Assessing Generative Pretrained Transformers (GPT) in Clinical Decision-Making: A Comparative Analysis of GPT-3.5 and GPT-4

Adi Lahat, Kassem Sharif, Narmin Zoabi, Yonatan Shneor Patt, Yusra Sharif, fisher Fisher, Uria shani, Mohamad Arow, Roni Levin, Eyal Klang

Submitted to: Journal of Medical Internet Research
on: November 15, 2023

Table of Contents

Assessing Generative Pretrained Transformers (GPT) in Clinical Decision-Making: A Comparative Analysis of GPT-3.5 and GPT-4

Adi Lahat[1*] MD; Kassem Sharif[1*] MD; Narmin Zoabi[1] MD; Yonatan Shneor Patt[1] MD; Yusra Sharif[2]; fisher Fisher[1] MD; Uria shani[1] MD; Mohamad Arow[1] MD; Roni Levin[1] MD; Eyal Klang[1] MD

[1]Chaim Sheba Medical Center Ramat Gan IL
[2]Hadassah Medical Cenrtre Jerusalem IL
[*]these authors contributed equally

Corresponding Author:
Adi Lahat MD
Chaim Sheba Medical Center
Tel Hashomer
Ramat Gan
IL

Abstract

Background: Artificial Intelligence (AI), particularly chatbot systems, is becoming an instrumental tool in healthcare, aiding clinical decision-making and patient engagement

Objective: To analyze the performance of Chat GPT-3.5 and Chat GPT-4 in addressing complex clinical and ethical dilemmas, and to illustrate their potential role in healthcare decision-making

Methods: Four specialized physicians formulated 176 real-world clinical questions. Both senior physicians and residents evaluated the answers generated by GPT-3.5 and GPT-4 on 1-5 scale in 5 categories: accuracy, relevance, clarity, beneficial, Completeness.

Results: Both GPT models received high scores (4.4 ± 0.8 for GPT-4 ,4.1 ± 1.0 for GPT-3.5).GPT-4 outperformed GPT-3.5 across all rating dimensions, with seniors consistently rating responses higher than residents for both models. Specifically, seniors rated GPT-4 as more beneficial and complete (4.6 vs 4.0 and 4.6 vs 4.1, respectively, $p<0.001$), and GPT-3.5 similarly (4.1 vs 3.7 and 3.9 vs 3.5, $p<0.001$). Ethical queries received the highest ratings for both models, with mean scores reflecting consistency across accuracy and completeness criteria. Distinctions among question types were significant, particularly for GPT-4's completeness across emergency, internal, and ethical questions (4.2 ± 1.0, 4.3 ± 0.8, 4.5 ± 0.7; $p < 0.001$), and for GPT-3.5's accuracy, beneficial, and completeness dimensions

Conclusions: Chat GPT's potential to assist physicians with medical issues is promising, with prospects to enhance diagnostics, treatments, and ethics. While integration into clinical workflows may be valuable, it must complement, not replace, human expertise. Continued research is essential to ensure safe and effective implementation in clinical environments. Clinical Trial: N/A

(JMIR Preprints 15/11/2023:54571)
DOI: https://doi.org/10.2196/preprints.54571

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 v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Assessing Generative Pretrained Transformers (GPT) in Clinical Decision-Making: A Comparative Analysis of GPT-3.5 and GPT-4

Adi Lahat MD [1,2*], Kassem Sharif MD *[1,3], Narmin Zoabi MD[1], Yonatan Shneor Patt MD[3], Yusra Sharif MD[4], Lior Fisher MD[3] , Uria Shani MD[3] , Mohamad Arow MD[3], Roni Levin MD[3] , Eyal Klang MD [5]

*Equal contribution

1- Department of Gastroenterology, Chaim Sheba Medical Center, affiliated with Tel Aviv University, Tel Aviv, Israel

2- Department of Gastroenterology, Samson Assuta Ashdod Medical Center, affiliated with Ben Gurion University of the Negev, Be'er Sheva, Israel

3- Department of Internal Medicine B, Sheba Medical Centre, Tel Aviv, Israel

4-Department of Internal Medicine C, Hadassah Medical Center, Jerusalem, Israel

5 – Division of Data-Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, New York, USA

Corresponding Author:
Adi Lahat, MD
Department of Gastroenterology, Chaim Sheba Medical Center, affiliated with Tel Aviv University, Tel Aviv, Israel
zokadi@gmail.com
word count: 3760

Keywords: chatGPT, natural language processing (NLP), artificial intelligence (AI), Internal medicine, Ethics, emergency medicine
Financial support: The manuscript had no financial support
Potential competing interests: None of the authors has any conflict of interest to declare
Author contributions: AL, KS and EK designed and performed the research and wrote the paper; NZ,YSP,YS,LF,US,MA,RL answered the questions, contributed to the analysis, and revised the paper critically. All authors approved the version to be published.
Data availability statement: The authors declare that there is no relevant data available for this study. All data used in the analysis and preparation of this manuscript have been included in the manuscript.

Abstract

Background*:* Artificial Intelligence (AI), particularly chatbot systems, is becoming an instrumental tool in healthcare, aiding clinical decision-making and patient engagement.

Aim: To analyze the performance of Chat GPT-3.5 and Chat GPT-4 in addressing complex clinical and ethical dilemmas, and to illustrate their potential role in healthcare decision-making, while comparing senior's and residents' rating, and specific questions types.

Methods: Four specialized physicians formulated 176 real-world clinical questions. Eight senior physicians and residents assessed responses from GPT-3.5 and GPT-4 on a 1-5 scale across five categories: accuracy, relevance, clarity, utility, and comprehensiveness. Evaluations were conducted within internal medicine, emergency medicine, and ethics. Comparisons were made globally, between seniors and residents, and across classifications.

Results: Both GPT models received high scores (4.4 ± 0.8 for GPT-4 ,4.1 ± 1.0 for GPT-3.5). GPT-4 outperformed GPT-3.5 across all rating dimensions, with seniors consistently rating responses higher than residents for both models. Specifically, seniors rated GPT-4 as more beneficial and complete (4.6 vs 4.0 and 4.6 vs 4.1, respectively, $p<0.001$), and GPT-3.5 similarly (4.1 vs 3.7 and 3.9 vs 3.5, $p<0.001$). Ethical queries received the highest ratings for both models, with mean scores reflecting consistency across accuracy and completeness criteria. Distinctions among question types were significant, particularly for GPT-4's completeness across emergency, internal, and ethical questions (4.2 ± 1.0, 4.3 ± 0.8, 4.5 ± 0.7; $p < 0.001$), and for GPT-3.5's accuracy, beneficial, and completeness dimensions.

