

Comparative Performance of Medical Students, ChatGPT-3.5 and ChatGPT-4.0 in Answering Questions from a Brazilian National Medical Exam: An Observational Study

Mateus Rodrigues Alessi, Heitor Augusto Gomes, Gabriel Felipe Contin de Oliveira, Matheus Lopes de Castro, Fabiano Grenteski, Leticia Miyashiro, Camila do Valle Pavanelo, Leticia Tozzini Tavres da Silva, Cristina Terumy Okamoto

Submitted to: JMIR AI
on: September 19, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 21

..... 21

Figures 22

Figure 1..... 23

Figure 2..... 24

Figure 3..... 25

Figure 4..... 26

Figure 5..... 27

Figure 6..... 28

Figure 7..... 29

Figure 8..... 30

Comparative Performance of Medical Students, ChatGPT-3.5 and ChatGPT-4.0 in Answering Questions from a Brazilian National Medical Exam: An Observational Study

Mateus Rodrigues Alessi^{1*}; Heitor Augusto Gomes^{1*}; Gabriel Felipe Contin de Oliveira^{1*}; Matheus Lopes de Castro^{1*}; Fabiano Grenteski¹; Leticia Miyashiro^{1*}; Camila do Valle Pavanelo^{1*}; Leticia Tozzini Tavres da Silva^{1*}; Cristina Terumy Okamoto^{1*}

¹ Universidade Positivo Curitiba BR

*these authors contributed equally

Corresponding Author:

Mateus Rodrigues Alessi

Universidade Positivo
Rua Carlos Gelenski 71
House 31
Curitiba
BR

Abstract

Background: Artificial intelligence (AI) has advanced significantly in various fields, including medicine, where tools like ChatGPT (GPT) have demonstrated remarkable capabilities in interpreting and synthesizing complex medical data. Since its launch in 2019, GPT has evolved, with version 4.0 offering enhanced processing power, image interpretation, and more accurate responses. In medicine, GPT has been used for diagnosis, research, and education, achieving significant milestones like passing the USMLE board exam. Recent studies show that GPT 4.0 outperforms its earlier versions and medical students on medical exams.

Objective: This study aimed to evaluate and compare the performance of GPT versions 3.5 and 4.0 on Brazilian Progress Tests (PT) from 2021 to 2023, analyzing their accuracy compared to medical students.

Methods: A cross-sectional observational study was conducted on 333 multiple-choice questions from the PT, excluding questions with images, nullified, or repeated items. All questions were presented sequentially, without any modification to their structure. The performance of GPT versions was compared using statistical methods, and medical students' scores were included for context.

Results: GPT 4.0 showed a significant improvement in accuracy, achieving an 87.2% accuracy rate compared to 68.4% for GPT 3.5 (p-value = 0.028), representing a relative improvement of 27.4%. Despite these differences, both GPT 3.5 and 4.0 achieved a higher score than students from all years of medical school.

Conclusions: GPT 4.0 demonstrates superior accuracy compared to its predecessor in answering medical questions on the PT. The results of this article are similar to other researches, indicating that we are getting closer to new revolution in Medicine.

(JMIR Preprints 19/09/2024:66552)

DOI: <https://doi.org/10.2196/preprints.66552>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to the public.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>, I will be able to make my accepted manuscript PDF available to anyone.

No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

Comparative Performance of Medical Students, ChatGPT-3.5 and ChatGPT-4.0 in Answering Questions from a Brazilian National Medical Exam: An Observational Study

Abstract

Introduction: Artificial intelligence (AI) has advanced significantly in various fields, including medicine, where tools like ChatGPT (GPT) have demonstrated remarkable capabilities in interpreting and synthesizing complex medical data. Since its launch in 2019, GPT has evolved, with version 4.0 offering enhanced processing power, image interpretation, and more accurate responses. In medicine, GPT has been used for diagnosis, research, and education, achieving significant milestones like passing the USMLE board exam. Recent studies show that GPT 4.0 outperforms its earlier versions and medical students on medical exams.

Objective: This study aimed to evaluate and compare the performance of GPT versions 3.5 and 4.0 on Brazilian Progress Tests (PT) from 2021 to 2023, analyzing their accuracy compared to medical students.

Methodology: A cross-sectional observational study was conducted on 333 multiple-choice questions from the PT, excluding questions with images, nullified, or repeated items. All questions were presented sequentially, without any modification to their structure. The performance of GPT versions was compared using statistical methods, and medical students' scores were included for context.

Results: There was a statistically significant difference in total performance scores across the 2021, 2022, and 2023 exams between AI versions 3.5 and 4.0 ($p = 0.028$). However, this significance did not remain after Bonferroni correction. On average, version 3.5 scored 68.4%, whereas version 4.0 achieved an impressive 87.2%, reflecting an absolute improvement of 18.8% and a relative increase of 27.4% in accuracy. When broken down by subject, the average scores for versions 3.5 and 4.0, respectively, were: Surgery (73.5% vs. 88.0%, $p = 0.032$), Basic Sciences (77.5% vs. 96.2%, $p = 0.004$), Internal Medicine (61.5% vs. 75.1%, $p = 0.137$), Gynecology and Obstetrics (64.5% vs. 94.8%, $p = 0.002$), Pediatrics (58.5% vs. 80.0%, $p = 0.022$), and Public Health (77.8% vs. 89.6%, $p = 0.021$). After Bonferroni correction, only Basic Sciences and Gynecology and Obstetrics retained statistically significant differences.

Conclusion: GPT 4.0 demonstrates superior accuracy compared to its predecessor in answering medical questions on the PT. The results of this article are similar to other researches, indicating that we are getting closer to new revolution in Medicine.

IRB approval: The research did not require an approval.

Keywords: Artificial Intelligence, Intelligent Systems, Biomedical Technology, Medical Ethics, Exam Questions, Academic Performance

Introduction:

Artificial intelligence (AI) is a branch of science centered on the development of systems and algorithms capable of performing complex tasks that typically require human cognition. One famous example is ChatGPT (GPT), introduced by OpenAI in 2019 [1]. Unlike other AIs, it relies on large language models and deep learning, which, in general terms, means that the tool uses vast amounts of text through deep neural networks to analyze and generate natural language with a high degree of complexity and precision, learning and evolving from its own mistakes. Its popularity stems from its ability to synthesize and interpret complex texts, responding to users in seconds. Instead of retrieving information directly from internet data source, its responses are generated based on the probabilistic pattern of two or more words appearing together in a sentence. Since its launch, OpenAI has been making improvements and updating the tool, releasing version 4.0 in March 2023, surpassing its predecessor, 3.5. The new improvements include: image interpretation, greater processing power and the ability to solve complex problems, interpretation of more words in a single query, and more accurate responses, understanding the subtleties of human language.

Since AI development, it has been used in various fields of medicine, with new studies focusing in medical education [2–5]. When GPT 3.0 was challenged with the United States Medical Licensing Examination (USMLE), it scored around 60%, enough to pass the exam, a milestone achievement in the academic field, as the tool passed all three exams, which together contain more than 1,000 questions and are recognized worldwide for their difficulty. This article also suggested that GPT 3.0 has a similar intelligence of a third-year medical student [6]. In addition, Liu et al. in 2024, performed a Meta-analysis evaluating the performance of GPT in 45 medical examinations worldwide [7]. The study found that version 4.0 achieved an overall accuracy rate of 81%, outperforming GPT 3.5, and, in most cases, passing the medical examinations, surpassing the average scores of medical students. In this research, 29 of the articles used version 4.0, 26 tested the performance of version 3.5 and 14 studies tested both GPT 4.0 and GPT 3.5. This study also found that translating the test to English significantly improves the accuracy of GPT 3.5 but not 4.0. Only one Brazilian study was included in the meta-analysis, finding 87.7% of correct answers for GPT 4.0 when exposed to the Brazilian National Examination for Medical Degree Revalidation (Revalida), but only version 4.0 of the AI was evaluated [8].

