fundamental for the development of future tools aimed at specific skills.

Finally, LLM-based chatbots have significant potential to aid in medical education. Kung et al. (2023) demonstrated that ChatGPT provided highly coherent explanations and valuable insights that facilitate the learning process. The answers showed high internal agreement and modeled deductive reasoning, introducing new and non-obvious concepts to the students, enriching understanding and clinical reasoning. Similarly, Cascella et al. (2023) evidenced ChatGPT's ability to assist in the preparation of medical notes and the communication of complex information in an understandable way. The chatbot correctly categorized clinical parameters presented in a disordered or abbreviated way, provided relevant suggestions for further treatments, and adapted the language according to the audience, whether among

health professionals or patients. This evidence, added to that obtained in our study, suggests that the integration of chatbots in medical education can complement traditional methods, offering personalized support, promoting the development of clinical skills, and facilitating the understanding of complex materials, contributing to a more autonomous and effective learning process[20,46].

CONCLUSION

The results of this study show that the chatbots evaluated, especially Bing AI and ChatGPT-4, have a superior ability to solve medical assessment questions compared to students in the last semester of medical school. This finding not only reinforces the potential role of AI tools as valuable complements to traditional teaching, but also opens a promising horizon for the integration of these technologies into educational practices aimed at improving learning and medical assessment.

The segmented analysis of the areas of knowledge highlighted that chatbots, in addition to outperforming students in general terms, demonstrated consistent performance in critical areas such as basic sciences, gynecology and obstetrics, and surgical clinic. These data suggest that the application of chatbots can be particularly useful in reinforcing teaching in disciplines where students face greater challenges, thus offering significant support to the medical training process. In addition, the superiority observed in areas of therapeutic and diagnostic management points to the feasibility of integrating these tools into clinical simulation scenarios, expanding the range of resources available for medical education.

However, it is essential to recognize that limitations permeate this study, such as the memory bias of chatbots and the reduced motivation of students in the Progress Test, which highlights the need for a careful and judicious implementation of these technologies. In order for the identified benefits to be fully employed, future research should focus on strategies to mitigate these biases and on adaptations of chatbots to the curricular and cultural specificities of different educational institutions. Therefore, with a strategic and contextualized approach, chatbots can become potential allies in the training of doctors who are better prepared for the current challenges of clinical practice.

AUTHOR CONTRIBUTIONS

Douglas Augusto Pasqual La Maison contributed to the collaborative creation of the article's topic; Participation in the delimitation of the work methodology; Collaborative participation in sending most of the questions to the chatbots and recording the respective answers; Transcription of the answers into a table; Construction of graphs and tables present in the text; Writing the methodology; Writing the results; Literature review; Article formatting. Amanda Carolina Fonseca da Silva contributed to the joint idealization of the topic of the article; Collaborative participation in the delimitation of the work methodology; Prompt development; Request for the use of Napisul-II data; Collaborative participation in sending questions to chatbots and recording the respective answers; Writing of the introduction; Writing of the discussion; Assistance in the construction of graphs; Literature review; Organization of work references. Ari Ojeda Ocampo Moré contributed by being the creator and advisor of the study, being present in all stages of scientific research and writing. Bruno Ferreira Glir contributed with participation in data collection; literature review regarding the methodology; Review of the article formatting.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

SOURCES OF FUNDING

The authors declare no sources of funding.

DATA AVAILABILITY STATEMENT

Research data are available in the body of the document

REFERENCES

1. Wang X. Experiences, challenges, and prospects of National Medical Licensing Examination in China. BMC Med Educ. 2022;22(1):349.
2. Huber-Lang M, Palmer A, Grab C, Boeckers A, Boeckers TM, Oechsner W. Visions and reality: the idea of competence-oriented assessment for German medical students is not yet realised in licensing examinations. GMS J Med Educ. 2017;34(2):Doc25 [acesso em 13 jan 2024]. Disponível em: http://www.egms.de/en/journals/zma/2017-34/zma001102.shtml.
3. Babla K, Crampton P, Kronfli M. National licensing examinations: what are they good for? Clin Teach. 2020;17(3):323-5.
4. Reberti AG, Monfredini NH, Ferreira Filho OF, Andrade DFD, Pinheiro CEA, Silva JC. Progress Test in medical school: a systematic review of the literature. Rev Bras Educ Med. 2020;44(1):e014.
5. Sakai MH, Ferreira Filho OF, Almeida MJD, Mashima DA, Marchese MDC. Teste de Progresso e avaliação do curso: dez anos de experiência da medicina da Universidade Estadual de Londrina. Rev Bras Educ Med. 2008;32(2):254-63.
6. Bollela VR, Borges MDC, Troncon LEDA. Avaliação somativa de habilidades cognitivas: experiência envolvendo boas práticas para a elaboração de testes de múltipla escolha e a composição de exames. Rev Bras Educ Med. 2018;42(4):74-85.
7. Sakai MH, Ferreira Filho OF, Matsuo T. Avaliação do crescimento cognitivo do estudante de Medicina: aplicação do teste de equalização no Teste de Progresso. Rev Bras Educ Med. 2011;35(4):493-501.
8. Baldim TL, Arcuri MB, Aparecida C. O TESTE DE PROGRESSO SOB A VISÃO DO DISCENTE. Revista da Faculdade de Medicina de Teresópolis [Internet]. 2018 [cited 2024 Aug 12];2(1):41–54. Available from: https://revista.unifeso.edu.br/index.php/faculdadedemedicinadeteresopolis/article/view/608
9. Presta PM, Oliveira A de P, Moreira MLRM, Costa GOF da, Peixoto RAC. Conhecimento em clínica médica: resultados de Teste de Progresso de uma instituição do Nordeste. Revista Interagir. 2024;(126):36-43.