Conclusion: Chat GPT's potential to assist physicians with medical issues is promising, with prospects to enhance diagnostics, treatments, and ethics. While integration into clinical workflows may be valuable, it must complement, not replace, human expertise. Continued research is essential to ensure safe and effective implementation in clinical environments.

Introduction

Artificial Intelligence (AI), particularly chatbot systems, is becoming an instrumental tool in healthcare, aiding clinical decision-making and patient engagement [1,2]. These systems, utilizing natural language processing (NLP), offer support to physicians in various clinical scenarios [3].

OpenAI's GPT technology represents a leading example of this innovation [4].

The integration potential of AI in healthcare is not only a technological advancement; it's a paradigm shift in how medical professionals approach patient care. Leveraging AI, clinicians can access real-time data analysis, personalized treatment recommendations, and even predictive insights into patient health trends. The convergence of AI with traditional medical practices is fostering a new era of precision medicine, where treatments are tailored to individual patient needs and preferences. This personalized approach enhances patient satisfaction and outcomes, while also optimizing resource allocation within the healthcare system.

Earlier GPT models like GPT-3.5 were recognized for generating human-like text and have potential across various fields, including medicine, where they can assist in diagnostics, therapeutic planning, telehealth, and patient education [5-15]. Recent AI algorithms ability to synthesize vast information and recognize complex patterns, can enhance clinical accuracy, and reduce errors [16].

With the introduction of GPT-4, an advancement in AI's capabilities, it becomes imperative to evaluate its comparative effectiveness in clinical scenarios [17]. This model, with improved language comprehension and generation, presents a promising opportunity for refined healthcare applications. Past studies have highlighted the clinical potential of ChatGPT (GPT-3.5), which managed to score approximately 60% on the USMLE [18]. Moreover, the more advanced GPT-4 achieved an impressive 87% [19].

Recent advancements in AI technology have paved the way for innovative applications in medical education and patient support, particularly through AI-powered chatbots. Chow et al. [20] developed a chatbot using the IBM Watson Assistant platform to make radiotherapy knowledge more accessible to the general public. This chatbot, designed with a user-friendly interface, engages users in an interactive question-and-answer format, enhancing their understanding of radiotherapy. Despite its positive feedback, the study highlights the need for improvements in conversational approaches and language inclusivity. In another study, Rebelo et al. [21] created a novel virtual assistant to explain the radiation treatment process comprehensively. Utilizing IBM Watson's natural language processing features, this chatbot guides users through the complex radiotherapy process, from diagnosis to treatment delivery, and has been tested for effectiveness in knowledge transfer. Both studies underscore the potential of AI chatbots in transforming medical communication and education, offering personalized, interactive, and reassuring platforms for patients and the general public to access critical health information. These initiatives reflect the growing importance of AI in addressing the need for accessible and accurate medical information, particularly in areas like radiotherapy, where patient and family education is crucial.

However, the reliability of ChatGPT in delivering health information to patients and healthcare providers remains questionable. ChatGPT differs from specialized medical chatbots as it is not trained on datasets curated by medical professionals, raising concerns about the accuracy of its medical advice and ethical issues related to patient safety and data privacy. As ChatGPT evolves as a disruptive technology in healthcare, it faces several challenges. These include its database being possibly outdated, risks of misinformation, and the need for its integration into the medical field to be carefully monitored and guided by ethical frameworks and professional standards. [22]

In light of all of the above, the goal our current study is to analyze the performance of Chat GPT-3.5 and Chat GPT-4 in addressing complex clinical and ethical dilemmas, and to illustrate their potential

role as allies in healthcare decision-making.

Our study aims to compare the free version of GPT-3.5 with the paid GPT-4 in a medical context, focusing on accessibility and performance for a diverse audience. This comparison is vital for understanding the practicality and effectiveness of AI in healthcare, especially for users with limited resources. By evaluating both versions, we can assess the performance gap, guide resource allocation, and address the democratization of AI technology in medicine. This approach not only helps in setting realistic expectations for the free version's capabilities but also underscores the ethical and social implications of AI accessibility. Such a comparison is crucial for informing both users and developers about the practical use of AI in medical contexts, ensuring the study remains relevant and beneficial to a wide range of potential users, including those without access to the paid version.

Furthermore, we aimed to assess and compare the performance of ChatGPT in responding to medical questions across different categories, namely emergency medicine, internal medicine, and ethical dilemmas. We specifically focus on comparing the ratings provided by two distinct groups: medical residents and senior physicians. The rationale behind this comparison lies in the differing levels of clinical experience and expertise between these groups. Medical residents, being in the earlier stages of their training, may approach medical questions with a fresh perspective and rely more on foundational knowledge and guidelines. On the other hand, senior physicians, with their extensive clinical experience, may offer nuanced insights and consider broader contextual factors in their assessments. By evaluating responses from both groups, we aimed to gain a comprehensive understanding of ChatGPT's performance across various medical domains and discern potential differences in the depth and accuracy of responses based on clinical experience levels. We believe that this comparison can provide valuable insights into the utility and limitations of using AI language models like ChatGPT in supporting medical decision-making across different stages of clinical training and practice, and to contribute to a broader understanding of AI's expanding role in

the medical field.

Methods

Study Design:

In this study, three forms of comparisons were conducted to evaluate the performance of ChatGPT in responding to medical questions. Firstly, we compared the responses generated by two versions of the model, GPT-3.5 and GPT-4, to assess any differences in the quality and accuracy of responses between the two iterations. This comparison aimed to elucidate potential improvements or variations in performance introduced by advancements in the underlying AI architecture. Secondly, we compared the evaluations provided by senior physicians and medical residents on the responses generated by ChatGPT. This analysis aimed to explore potential disparities in the perceived quality and usefulness of responses based on the level of clinical experience and expertise of the evaluator. Lastly, we analyzed responses across different types of medical questions, including emergency medicine, internal medicine, and ethical dilemmas, to assess the model's performance across diverse clinical scenarios. This comparison aimed to identify any variations in the model's ability to provide relevant and accurate information across various medical domains. By conducting these three forms of comparisons, we aimed to comprehensively evaluate the capabilities of ChatGPT in addressing medical queries across different contexts and user perspectives.