The present article builds upon the work of Rodrigues Alessi et al. in 2024, who applied GPT 3.5 to the Progress Tests (PT) of 2021, 2022, and 2023, finding an average accuracy of 68.4%, surpassing medical students of all years [9]. Even though Brazil does not have a national medical exam for resident selection or for newly graduated doctors to practice the profession, PT is a national exam with over 50,000 medical students participating in recent editions. It consists of 120 multiple-choice questions with five alternatives and only a single correct answer, equally distributed across the areas of, Surgery, Pediatrics, Social Sciences, Internal Medicine, Basic Sciences, and Gynecology and Obstetrics. In 2021, the Brazilian Association of Medical Education (ABEM) administered a single national exam, while in 2022 and 2023, the exams were divided by regions of Brazil. These last two evaluations have a regional scope, encompassing certain conglomerates of states in the country. One of these nuclei is the South II Institutional Pedagogical Support Center (Napisul II), which includes a total of 13 universities from Paraná and Santa Catarina. It is worth noting that this exam is administered by the same group, with the same number of questions and covering the same six topics. The difficulty and the way the questions are evaluated are similar.

The exact evolution between GPT 3.5 and 4.0 in answering Brazilian medical questions remains uncertain. Therefore, the study aimed to assess and compare the performance of medical students and the two AI versions in answering questions from a Brazilian national medical exam in

its native language.

Methodology

Overview

The study utilized an observational, cross-sectional design to evaluate the performance of GPT 3.5 and 4.0 on 333 questions from the PT of 2021, 2022 and 2023 on its original language, excluding questions with image, nullified questions, and repeated questions. Each question was manually inputted into the AI platform, and the tool's history was cleared and restarted after each question to avoid memory bias. Responses were categorized as correct and incorrect. In addition, first to sixth-year medical students' mean score was also compared, using appropriate statistical methods to assess accuracy and effectiveness. Unfortunately, ABEM didn't release the number of students by each year of medicine school who participated in these tests, but the approximate number of all tests together surpass 50,000 medical students. All questions and the answer keys used in this article are public available.

Inclusion and exclusion criteria:

Questions from the 2021 National PT and the 2022 and 2023 Regional Tests (NAPISUL II), totaling 360 questions (120 from each test) were included. We excluded questions that included images or figures containing graphs; questions that were repeated across the three tests evaluated and questions that were invalidated by the test organizers (those that, after the exam was administered, were found to contain errors or issues, such as ambiguous wording, incorrect answers, or flawed reasoning).

As a result, 333 multiple-choice questions were used: 109 from the 2021 PT, 117 from the 2022 NAPISUL II Test, and 107 from the 2023 NAPISUL II Test, each containing only one correct answer out of five possible alternatives (A, B, C, D, or E).

Procedures:

First, each question included in the study criteria was asked to the virtual platform GPT 3.5 and 4.0 (separately), and the responses were categorized into two possible outcomes based on the official answer key for each test: correct answer or incorrect answer. To avoid memory bias on the part of the AI, the platform's history was deleted and the site was restarted after each question was presented.

In cases where the platform considered more than one answer to be correct, the question "Which is the most correct alternative?" was asked to obtain a single answer and improve statistical interpretation. If the new AI response matched the official answer key, it was considered correct; if it differed, it was considered incorrect.

Thus, three possible outcomes were obtained for each question asked of GPT: (A) Correct; (B) Incorrect but agreed with answer key after it was presented; (C) Incorrect and did not agree with answer key after it was presented.

Results were compared between versions 3.5, 4.0, and the students, divided into overall average scores (average scores of students from the first to sixth year) and only sixth-year students. The results were also divided into the six main subject areas of the test. Unfortunately, for the 2021

test, the average percentage of correct answers per subject for students from the first to sixth year was not provided by ABEM, only the percentage per subject for sixth-year students was made public. Since the data were only available in averages, a statistical significance test could not be conducted.

Data Evaluation

The data were organized in an Excel® spreadsheet and analyzed using IBM SPSS Statistics v.29.0.0 software. For the descriptive analysis of quantitative variables, the mean, standard deviation, median, minimum, and maximum were presented. The Wilcoxon non-parametric test was used to compare the accuracy rates between versions 3.5 and 4.0, and Bonferroni corrections was performed for the results. P-values <0.05 were considered statistically significant.

Ethical Considerations

This study only used information that was already published on the internet and did not involve human subjects; rather, an analysis of the PT was performed. Therefore, approval by the Institutional Review Board of Shimane University was not required.

Results:

Based on the questions that met the inclusion criteria, the results showed that there is a statistically significant difference between the total performance scores of the three tests (2021, 2022, and 2023) between the AI versions 3.5 and 4.0 (p-value = 0.028). However, after Bonferroni corrections, we found a non-significant value. The average score of versions 3.5 was 68.4%, while version 4.0 achieved an outstanding 87.2% accuracy rate. This indicates an absolute difference of 18.8% between the two versions and a relative improvement of 27.4% in accuracy for the latest platform version.

The mean score of the three years tests by subject for version 3.5 and 4.0 were, respectively: 73.5% vs. 88% for Surgery (p-value = 0.032), 77.5% vs. 96.2% for Basic Sciences (p-value = 0.004), 61.5% vs. 75.1% for Internal Medicine (p-value = 0.137), 64.5% vs. 94.8% for Gynecology and Obstetrics (p-value = 0.002), 58.5% vs. 80% for Pediatrics (p-value = 0.022), and finally, 77.8% vs. 89.6% for Public Health (p-value = 0.021) as shown in (Table 1). After Bonferroni corrections, only Basic Sciences and Gynecology and Obstetrics continued with statistically significant results.

Table 1 Title: GPT 3.5 Vs GPT 4.0 Statistical Significance Evaluation

Subject	Version	Mean	p*	p**	Number of
---------	---------	------	----	-----	-----------

					Questions
Surgery	v3.5	73.5	0.032	0.224	53
	v4.0	88.0			
Basic Sciences	v3.5	77.5	0.004	0.028	53
	v4.0	96.2			
Internal Medicine	v3.5	61.5	0.137	0.959	55
	v4.0	75.1			
Gynecology and Obstetrics	v3.5	64.5	0.002	0.014	57
	v4.0	94.8			
Pediatrics	v3.5	58.5	0.022	0.154	56
	v4.0	80.0			
Public Health	v3.5	77.8	0.021	0.147	59
	v4.0	89.6			
Overall	v3.5	68.4	0.028	0.196	333
	v4.0	87.2			

TABLE 1 TEXT: The statistical analysis of the 6 subjects of the tests plus the overall score in percentage of GPT 3.5 and 4.0. *: The test used was the Non parametric Wilcoxon test, and the statistical significance value was defined as a $p < 0.05$. **: The test used was the Non parametric Wilcoxon test with Bonferroni correctios, and the statistical significance value was defined as a $p < 0.05$. The number of questions from all three tests together is also present in the right column.

Gynecology and Obstetrics showed the biggest absolute difference between the two versions of GPT, with a 30.3% absolute difference score and a relative difference of 46.9%. meanwhile, Public Health showed the lowest difference, with an absolute difference of 11.8%, and a relative difference of 15.1%.

Among the 333 questions selected for the study, the overall average accuracy of GPT 3.5 for the tests of 2021, 2022, and 2023 was 69.7%, 68.3%, and 67.2%, respectively. On the other hand, GPT 4.0, scored 87.8%, 86.4%, and 87.7% for the same tests during that period. Conversely, medical students scored 49.7%, 45%, and 57.4%. Regarding the performance of 6th-year students alone, the average accuracy was 66.3%, 56.5%, and 60% on the respective tests. These results are illustrated in (Figure 1-2).