10. Rosa MID, Isoppo CC, Cattaneo HD, Madeira K, Adami F, Ferreira Filho OF. O Teste de Progresso como indicador para melhorias em curso de graduação em Medicina. Rev Bras Educ Med. 2017;41(1):58-68.
11. Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. EdArXiv; 2023 [acesso em 1º fev 2024]. Disponível em: https://osf.io/5er8f.
12. Rasul T, Nair S, Kalendra D, Robin M, Santini FO, Ladeira WJ, et al. The role of ChatGPT in higher education: benefits, challenges, and future research directions. JALT. 2023;6(1):41-56 [acesso em 1º fev 2024]. Disponível em: https://journals.sfu.ca/jalt/index.php/jalt/article/view/787.
13. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare. 2023;11(6):887.
14. Rudolph, J., Tan, S., Tan, S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? JALT. 2023;6(1): 342-362 [acesso em 1º fev 2024]. Disponível em: https://journals.sfu.ca/jalt/index.php/jalt/article/view/689.
15. Costa MJM, Santos DWD, Bottentuit Junior JB. Inteligência artificial e metodologias ativas no ensino de medicina: percepções dos discentes de habilidades médicas de um centro universitário. Revista Intersaberes. 2024;19;e24do3003.
16. 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.
17. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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. 2023;9:e45312.
18. Shoja M M, Van de Ridder J, Rajput V (June 24, 2023) The Emerging Role of Generative Artificial Intelligence in Medical Education, Research, and Practice. Cureus 15(6): e40883. doi:10.7759/cureus.40883.
19. Ramos ASM. Inteligência Artificial Generativa baseada em grandes modelos de linguagem - ferramentas de uso na pesquisa acadêmica [Internet]. SciELO Preprints. 2023 [citado 1º de fevereiro de 2024]. Disponível em: https://preprints.scielo.org/index.php/scielo/preprint/view/6105
20. Kung TH, Cheatham M, Medenilla A, Sillos C, De 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.
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
22. Lee H. The rise of ChatGPT : exploring its potential in medical education. Anat Sci Educ. 2024;17:926–931. doi: 10.1002/ase.2270.
23. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9.
24. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc. 2023;86(7):653-8.
25. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first-year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service examination. Aesthet Surg J. 2023;43(12):NP1085-9.
26. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European exam in core cardiology: an artificial intelligence success story? Eur Heart J Digit Health. 2023;4(3):279-81.
27. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA Intern Med. 2023;183(9):1028-1030.
28. Alessi MR, Gomes HA, Castro ML de, Okamoto CT. Performance of ChatGPT in solving questions from the Progress Test (Brazilian National Medical Exam): a potential artificial intelligence tool in medical practice. 2024 Jul 19;16(7):e64924. [acesso em 25 maio 2025]. Disponível em: https://www.cureus.com/articles/272161-performance-of-chatgpt-in-solving-questions-from-the-progress-test-brazilian-national-medical-exam-a-potential-artificial-intelligence-tool-in-medical-practice.
29. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online. 2023;28(1):2220920.
30. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the possible use of AI chatbots in public health education: feasibility study. JMIR Med Educ. 2023;9:e51421.
31. Chat GPT 4.0 [acesso em Novembro de 2023]. Disponível em: https://chatgpt.com.
32. Copilot AI [acesso em Novembro de 2023]. Disponível em: https://copilot.microsoft.com.
33. Gemini AI [acesso em Novembro de 2023]. Disponível em: https://gemini.google.com.
34. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. Eur J Orthod 2024;46: cjae017.
35. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? arXiv; 2023 [acesso em 3 jul 2024]. Disponível em: http://arxiv.org/abs/2207.08143.
36. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg. 2023;31(23):1173-9.
37. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589-597.
38. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof. 2023;20:30.
39. Resultados Teste de Progresso [acesso em Julho de 2024]. Disponível em: https://medicina.ufsc.br/?page_id=2462.
40. Morreel S, Verhoeven V, Mathysen D. Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. PLOS Digit Health. 2024;3(2):e0000349.
41. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare. 2023;11(14):2046.
42. Tong W, Guan Y, Chen J, Huang X, Zhong Y, Zhang C, et al. Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT, in the Chinese National Medical Licensing Examination. Front Med. 2023;10:1237432.
43. Aljindan FK, Al Qurashi AA, Albalawi IAS, et al. ChatGPT conquers the Saudi medical licensing exam: exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus. 2023 Sep;15(9):e45043. doi: 10.7759/cureus.45043. doi. Medline
44. 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. 2023;30(1):e100815.
45. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med. 2023;10:1240915.
46. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33.