Question Development and Categorization:

Four experienced physicians specialized in internal medicine and emergency medicine collaborated to formulate 176 questions. Questions were generated based on clinical experience and designed to follow clinical guidelines.

The questions were written by two senior physicians in internal medicine and emergency medicine with reference to board questions used for the assessment of doctors in training, questions were chosen to reflect different aspects of medicine as seen in everyday practice. Verification was done independently by two different physicians and answers were checked to comply with updated

guidelines.

Questions were mainly classified to one of three categories -

ER: Questions addressing emergency scenarios.

Internal Medicine: Questions concerning diagnosis, treatment, and management of internal medicine conditions.

Ethical: Questions focusing on ethical considerations in medical practice.

These questions were created to reflect real-world clinical scenarios, aligning with prevailing clinical guidelines and ethical norms.

Selection and Configuration of Models:

The study engaged OpenAI's GPT-3.5 and GPT-4 models, recognized for their advanced capabilities in natural language processing. They were queried through OpenAI's API to generate answers to the formulated questions.

Prompting the Models:

The models were prompted using the specific instruction:

"Please answer this {question_type} question in clear, concise, concrete, full, bullets: {question}".

Grading Participants and Process:

Eight clinicians from two tertiary medical centers, including four senior physicians and four residents, independently graded the responses. Senior physicians were all active in daily clinical practice, including emergency care within the Emergency Room (ER) department, between 2-4 years subsequent to the completion of residency training. All residents were in their last year of residency (3 out of 4 years), and experienced with internal and emergency medicine.

Grading criteria were as follows:

Accuracy: Does it reflect medical understanding (Scale 1-5). The answers were graded as follows:

1: Very Inaccurate - The response shows a fundamental misunderstanding of medical concepts.

2: Somewhat Inaccurate - The answer contains more incorrect than correct medical information.

3: Moderately Accurate - The response is generally correct but includes some inaccuracies.

4: Mostly Accurate - The answer is largely accurate with minor errors or omissions.

5: Completely Accurate - The response reflects a high level of medical understanding with accurate and precise information.

Relevance: Does the response directly address the asked question, or does it deviate to unrelated subtopics (Scale 1-5). The answers were graded as follows:

1: Not Relevant - The response is completely off-topic or unrelated to the question.

2: Slightly Relevant - The answer addresses the question but includes significant unrelated information.

3: Moderately Relevant - The response is relevant but includes some tangential content.

4: Highly Relevant - The answer is directly related to the question with minimal unrelated details.

5: Completely Relevant - The response precisely addresses the question without any deviation.

Clarity: How clear is the provided information (Scale 1-5). The answers were graded as follows:

1: Very Unclear - The response is confusing, poorly articulated, or difficult to comprehend.

2: Somewhat Clear - The answer has some clarity but may require additional explanation.

3: Clear - The response is understandable with a reasonable level of clarity.

4: Very Clear - The answer is well-explained and easy to follow.

5: Exceptionally Clear - The response is articulated in an exceptionally straightforward and comprehensible manner.

Beneficial: Does the response significantly aid the decision-making process (Scale 1-5). The answers were graded as follows:

1: Not Beneficial - The response provides no useful aid for decision-making.

2: Slightly Beneficial - The answer offers limited assistance in the decision-making process.

3: Moderately Beneficial - The response is somewhat helpful for decision-making.

4: Highly Beneficial - The answer significantly aids the decision-making process.

5: Extremely Beneficial - The response is exceptionally valuable and decisively aids in making informed decisions.

Completeness: Does the response cover all necessary information required to fully answer the question (Scale 1-5). The answers were graded as follows:

1: Very Incomplete - The response leaves out crucial information necessary to answer the question.

2: Somewhat Complete - The answer includes some necessary information but is missing significant aspects.

3: Moderately Complete - The response covers a fair amount of the necessary information.

4: Mostly Complete - The answer is almost complete with only minor omissions.

5: Fully Complete - The response comprehensively covers all information required to fully answer the question.

A comprehensive grading guide ensured uniformity in the application of the criteria.

Statistical Analysis:

To evaluate the performance differences between GPT-3.5 and GPT-4, the non-parametric Mann-Whitney U tests were utilized. The Mann-Whitney U test was chosen to compare GPT-3.5 and GPT-4 scores, as our data are ordinal. This non-parametric test is ideal for data not meeting normal distribution assumptions. This non-parametric test was also employed to discern grading disparities between senior physicians and residents, representing two heterogeneous grader groups with potentially varying evaluation patterns. Furthermore, to probe into the potential differences in the responses based on the three distinct question categories (Emergency Medicine, Internal Medicine, and Ethical Questions), Mann-Whitney U tests were once again applied.

For all statistical tests, a p-value of less than .05 was considered indicative of a statistically significant difference.

Results

Overview of ChatGPT performance

An evaluation was conducted to assess the performance of two generative models, GPT-4 and GPT-3.5, in terms of accuracy, relevance, clarity, benefit to the decision-making process, and completeness of the response. A total of 176 questions were analyzed by both residents and seniors, yielding a consistent and objective assessment. Examples of Emergency Medicine, Internal Medicine, and Ethical Questions and the corresponding GPT-3.5 and GPT-4 detailed answers are presented in **Table 1**.

Table 1 – The table below illustrates a comparison between the responses of ChatGPT-4 and ChatGPT-3 to questions from three different disciplines: Emergency Medicine, Internal Medicine, Ethical dilemmas

Discipline	Question	ChatGPT-4	ChatGPT-3
Emergency medicine	How would you manage a patient presenting with signs of necrotizing fasciitis?	• Immediately initiate broad-spectrum intravenous antibiotics, including coverage for gram-positive, gram-negative, and anaerobic bacteria. • Administer intravenous fluid resuscitation to maintain hemodynamic stability. • Provide the patient with adequate analgesia for pain management. • Consult a surgeon urgently for possible surgical debridement. Necrotizing fasciitis is a surgical emergency and tissue debridement is critical in management. • Order laboratory tests including complete blood count, electrolytes, renal function, liver function, coagulation profile, and blood cultures. • Obtain imaging such as X-ray, CT scan, or MRI if diagnosis is uncertain. However, do not delay treatment for imaging. • Consider administration of intravenous immunoglobulin	• Call for emergency medical assistance immediately • Begin supportive care, including oxygenation and intravenous fluids • Administer broad-spectrum antibiotics that cover both gram-positive and gram-negative bacteria • Obtain blood cultures and wound cultures • Surgical consultation should be obtained immediately • Consider early surgical intervention, including debridement of affected tissue • Monitor the patient closely for signs of sepsis and organ dysfunction • Administer pain medication as needed • Educate the patient and family about