In the 2021 test, due to the unavailability of students' subject-specific accuracy data, it was not possible to compare the students' accuracy by subject against GPT. Nevertheless, when comparing the overall score, the results were 49.7% for the students, 69.7% for GPT 3.5, and 87.8% for GPT 4.0. The values found for versions 3.5 and 4.0, respectively, were: 94.1% vs. 94.1% (Basic Sciences), 68.7% vs. 81.2% (Surgery), 66.6% vs. 66.6% (Internal Medicine), 50% vs. 100% (Gynecology and Obstetrics), 59% vs. 85% (Pediatrics), and 90% vs. 100% (Public Health), as presented in (Figure 3).

In the 2022 test, the overall score was 45.2% for the students, 68.3% for GPT 3.5, and 86.4% for GPT 4.0. When evaluating the average accuracy per subject for students and AI versions 3.5 and 4.0, the results were, respectively: 46.1% vs. 73.6% vs. 94.7% (Basic Sciences); 44% vs. 68.4% vs. 94.7% (Surgery); 43.3% vs. 65% vs. 70% (Internal Medicine); 50.2% vs. 68.4% vs. 89.4% (Gynecology and Obstetrics); 42.3% vs. 60% vs. 80% (Pediatrics); and 45.1% vs. 75% vs. 90% (Public Health), summarized graphically in (Figure 4).

In 2023, the overall student score was 57.4%, compared to 67.2% for GPT 3.5 and 87.7% for GPT 4.0. The scores by subject for medical students, GPT 3.5 and 4.0 of GPT were, respectively: 56.8% vs. 64.7% vs. 100% (Basic Sciences); 57.9% vs. 83.3% vs. 88.8% (Surgery); 56% vs. 52.9% vs. 88.8% (Internal Medicine); 62.3% vs. 75% vs. 95% (Gynecology and Obstetrics); 56.9% vs. 56.2% vs. 75% (Pediatrics); and 54.7% vs. 68.4% vs. 78.9% (Public Health), graphically illustrated in (Figure 5).

The score analysis of the 2022 and 2023 tests together, separated by medical students vs. GPT 3.5 vs. GPT 4.0 was, respectively: 51.3% vs. 67.8% vs. 87% (Overall); 51.5% vs. 69.2% vs. 97.3% (Basic Sciences); 51% vs. 75.9% vs. 91.7% (Surgery); 49.7% vs. 59% vs. 79.4% (Internal Medicine); 56.3% vs. 71.7% vs. 92.2% (Gynecology and Obstetrics); 49.6% vs. 58.1% vs. 77.5% (Pediatrics); and 49.9% vs. 71.7% vs. 84.4% (Public Health), as presented in (Figure 6).

Data on average subject accuracy for 6th-year students in the three tests were made available by ABEM, which allowed to compare the accuracy of 6th-year students with GPT versions 3.5 and 4.0, yielding the following results, respectively: 60.9% vs. 68.4% vs. 87.2% (Overall); 54.8% vs. 77.5% vs. 96.2% (Basic Sciences); 51.2% vs. 73.5% vs. 88% (Surgery); 53.8% vs. 61.5% vs. 75.1% (Internal Medicine); 56.7% vs. 64.5% vs. 94.6% (Gynecology and Obstetrics); 55.1% vs. 58.5% vs. 80% (Pediatrics); and 55% vs. 77.8% vs. 89.3% (Public Health), as shown in (Figure 7-8).

The latest AI version answered two new incorrect questions and 12 new correct ones for Basic Sciences; four new incorrect questions and 11 new correct ones for Surgery; four new incorrect questions and 11 new correct ones for Internal Medicine; two new incorrect questions and 18 new correct ones for Gynecology and Obstetrics; five new incorrect questions and 19 new correct ones for Pediatrics; and two new incorrect questions and 10 new correct ones for Public Health.

Example

For demonstrative purpose, an example of a real question that was present to GPT 3.5 and GPT 4.0 will be presented below. All the text was inserted in Portuguese, and latter translated to English for better worldwide comprehension, but it is important to highlight that in this particular research everything was done in the original language of the text.

Questions example: News reported that ovarian implantation technique can be successful in the field of fertility. A woman who suffered from leukemia had her ovaries removed and preserved in the laboratory and, after treatment for the disease, she underwent reimplantation of her ovaries, successfully producing oocytes. After the medical-surgical procedure, it was possible to verify the growth and maturation of the oocytes. The hormones secreted by the pituitary gland and ovary, respectively, that induce ovulation are:

Response:

1. Follicle stimulating and estrogen; luteotrophic hormone.
2. B Luteinizing and follicle stimulating; progesterone.
3. C Stimulating and luteotrophic follicle; estrogens.
4. D Stimulating and luteotrophic follicle; progesterone.

GPT 3.5 answer: "The correct answer is option A: Follicle stimulating and estrogens; luteotrophic hormone. Here is the explanation:

1. **Follicle Stimulating Hormone (FSH):** Secreted by the pituitary gland, FSH stimulates the growth and maturation of ovarian follicles, which contain oocytes. During this process, developing follicles secrete estrogens.

2. **Estrogens:** These are hormones produced by the ovaries (specifically by maturing follicles) and play a crucial role in regulating the menstrual cycle, promoting the growth of the endometrium and preparing the body for a possible pregnancy.

3. **Luteinizing Hormone (LH) or Luteotrophic Hormone:** Also secreted by the pituitary gland, LH is responsible for triggering ovulation (the release of the mature oocyte from the follicle). After ovulation, LH promotes the formation of the corpus luteum, which secretes progesterone.

Therefore, ovulation is induced by pituitary hormones, such as FSH (which stimulates the maturation of follicles and the production of estrogen) and LH (luteotrophic hormone, which triggers ovulation)."

GPT 4.0 answer: "The correct answer is option C: Stimulating and luteotrophic follicle; estrogens. Explanation:

- **Follicle stimulating (FSH):** Hormone secreted by the pituitary gland, responsible for stimulating the growth and maturation of ovarian follicles, which contain oocytes.
- **Luteotrophic (LH):** Also secreted by the pituitary gland, it is responsible for the ovulation process, where the mature follicle releases the oocyte.
- **Estrogens:** Hormones produced by the ovary, important in the development and maintenance of the female reproductive system and in regulating the menstrual cycle. They also play a crucial role in inducing the LH surge, which results in ovulation.

Thus, the correct sequence is FSH (follicle stimulating) and LH (luteotrophic) by the pituitary, and estrogens by the ovary."

The answer key indicated that letter C was the correct one, meaning GPT 4.0 choose the correct answer with a meaningful explanation, when GPT 3.5 made a mistake, but still tried to explain why it choose that alternative.

Discussion:

Principal Findings

GPT 4.0 displayed a high degree of superiority over its predecessor, improving its score by at least 10% in all subjects, with results similar to those found in other studies. The average score of the three tests together was 87.2%, with a relative increase of 27.4% from the 3.5 version. Comparing the results obtained with those in the literature, which assessed the percentage of correct answers between the two versions in various medical knowledge exams from different countries, it is noted that all found a score close to 70% accuracy for version 4.0, with a better performance for the new version, with accuracy differences ranging from 5.34% to 32.1% [10–17]. A new Meta-analysis reviewing 45 studies that compared these two platforms in licensing examination tests across the globe found an accuracy rate above 80% for the 4.0 version, similar to this study [7]. Like our results, Liu et al in 2024 found many results surpassing human performance on the exams [7].

Reflections

However, the question remains: Is this improvement aligned with expectations? Analysts, including Bill Gates, have raised concerns that AI platforms like GPT may be nearing a plateau, with each successive update yielding diminishing returns [18]. While there's no direct comparison between GPT 2.5 or 3.0 on PT questions, the leap from version 3.5 to 4.0 was notably substantial, as evidenced by a 27.6% relative score increase in our study. This evolution doesn't necessarily mean that future updates won't encounter stagnation, but we may not have reached that point yet. As new updates and studies emerge, further comparison across applications will be crucial.

While the results were impressive, the most important takeaway from this study is that GPT can be a reliable source for learning, if used in conjunction with other tools. Its applications in medical education include communication, knowledge retention, writing and interpreting texts and images, individualized and personalized study, creation of clinical cases, and hypothesis generation for diseases, especially the rarer and less remembered ones [19–23].