		(IVIG) as adjunctive therapy. • Monitor the patient closely in an intensive care setting. • Provide supportive care including nutritional support and prevention of secondary infections. • Consider tetanus prophylaxis if the patient's immunization status is unknown or not up-to-date. • Arrange for psychosocial support as necrotizing fasciitis can have significant impact on mental health. • Implement contact precautions to prevent spread of infection to other patients and healthcare workers.	the disease and potential complications
Internal Medicine	How should you handle a patient with known cirrhosis presenting with new-onset ascites?	- Confirm the diagnosis: Ascertain the patient's symptoms such as abdominal bloating, weight gain, and discomfort. Perform physical examination to assess for fluid wave or shifting dullness. - Order diagnostic tests: Arrange for ultrasound or CT scan to confirm ascites. A paracentesis may be performed to analyze the ascitic fluid and rule out infection or malignancy. - Manage underlying cirrhosis: Ensure the patient is on optimal therapy for cirrhosis, which may include medications like diuretics, beta-blockers, or lactulose. - Sodium restriction: Advise the patient to limit their sodium intake to less than 2 grams per day. - Diuretics: Start the patient on a diuretic regimen, often spironolactone with or without furosemide, to help mobilize the fluid. - Monitor for complications: Regularly monitor the patient for complications of cirrhosis and ascites such as hepatic encephalopathy, spontaneous bacterial peritonitis, and hepatorenal syndrome. - Consider referral for liver transplantation: If the patient's condition continues to worsen despite optimal medical therapy, consider referral to a transplant center for evaluation for liver transplantation..	- Obtain a detailed medical history, including the duration and severity of ascites, any recent changes in medication or diet, and any history of alcohol or drug abuse. - Perform a thorough physical examination, including a careful assessment of the abdomen for fluid accumulation, signs of liver disease, and any evidence of hepatic encephalopathy. - Order laboratory tests, including liver function tests, complete blood count, electrolytes, and renal function tests. - Monitor the patient closely for signs of complications, such as infection, renal failure, or hepatic encephalopathy, and adjust treatment as needed. - Educate the patient about the importance of adhering to a low-sodium diet, avoiding alcohol and other hepatotoxic substances, and following up regularly with their healthcare provider to monitor their condition.

		- Patient education: Educate the patient about the importance of medication adherence, dietary sodium restriction, and regular follow-up. Also inform about the signs of complications	
Ethical questions	Should a healthcare provider give placebos to patients who insist on receiving treatment when no treatment is necessary?	"- Ethically, healthcare providers should not give placebos to patients who insist on receiving treatment when no treatment is necessary. - The use of placebos without patient consent is deceptive and violates the principle of patient autonomy and informed consent, which are fundamental to medical ethics. - It can undermine the trust between the patient and the healthcare provider. If discovered, it could damage the patient's confidence in their healthcare provider or the healthcare system as a whole. - It may also delay necessary treatment if the patient's condition changes or worsens, as the placebo effect is temporary and does not treat underlying conditions - However, some argue that if a placebo can alleviate patient anxiety or perceived symptoms, it may be considered in the patient's best interest. This approach should only be taken with full transparency and patient consent. - In conclusion, it is generally considered unethical to administer placebos without patient knowledge or consent, and it is crucial to respect patient autonomy, promote open communication, and provide appropriate care based on medical necessity."	- • No, healthcare providers should not give placebos to patients who insist on receiving treatment when no treatment is necessary. • Giving placebos is unethical and goes against the principle of informed consent. • It can also harm the patient's trust in healthcare providers and the medical profession as a whole. • Healthcare providers should educate patients on the lack of necessity for treatment and offer alternative forms of support, such as counseling or therapy. • If a patient continues to insist on treatment, it is important to have an open and honest conversation about the potential risks and benefits of treatment and the lack of evidence supporting the use of placebos.

GPT-4 demonstrated statistically significant superiority in all assessed criteria compared to GPT-3.5. The analysis shows that the models differed notably in their ability to provide clear, relevant, and beneficial responses, with GPT-4 consistently outperforming GPT-3.5. The results of GPT-4 and GPT-3.5 are summarized in **Table 2**.

Table 2: Comparative Analysis of GPT-4 and GPT-3.5 Performance

Criteria	GPT-4 mean ±SD	GPT-3.5 mean ±SD	p-value
Accuracy	4.5 ± 0.8	4.2 ± 0.9	<.001
Relevance	4.5 ± 0.7	4.4 ± 0.8	<.001
Clarity	4.6 ± 0.7	4.4 ± 0.9	<.001
Beneficial	4.3 ± 0.9	3.9 ± 1.0	<.001
Completeness	4.3 ± 0.9	3.7 ± 1.1	<.001
Total	4.4 ± 0.8	4.1 ± 1.0	<.001