Despite all of these, some schools are working to restrict the use of IA in classrooms, fearing that students will misuse it to complete assignments faster without learning. This view is flawed, since the focus should be on educating professors and institutions on how to integrate AI into curricula effectively. If a task given by a professor can be easily carried out by GPT, the error is not in the student using it, but in the institution that assigned it. What is the point of expending time and resources on a task that is already being done better in AI? This new type of technology is too new, and professors are still learning how to use them, therefore, more time is necessary to train teachers on balancing the use of AI in class, as Machine learning might prove as important as the internet revolution in academic settings.

In Medicine, AI is already being used in the creation of medical forms, assistance in diagnoses and decision-making, interpretation and reporting of exams, patient monitoring, cost reduction, education, probabilistic prognostic elaboration, identification of treatment responses, medical record creation, and much more [2–5]. However, we cannot overlook the potential concerns involving its use. The main barriers are validation, usability, utility, and ethics [24]. More research around the world is still needed, generating more data on the use of AIs in Medicine. Only after extensive proof of superiority will its use be legalized [25]. Utility, in turn, refers to the functionality of a tool, being studied and improved through more research and the use in various functions, while usability refers to the ability of healthcare professionals to use the tool to achieve satisfactory results [25]. Finally, ethical considerations are a cornerstone in integrating AI in medicine [26]. AI systems can pose risks to privacy and confidentiality, as the handling of sensitive patient data may be vulnerable to misuse or unauthorized access. Additionally, biases inherent in the algorithms, such as those resulting from biased training data or biased model development, can perpetuate disparities in

healthcare, potentially leading to unfair outcomes for certain groups. Furthermore, deploying AI in areas without prior validation can raise concerns about its reliability and safety in clinical decision-making, as well as its impact on patient care. These challenges underscore the importance of addressing ethical principles such as transparency, liability, and fairness in the development and application of AI in healthcare, ensuring it serves all patients equitably and responsibly.

Another key issue with GPT is the phenomenon by which the tool creates fictitious and incorrect information, called "hallucinations" [27]. This action is not fully understood, but provides a plausible-sounding, yet wrong, answer for the user, who is not informed of such an act [25]. The biggest problem is that GPT can create detailed and local explanations that are completely wrong, but may look very convenient for the reader. This phenomenon may explain why the AI answered incorrectly some questions that it previously scored correctly.

There are still many unsolved questions. How will AI affect the doctor-patient relationship? Will the new generation of doctors be dependent on technology or will it enhance their abilities? What will happen to medical intuition? All these questions will only be answered with time, but we cannot forget the impact of a good patient-medical relationship in treating diseases or improving medication adherence [28]. And what about the technology dependence we are already living in? A physician must be able to diagnose appendicitis with physical examination, without always requiring a computer tomography. The future medical generations must understand how to use the technology in their favor, without forgetting the basics of Medicine. Finally, medical intuition has been used for years, assisting health care professionals when machines could not detect abnormalities [29]. The instinct, or sixth sense one gets with years of medical practice cannot be taught or left aside at the expense of algorithms and probability formulas. The challenge will be integrating AI into Medicine while preserving the essential human touch in patient care.

Though there is still a period of study and adaptation for the implementation of these new technologies in Medicine, it is evident that AI will replace humans in certain areas. Daily, new research presents the use of machine learning in image interpretation, clinical reasoning, laboratory test analysis, drug development, and much more [2–5,23,24]. Unfortunately, those who fail to stay updated in the job market will be at great risk of being left behind. Most likely, individual qualities will change drastically, with more human characteristics such as empathy, creativity, abstraction, leadership, communication, and flexibility being more sought after. Professionals who learn to work alongside new technologies, forming a human-machine symbiosis to take advantage of their functions while remaining empathetic and human, will likely be the most valued in the future.

Limitation

Several limitations need consideration for this article, like the use of average student scores from ABEM, which limits the ability to perform detailed statistical significance analyses. The AI's last data update was in 2021, which could also affect its performance on more recent topics. Additionally, the optional nature of the PT introduces potential selection bias, as not all students or universities participate. Finally, LLMs can generate different responses each time due to inherent stochasticity, it is important to note that this variability was not systematically assessed in the present study and should be addressed in future research. Addressing these limitations in new studies, could provide even more insightful results.

Conclusion

GPT 4.0 demonstrates superior accuracy compared to its predecessor in answering medical questions on the PT. Its percentage of correct answers was higher in all subjects of the exam, even surpassing the scores of students from the first to sixth years of medical school. Even though the results of this article are similar to other researches, more scientific work is required for AI full implementation in Medicine, but one thing is certain, this technology is here to stay.

Acknowledgements

This article did not receive any funding, and all financial aspects were covered by the authors. No sponsors were present. All data used is publicly available.

Author's contribution

Conceptualization or Study Design: Mateus Rodrigues Alessi, Matheus Lopes de Castro, Heitor Augusto Gomes, Gabriel Felipe Contin de Oliveira, Fabiano Grenteski, Leticia Miyashiro, Leticia Tozzini Tavares da Silva, Camila do Valle Pavanelo

Data Collection: Mateus Rodrigues Alessi, Matheus Lopes de Castro, Heitor Augusto Gomes, Gabriel Felipe Contin de Oliveira, Fabiano Grenteski, Leticia Tozzini Tavares da Silva, Camila do Valle Pavanelo

Data Analysis and Interpretation: Mateus Rodrigues Alessi, Matheus Lopes de Castro, Heitor Augusto Gomes, Gabriel Felipe Contin de Oliveira, , Leticia Tozzini Tavares da Silva, Camila do Valle Pavanelo

Methodology Development: Mateus Rodrigues Alessi, Matheus Lopes de Castro, Heitor Augusto Gomes, Fabiano Grenteski, Leticia Miyashiro, Camila do Valle Pavanelo

Writing – Original Draft: Mateus Rodrigues Alessi, Matheus Lopes de Castro, Heitor Augusto Gomes, Gabriel Felipe Contin de Oliveira, Fabiano Grenteski, Leticia Miyashiro

Writing – Review & Editing: Mateus Rodrigues Alessi, Heitor Augusto Gomes, Leticia Miyashiro, , Leticia Tozzini Tavares da Silva, Camila do Valle Pavanelo

Project Administration: Mateus Rodrigues Alessi, Cristina Terumy Okamoto

Conflicts of Interest

None declared.

Abbreviations

AI: Artificial Intelligence

GPT: ChatGPT

PT: Progress Test



References

1. ChatGPT. Accessed: July 28, 2024. <https://chatgpt.com/>.
2. Au K, Yang W: Auxiliary use of ChatGPT in surgical diagnosis and treatment. *Int J Surg*. 2023, 109:.. 10.1097/JS9.0000000000000686
3. Hassan AM, Rajesh A, Asaad M, Nelson JA, Coert JH, Mehrara BJ, Butler CE: Artificial Intelligence and Machine Learning in Prediction of Surgical Complications: Current State, Applications, and Implications. *American Surgeon*. 2023, 89:.. 10.1177/00031348221101488
4. Koohi-Moghadam M, Bae KT: Generative AI in Medical Imaging: Applications, Challenges, and Ethics. *J Med Syst*. 2023, 47:.. 10.1007/s10916-023-01987-4
5. Chenais G, Lagarde E, Gil-Jardiné C: Artificial Intelligence in Emergency Medicine: Viewpoint of Current Applications and Foreseeable Opportunities and Challenges. *J Med Internet Res*. 2023, 25:.. 10.2196/40031
6. Kung TH, Cheatham M, Medenilla A, et al.: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*. 2023, 2:.. 10.1371/journal.pdig.0000198
7. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, Kiuchi T: Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res*. 2024, 26:e60807. 10.2196/60807
8. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, Belfort R: Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. *Rev Assoc Med Bras*. 2023, 69:.. 10.1590/1806-9282.20230848
9. Rodrigues Alessi M, Gomes HA, Lopes de Castro M, Terumy Okamoto C: Performance of ChatGPT in Solving Questions From the Progress Test (Brazilian National Medical Exam): A Potential Artificial Intelligence Tool in Medical Practice. *Cureus*. Published Online First: 19 July 2024. 10.7759/cureus.64924
10. 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:.. 10.2196/48002
11. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S: Exploring the Performance of ChatGPT Versions 3.5, 4, and 4 With Vision in the Chilean Medical Licensing Examination: Observational Study. *JMIR Med Educ*. 2024, 10:e55048–e55048. 10.2196/55048
12. Lim ZW, Pushpanathan K, Yew SME, et al.: Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*. 2023, 95:.. 10.1016/j.ebiom.2023.104770
13. Lee TJ, Rao AK, Campbell DJ, Radfar N, Dayal M, Khrais A: Evaluating ChatGPT-3.5 and ChatGPT-4.0 Responses on Hyperlipidemia for Patient Education. *Cureus*. Published Online First: 25 May 2024. 10.7759/cureus.61067

14. Choi J, Oh AR, Park J, Kang RA, Yoo SY, Lee DJ, Yang K: Evaluation of the quality and quantity of artificial intelligence-generated responses about anesthesia and surgery: using ChatGPT 3.5 and 4.0. *Front Med (Lausanne)*. 2024, 11:. 10.3389/fmed.2024.1400153
15. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorgia V, Giannaccare G: Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Sci Rep*. 2023, 13:. 10.1038/s41598-023-45837-2
16. Lechien JR, Briganti G, Vaira LA: Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. *European Archives of Oto-Rhino-Laryngology*. 2024, 281:. 10.1007/s00405-023-08441-8
17. Liang R, Zhao A, Peng L, et al.: Enhanced Artificial Intelligence Strategies in Renal Oncology: Iterative Optimization and Comparative Analysis of GPT 3.5 Versus 4.0. *Ann Surg Oncol*. 2024, 31:. 10.1245/s10434-024-15107-0
18. Bill Gates: „Mit KI können Medikamente viel schneller entwickelt werden". Accessed: July 27, 2024. <https://www.handelsblatt.com/technik/ki/bill-gates-mit-ki-koennen-medikamente-viel-schneller-entwickelt-werden/29450298.html>.
19. Eysenbach G: The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. *JMIR Med Educ*. 2023, 9:. 10.2196/46885
20. Sedaghat S: Early applications of ChatGPT in medical practice, education and research. *Clinical Medicine, Journal of the Royal College of Physicians of London*. 2023, 23:. 10.7861/clinmed.2023-0078
21. Lee H: The rise of ChatGPT: Exploring its potential in medical education. *Anat Sci Educ*. 2024, 17:. 10.1002/ase.2270
22. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, Househ M: The Pros and Cons of Using ChatGPT in Medical Education: A Scoping Review. In: *Studies in Health Technology and Informatics*. 2023. 10.3233/SHTI230580
23. Hunter DJ, Holmes C: Where Medical Statistics Meets Artificial Intelligence. *New England Journal of Medicine*. 2023, 389:. 10.1056/nejmra2212850
24. Hamet P, Tremblay J: Artificial intelligence in medicine. *Metabolism*. 2017, 69:S36-40. 10.1016/j.metabol.2017.01.011
25. Sallam M: ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Switzerland)*. 2023, 11:. 10.3390/healthcare11060887
26. Borenstein J, Howard A: Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*. 2021, 1:. 10.1007/s43681-020-00002-7
27. Dave T, Athaluri SA, Singh S: ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023, 6:. 10.3389/frai.2023.1169595

28. Diamond-Brown L: The doctor-patient relationship as a toolkit for uncertain clinical decisions. Soc Sci Med. 2016, 159:. 10.1016/j.socscimed.2016.05.002
29. Duarte-Rojo A, Sejdic E: Artificial Intelligence and the Risk for Intuition Decline in Clinical Medicine. American Journal of Gastroenterology. 2022, 117:. 10.14309/ajg.0000000000001618



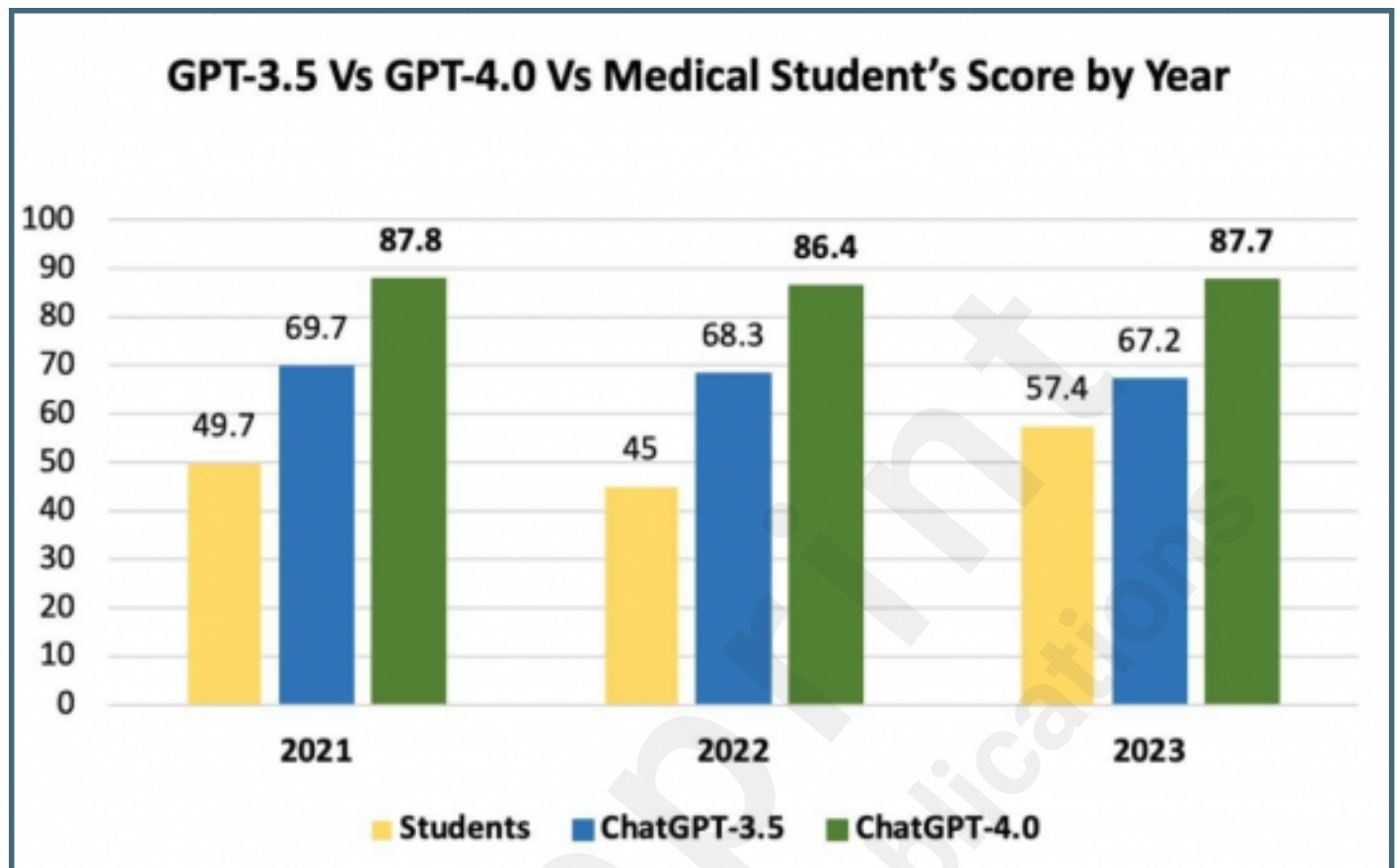
Supplementary Files

Untitled.