Distribution of Ratings for GPT-4 and GPT-3.5

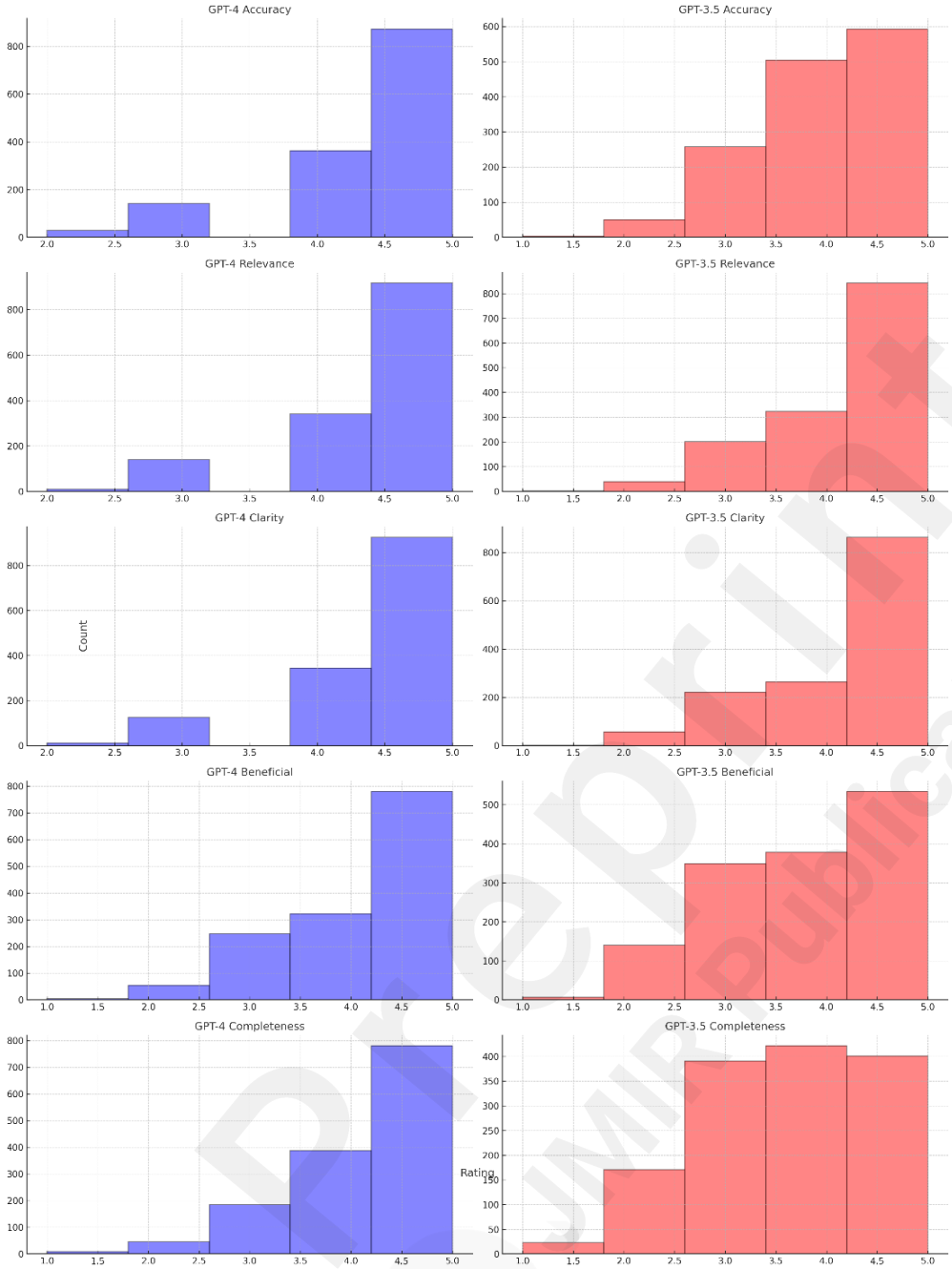
A visual inspection of the distributions of ratings for GPT-4 and GPT-3.5, provided through histograms (**Figure 1**), further elucidates the comparative performance of these models across all five rating dimensions: Accuracy, Relevance, Clarity, Benefit, and Completeness.

For all the rating dimensions, GPT-4 generally received higher ratings than its counterpart, GPT-3.5. The mode - the most frequently given rating - for GPT-4 consistently achieved the maximum score of 5 across all categories, with the exception of the Benefit category, where the mode was approximately 4.5.

In contrast, the distribution of ratings for GPT-3.5 presented a wider spread, with modes fluctuating between scores of 3 and 5, contingent on the category. In particular, the categories of Benefit and Completeness exhibited a broad spread of ratings for GPT-3.5, indicative of a greater variability in

the responses.

Figure 1: A visual inspection of the distributions of ratings for GPT-4 and GPT-3.5



Comparison Between Residents and Seniors

In evaluating the models' responses based on the reviewer type (**Table 3**), significant differences were discerned between residents and senior physicians.

Table 3: Comparison between residents and seniors' assessments of GPT-4 and GPT-3.5 answers.

Criteria	Residents GPT-4 mean ±SD	Seniors GPT-4 mean ±SD	p-value GPT-4	Residents GPT-3.5 mean ±SD	Seniors GPT-3.5 mean ±SD	p-value GPT-3.5
Accuracy	4.3 ± 0.9	4.7 ± 0.5	<.001	4.1 ± 1.0	4.3 ± 0.7	<.001
Relevance	4.3 ± 0.8	4.7 ± 0.6	<.001	4.2 ± 0.9	4.6 ± 0.8	<.001
Clarity	4.4 ± 0.7	4.7 ± 0.6	<.001	4.2 ± 0.9	4.5 ± 0.9	<.001
Beneficial	4.0 ± 1.0	4.6 ± 0.7	<.001	3.7 ± 1.0	4.1 ± 1.0	<.001
Completeness	4.1 ± 1.0	4.6 ± 0.6	<.001	3.6 ± 1.2	3.9 ± 0.9	<.001

For both GPT-4 and GPT-3.5, seniors consistently rated the responses higher across all criteria. The variation was most notable in the assessment of the models' benefits and completeness.

In the case of GPT-4, senior physicians rated it as more beneficial (4.6 vs 4.0, p<.001) and more complete (4.6 vs 4.1, p<.001) compared to the residents. A similar trend was observed for GPT-3.5, with seniors appreciating its benefits (4.1 vs 3.7, p<.001) and completeness (3.9 vs 3.5, p<.001) more than the residents. This finding suggests a potential influence of clinical experience on the perception of AI-generated responses.

Performance Across Question Types

In the evaluation of both GPT-4 and GPT-3.5, distinctions were evident across question types

(Emergency Medicine, Internal Medicine, Ethical Questions). Ethical queries consistently received the highest ratings (**Figure 2** and **Figure 3** presents analysis for GPT-4 and GPT-3.5 across all reviewers, while **table 4** and **table 5** presents analysis separately for residents and seniors). Higher grades were given for ethical answers by residents and seniors alike, and in both models of GPT. (**Table 4** and **Table 5**).

Figure 2: Performance of GPT-3.5 according to question subject

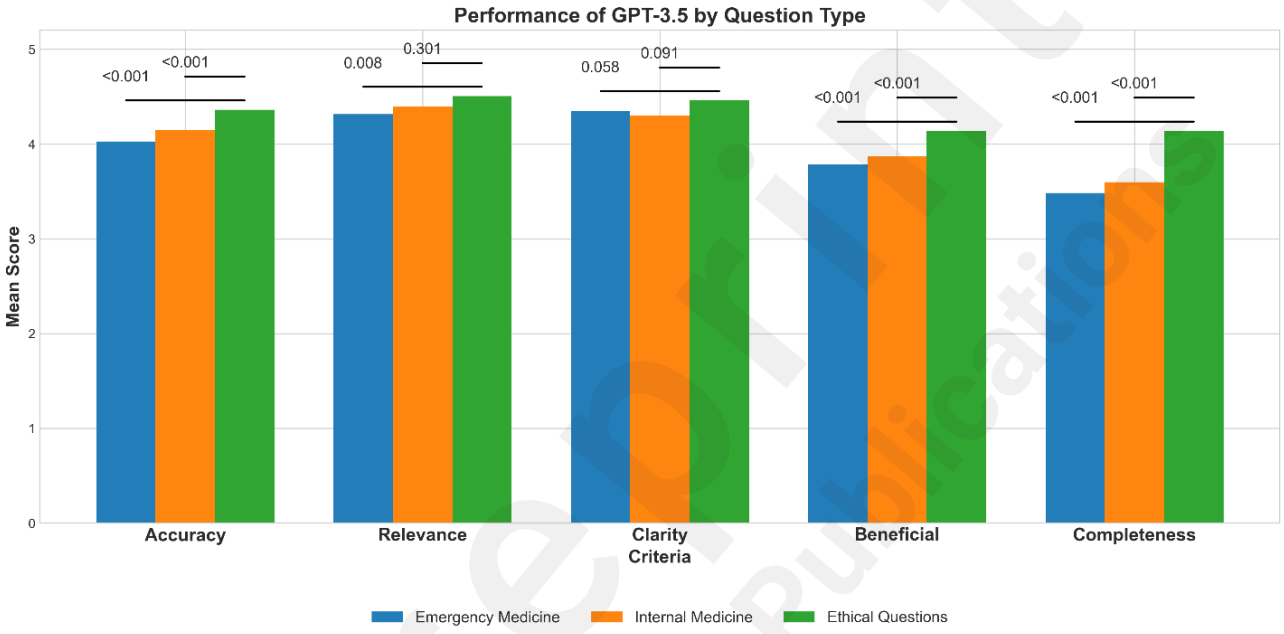


Figure 3: Performance of GPT-4 according to question subject

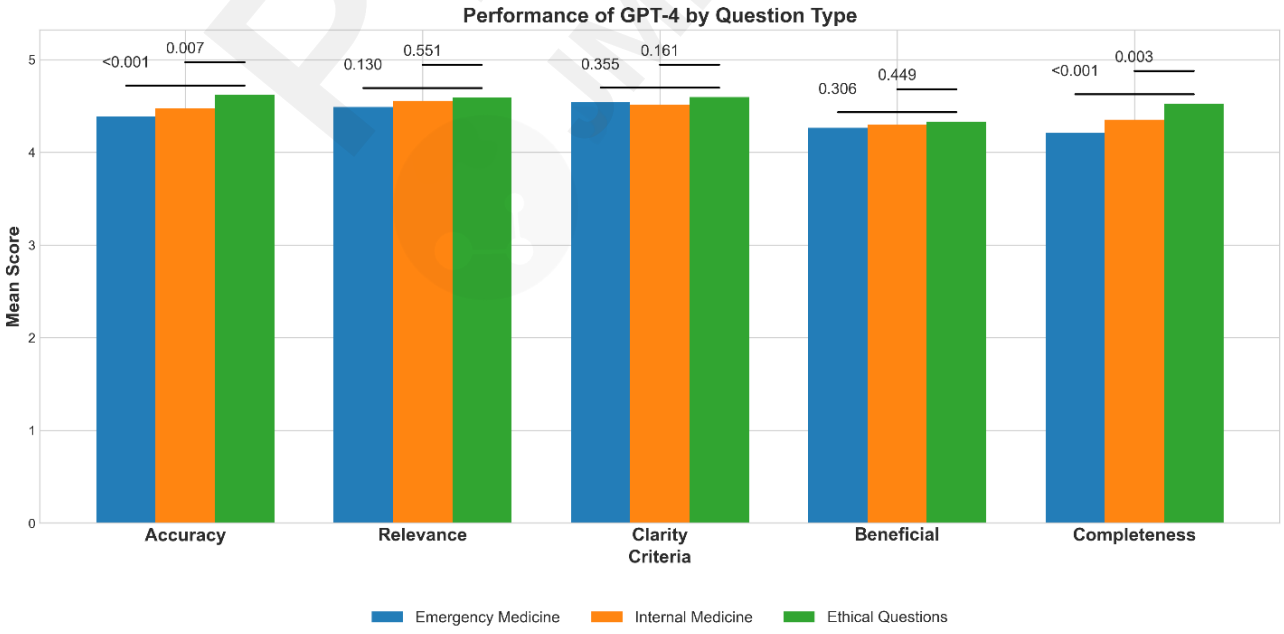


Table 4a: Comparison between residents' assessment of medical and ethical GPT-4 answers

Criteria	Medical_GPT-4 mean ±SD	Ethical_GPT-4 mean ±SD	p-value
Accuracy	4.2±0.9	4.4±0.8	.172
Relevance	4.3±0.8	4.4±0.7	.522
Clarity	4.4±0.7	4.5±0.7	.167
Beneficial	3.9±1.0	4.1±0.9	.028
Completeness	4.0±1.0	4.3±0.8	<.001
Total	4.2±0.9	4.3±0.8	<.001

Table 4b: Comparison between seniors' assessment of medical and ethical GPT-4 answers

Criteria	Medical_GPT-4 mean ±SD	Ethical_GPT-4 mean ±SD	p-value
Accuracy	4.6±0.6	4.9±0.3	<.001
Relevance	4.7±0.6	4.8±0.5	.135
Clarity	4.7±0.7	4.7±0.6	0.852
Beneficial	4.6±0.7	4.5±0.8	.633
Completeness	4.5±0.7	4.8±0.5	<.001
Total	4.6±0.7	4.7±0.6	<.001

Table 5a: Comparison between residents' assessment of medical and ethical GPT-3.5 answers

Criteria	Medical_GPT-3.5 mean ±SD	Ethical_GPT-3.5 mean ±SD	p-value
Accuracy	4.0±1.0	4.2±0.9	.079
Relevance	4.1±0.9	4.3±0.8	.004
Clarity	4.2±0.9	4.3±0.8	.105
Beneficial	3.6±1.0	3.9±1.0	<.001
Completeness	3.4±1.2	3.9±1.0	<.001
Total	3.9±1.0	4.1±0.9	<.001

Table 5b: Comparison between seniors' assessment of medical and ethical GPT-3.5 answers

Criteria	Medical_GPT-3.5 mean ±SD	Ethical_GPT-3.5 mean ±SD	p-value
Accuracy	4.1±0.8	4.6±0.5	<.001
Relevance	4.6±0.9	4.7±0.6	.699
Clarity	4.5±1.0	4.6±0.8	.161
Beneficial	4.0±1.0	4.4±1.0	<.001
Completeness	3.7±0.9	4.3±0.7	<.001
Total	4.2±1.0	4.5±0.8	<.001

For GPT-4, mean scores were fairly consistent across the criteria of Accuracy, and completeness, with significant differences among the question types (**Figure 1**). Specifically, Completeness revealed statistically significant differences among Emergency, internal, and ethical questions (4.2 ± 1.0, 4.3 ± 0.8, 4.5 ± 0.7; p <.001).