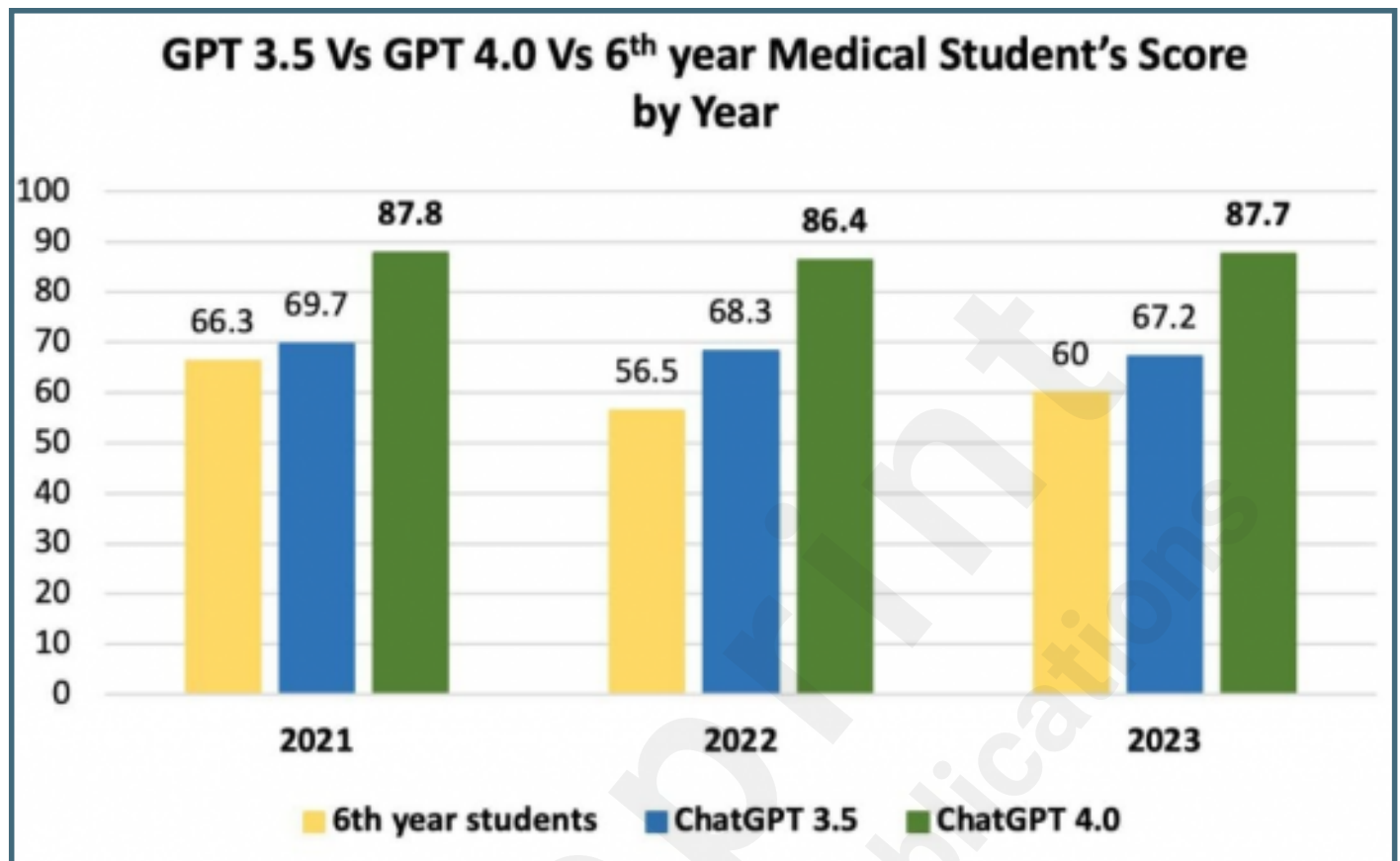
URL: <http://asset.jmir.pub/assets/dc7a198e7e6dcbb548b185893bd1727c.docx>

Figures

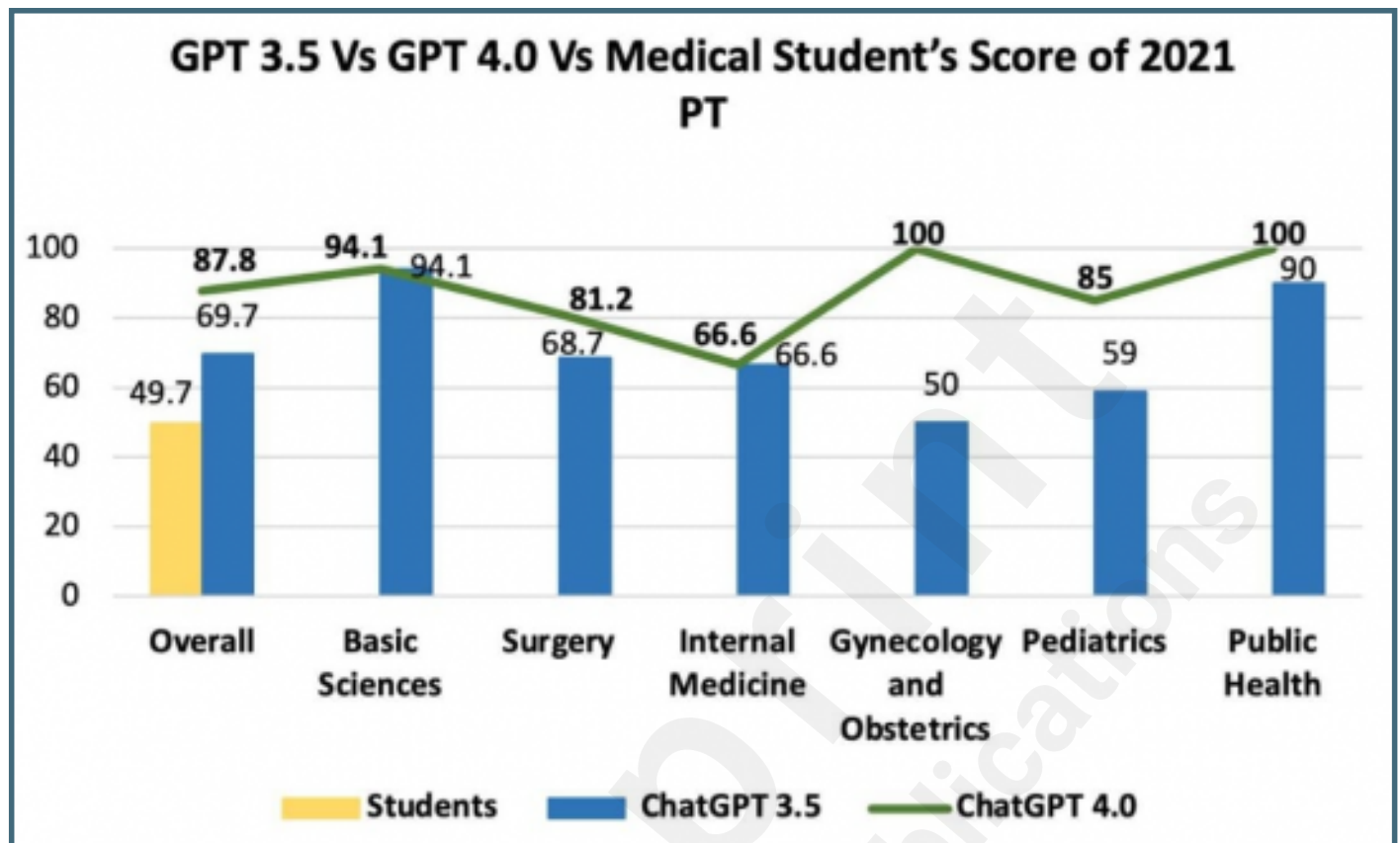
GPT 3.5 Vs GPT 4.0 Vs Medical Student's Score by Year.



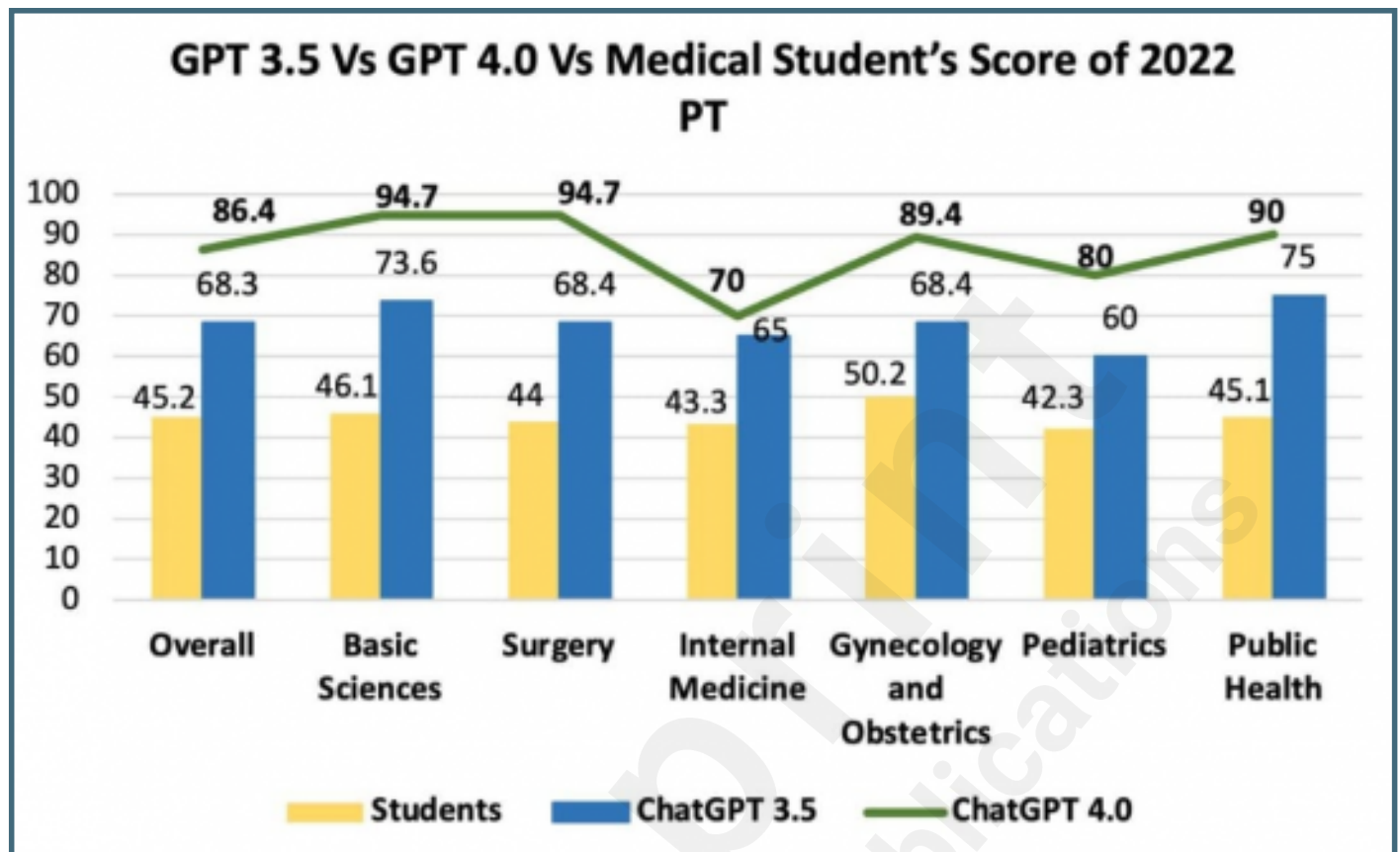
GPT 3.5 Vs GPT 4.0 Vs 6th year Medical Student's Score by Year.



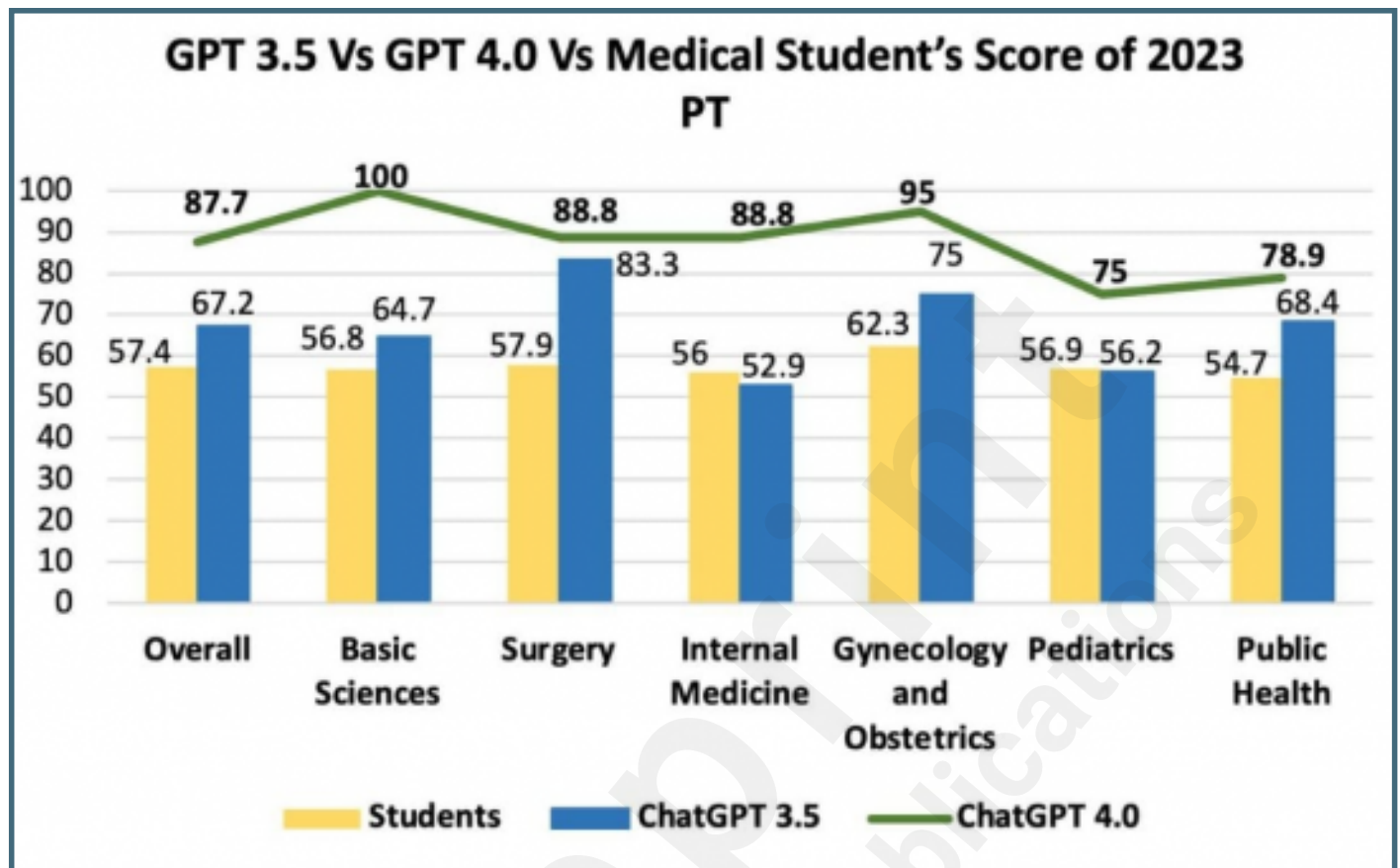
GPT 3.5 Vs GPT 4.0 Vs Medical Student's Score of 2021 PT.