For GPT-3.5, differences among question types were often statistically significant, as demonstrated in **Figure 2**. Accuracy ranged from 4.0 ± 0.9 for ED questions to 4.4 ± 0.8 for ethical questions (p <.001). Similar trends were observed in Beneficial (p <.001), and Completeness (p <.001).

Discussion

In our study, we evaluated Chat GPT-3.5 and Chat GPT-4's ability to address complex clinical and ethical dilemmas. Both models showed promise as tools to aid physicians in decision-making. Notably, Chat GPT-4 outperformed Chat GPT-3.5 across all parameters. This was validated by both resident and senior medical practitioners.

In an effort to provide a thorough, wide-ranging, and accurate evaluation, our study included an extensive compilation of queries, totaling 176 questions. These questions spanned various topics found within the domains of Emergency Medicine, Internal Medicine, and Ethical considerations, mirroring the real-life challenges that doctors frequently face in their daily clinical work. Both groups of evaluators - senior physicians and interns - assessed each question, aiming to capture diverse perspectives on the perceived benefits at different levels of expertise. To the best of our knowledge, this is the most exhaustive study conducted on this particular subject.

Overall, both Chat GPT models received high grades in terms of accuracy, relevance, clarity, benefit, and completeness. However, GPT-4 scored higher in all criteria assessed, including total mean grades (**Table 2**).

These findings are promising when considering the potential benefits of implementing an NLP model like Chat GPT into the field of medicine, and in agreement with current literature]10-15[. The

visible advantages are manifold, with one of the standout features being the rapid retrieval of information and examination of the literature. Chat GPT's strength lies in its capacity to quickly access a wide array of medical data from various sources. By offering doctors immediate entry to the newest findings, clinical standards, and specific cases, Chat GPT acts as a catalyst for keeping them aligned with the ever-changing medical landscape. This ability enhances physicians' capacity to make educated judgments when dealing with intricate or uncommon medical scenarios.

Additionally, Chat GPT's aptitude in understanding natural language equips it to thoroughly examine patient symptoms and medical backgrounds. This can lead to the suggestion of possible diagnoses, as well as offering alternative diagnoses for reflection. While it doesn't substitute for hands-on clinical expertise, Chat GPT proves to be an invaluable asset in assisting physicians to pinpoint diagnostic avenues and contemplate less evident conditions. [16]

Furthermore, Chat GPT's skill in scrutinizing medical publications and clinical studies empowers it to put forth suitable treatment recommendations rooted in the most up-to-date scientific proof. Doctors can leverage Chat GPT to weigh different treatment approaches, potential adverse effects, and counter-indications, thus paving the way for customized and well-informed therapeutic choices. [23]

The superior performance grades attributed to Chat GPT-4 align with expectations and are consistent with prior research comparing these two models within the healthcare domain. [19,24-27] This highlights the enhanced Language Model that Chat GPT-4 possesses, having benefitted from a broader and more varied dataset during its development phase. Such improvements allow Chat GPT-4 to detect more complex linguistic patterns, thereby improving its ability to comprehend and generate responses that are contextually relevant. Furthermore, the inclusion of a more substantial corpus of medical literature, scientific articles, and ethical guidelines equips Chat GPT-4 with a wider base of knowledge. This rich repository of information empowers the model to provide more comprehensive and nuanced answers when faced with medical inquiries and ethical dilemmas.

Our study also explored an intriguing research question related to Chat GPT's role in supporting decision-making in the nuanced area of Ethical Dilemmas. Since ethical considerations are a fundamental part of medical practice, they often create intricate scenarios for medical professionals. Chat GPT has the potential to act as a significant tool in this aspect, providing insights into ethical principles and the resolutions of past cases. While the final ethical determinations are the responsibility of the physician, Chat GPT's assistance can guide them through multifaceted ethical conundrums, thereby enhancing the focus on patient-centered care. Hence, our data reveals that questions related to ethics consistently garnered the highest evaluations (as seen in table 4, Figure 2, and Figure 3). These observations emphasize Chat GPT's impressive ability in grappling with ethical dilemmas, regardless of their innate complexity. The model's performance in this vital area is praiseworthy, even outperforming its evaluations in responding to information-based questions.

Interestingly, for both GPT-4 and GPT-3.5, senior physicians consistently rated the responses higher across all criteria. The variation was most notable in the assessment of the models' benefits and completeness. These differences might be explained by the combination of higher experience with complexity, higher familiarity with research and guidelines, higher critical analysis skills, more experience navigating ethical dilemmas, and a broader interdisciplinary perspective possessed by senior physicians that likely contributes to their higher ratings of ChatGPT's responses. However, further research is warranted to confirm our results.

Our study had several limitations. Firstly, the expert panel that formulated the questions was composed of only four experts, while the group that assessed the questions included eight physicians. Although the evaluating group was quite diverse, the findings might not accurately reflect the views of the broader physician community within these specialties. This small, potentially non-representative sample could lead to biases in focus areas, subjective interpretations, and variability in expertise. Subjectivity in assessing responses, especially in relevance and clarity, combined with individual preconceptions about AI's capabilities, might skew results. Expanding the expert pool and incorporating diverse perspectives from different subspecialties and healthcare settings, along with methodologies to adjust for

individual
biases, could mitigate these limitations in future studies.
Secondly, the assessment of ChatGPT's efficacy relied on subjective judgments from two groups of physicians, which could introduce bias and inconsistency. Nonetheless, medical inquiries often encompass intricate matters that defy simple quantification, and the intent of the study was to gauge its applicability to everyday medical work. Hence, personal assessment plays a vital role in this examination, ensuring that ChatGPT's responses are pertinent, lucid, evidence-supported, legitimate, and worthwhile.