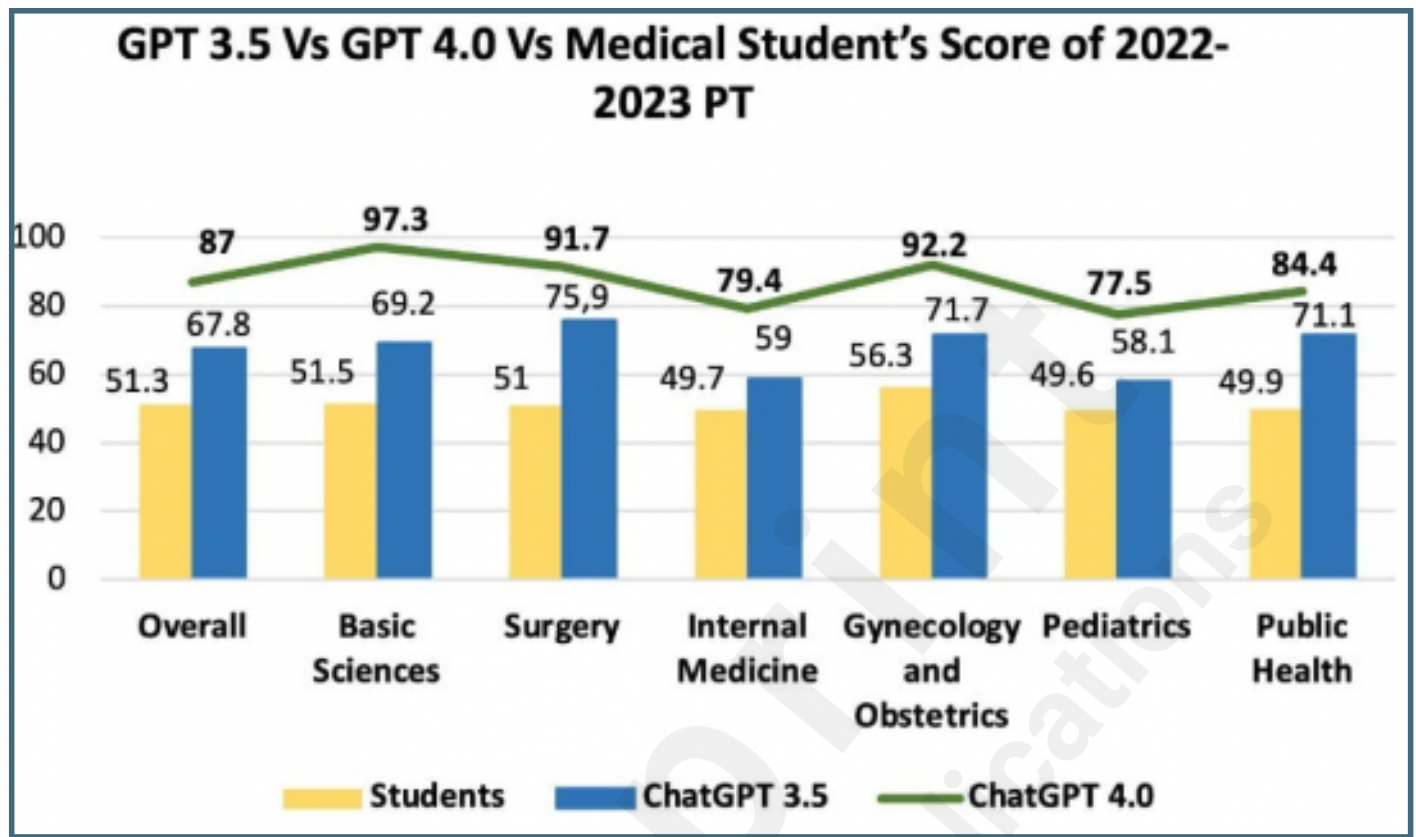
GPT 3.5 Vs GPT 4.0 Vs Medical Student's Score of 2022 PT.



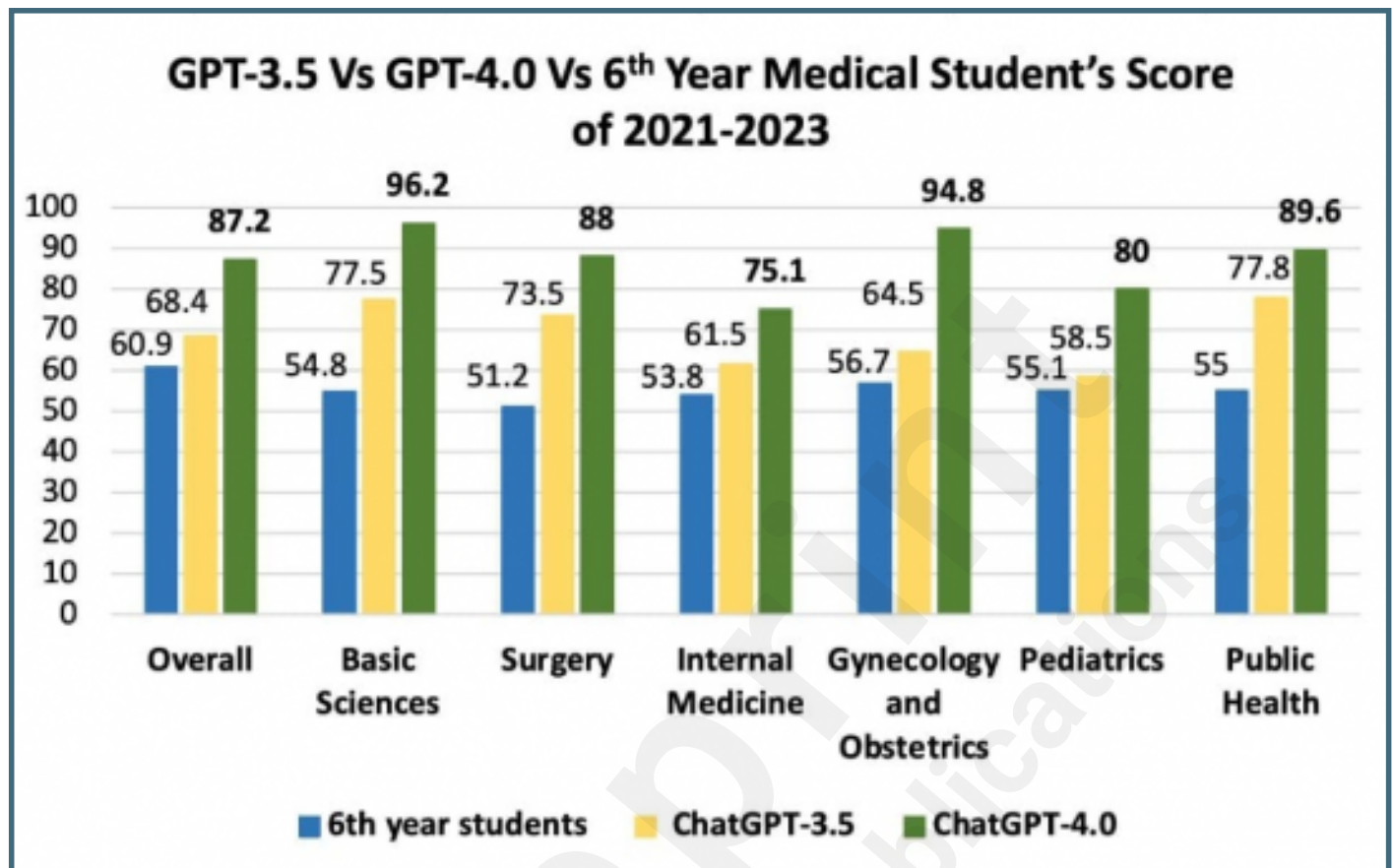
GPT 3.5 Vs GPT 4.0 Vs Medical Student's Score of 2023 PT.



GPT 3.5 Vs GPT 4.0 Vs Medical Student's Score of 2022-2023 PT.



GPT 3.5 Vs GPT 4.0 Vs 6th Year Medical Student's Score of 2021-2023 PT.



GPT 3.5 Vs GPT 4.0 Vs 6th Year Medical Student's Radar Score of 2022-2023 PT.