Thirdly, this study concentrated solely on ChatGPT's capability in responding to questions within specific sub-disciplines of internal medicine, emergency medicine (ER), and ethics, leaving its potential in other medical fields unexplored. Additional studies are required to scrutinize ChatGPT's performance across a broader spectrum of medical areas. Fourthly, our examination was restricted to both versions of ChatGPT, raising the possibility that the findings might have differed with an alternate language model. Further research is needed to ascertain how applicable our conclusions might be to other Language Model Models (LLMs) and varying contexts.

In conclusion, the potential of Chat GPT in aiding physicians in addressing common medical problems is promising. As technology continues to advance, integrating Chat GPT into clinical workflows may become a valuable asset, enhancing diagnostic accuracy, treatment decisions, and ethical considerations. Nevertheless, it is essential to acknowledge that Chat GPT's role should complement rather than replace clinical expertise and human judgment. As the technology evolves, further research and validation studies are warranted to optimize its abilities, ensuring safe and effective use in clinical settings.

References

[1] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2(4), 230-243.

[2] Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... & Coiera, E. (2018). Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association, 25(9), 1248-1258.

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Sutskever, I. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[5] Gwet, K. L. (2020). Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Advanced Analytics, LLC.

[6] Lahat, A.; Shachar, E.; Avidan, B.; Shatz, Z.; Glicksberg, B.S.; Klang, E. Evaluating the use of large language model in identifying top research questions ingastroenterology. Sci. Rep. 2023, 13, 4164.

[7] Ge, J.; Lai, J.C. Artificial intelligence-based text generators in hepatology: ChatGPT isjust the beginning. Hepatol. Commun. 2023, 7, e0097.

[8] Lahat A, Klang E. Can advanced technologies help address the global increase in demand for specialized medical care and improve telehealth services? J Telemed Telecare. 2023 Feb 9:1357633X231155520. doi: 10.1177/1357633X231155520

[9] Hirosawa, T.; Harada, Y.; Yokose, M.; Sakamoto, T.; Kawamura, R.; Shimizu, T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int. J. Environ. Res. Public Health 2023, 20, 3378. https://doi.org/10.3390/ijerph20043378.

[10] 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, e46885. https://doi.org/10.2196/46885.

[11] Rasmussen, M.L.R.; Larsen, A.C.; Subhi, Y.; Potapenko, I. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. Graefe's Arch. Clin. Exp. Ophthalmol. 2023. May 2. doi: 10.1007/s00417-023-06078-1

[12] Samaan, J.S.; Yeo, Y.H.; Rajeev, N.; Hawley, L.; Abel, S.; Ng, W.H.; Srinivasan, N.; Park, J.; Burch, M.; Watson, R.; et al. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. Obes. Surg. 2023, 33, 1790–1796. https://doi.org/10.1007/s11695-023-06603-5.

[13] Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic Surgery Advice and Counseling from Artificial Intelligence: A Rhinoplasty Consultation with ChatGPT. Aesthetic Plast Surg. 2023 Apr 24. doi: 10.1007/s00266-023-03338-7

[14] Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, Ayoub W, Yang JD, Liran O, Spiegel B, Kuo A. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023 Mar 22. doi: 10.3350/cmh.2023.0089

[15] Johnson, S.B.; King, A.J.; Warner, E.L.; Aneja, S.; Kann, B.H.; Bylund, C.L. Using ChatGPT to evaluate cancer myths and misconceptions: Artificial intelligence and cancer information. JNCI Cancer Spectr. 2023, 7, pkad015. https://doi.org/10.1093/jncics/pkad015.

[16] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

[17] Voss, C., Surendran, A., & LaValle, S. (2023). OpenAI GPT-4: Supercharging AI's language understanding. OpenAI blog.

[18] Valentin Liévin, Christoffer Egeberg Hother, Ole Winther. "Can large language models reason about medical questions?" 2023. arXiv preprint arXiv:2207.08143. https://doi.org/10.48550/arXiv.2207.08143.

[19] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv preprint arXiv:2303.13375. 2023. https://doi.org/10.48550/arXiv.2303.13375.

[20] Chow JCL, Wong V, Sanders L, Li K. Developing an AI-Assisted Educational Chatbot for Radiotherapy Using the IBM Watson Assistant Platform. Healthcare (Basel). 2023 Aug 29;11(17):2417

[21] Rebelo N, Sanders L, Li K, Chow JCL. Learning the Treatment Process in Radiotherapy Using an Artificial Intelligence-Assisted Chatbot: Development Study. JMIR Form Res. 2022 Dec 2;6(12):e39443

[22] Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell. 2023 Apr 5;6:1166014

[23] Chiesa-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M, Sanchez-Barrueco A, Saga-Gutierrez C. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol. 2023 Jul 5. doi: 10.1007/s00405-023-08104-8. Epub ahead of print. PMID: 37405455.

[24] 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 Jun 29;9:e48002. doi: 10.2196/48002. PMID: 37384388; PMCID: PMC10365615.

[25] Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. Cureus. 2023 Jun 22;15(6):e40822. doi: 10.7759/cureus.40822. PMID: 37485215; PMCID: PMC10362981.

[26] He N, Yan Y, Wu Z, Cheng Y, Liu F, Li X, Zhai S. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. J Telemed Telecare. 2023 Jun 22:1357633X231181922. doi: 10.1177/1357633X231181922. Epub ahead of print. PMID: 37350055.

[27] Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on Medical Challenge Problems. *ArXiv*. /abs/2303.13375

Table and figures legends

Tables

Figures

Summary points:

- ✓ Our aim was to examine Chat GPT-3.5 and GPT-4's efficacy in tackling intricate clinical and ethical challenges.
- ✓ Our methods included physicians' assessment of GPT-3.5 and GPT-4's answers to 176 practical clinical

queries.

- ✓ Our results: both versions achieved high grades: Average scores: GPT-4 (4.4 ± 0.8) and GPT-3.5 (4.1 ± 1.0), yet GPT-4 surpassed GPT-3.5 in every evaluation category. Seniors gave higher ratings than residents for both models. Ethical questions had the top scores for both versions.
- ✓ Our conclusions: Chat GPT's role in assisting doctors seems hopeful, especially in diagnostics, treatments, and ethics. Its integration into clinical processes is beneficial but shouldn't eclipse human knowledge. More studies are warned to establish its place in clinical practice.

Supplementary Files

Multimedia Appendixes

Original Questions and Responses Generated by GPT-3.5 and GPT-4.
URL: http://asset.jmir.pub/assets/33930e57099038b64027f96df91f7667.xlsx