ChatGPT-4 performance on USMLE Step 1 questions and its implications for medical education: A comparative study across systems and disciplines

Razmig Garabet (rgarabet@gmail.com)
Drexel University College of Medicine

Brendan P Mackey
Drexel University College of Medicine

James Cross, MBA
Drexel University College of Medicine

Michael Weingarten, MD, MBA
Drexel University College of Medicine

Research Article

Keywords: medical education, ChatGPT, USMLE, Step 1, machine learning, artificial intelligence

Posted Date: August 8th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3240108/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

We assessed the performance of OpenAI’s ChatGPT-4 on United States Medical Licensing Exam STEP 1 questions across the systems and disciplines appearing on the examination. ChatGPT-4 answered 86% of the 1300 questions accurately, exceeding the estimated passing score of 60% with no significant differences in performance across clinical domains. Findings demonstrated an improvement over earlier models as well as consistent performance in topics ranging from complex biological processes to ethical considerations in patient care. Its proficiency provides support for the use of artificial intelligence (AI) as an interactive learning tool and furthermore raises questions about how the technology can be used to educate students in the preclinical component of their medical education. Authors provide an example and discuss how students can leverage AI to receive real-time analogies and explanations tailored to their desired level of education. An appropriate application of this technology potentially enables enhancement of learning outcomes for medical students in the preclinical component of their education.

Introduction

Artificial intelligence (AI) is revolutionizing our world. As an emerging technology, it has shown enormous potential in numerous fields with the arrival of self-driving vehicles, smart devices, and automated personal assistants [1]. This has since sparked conversation about other possible implications, particularly in the realm of education.

The basis for AI’s use in education lies in its incorporation into language learning models (LLMs), or computational tools that can comprehend, learn, and generate human-like text. These models train on an extensive data set and abide by an algorithm to discern relationships between words and grammatical structure in order to establish language patterns [2]. As a result, LMMs establish rules for which they can create their own coherent and contextually relevant sentences.

A prime example is OpenAI’s ChatGPT, which in its fourth iteration builds upon a transformer-based model and introduces several pivotal improvements over prior versions and other models [3]. The most notable is its breadth of training data, which expanded from 175 billion parameters in GPT3.5 to 1 trillion in the 4.0 model [4]. Bing’s current AI model has 175 billion parameters and Microsoft’s Bard has 540 billion by comparison [5, 6]. ChatGPT therefore has a substantial knowledge base, which enables it to produce more nuanced and accurate responses than its competitors.

Due to its widespread acclaim as the fastest growing model, GPT-4 has since been subjected to numerous testing environments, passing the Uniform Bar Exam, Law School Admission Test, Scholastic Aptitude Tests, Graduate Record Examinations, and Advanced Placements exams with high percentiles [7, 8]. In this paper, we focus on GPT-4 performance with respect to the United States Medical Licensing Examination (USMLE) STEP 1 examination, which is the first of three board exams required of medical students. It is administered over 8 hours and consists of 7 blocks each composed of 40 questions.
pertaining to basic sciences of the practice of medicine [9]. In 2022, 29,039 examinees from US/Canadian schools sat the exam and 91% passed [10].

Earlier studies have demonstrated success of the prior models, ChatGPT-3 and ChatGPT-3.5, with scores above the passing threshold on USMLE STEP 1 questions, provoking discussion of its use as a question analyzer and educational resource [11, 12]. All of these studies, however, reported GPT’s overall performance. No studies have assessed performance within the subjects and disciplines present on the USMLE STEP 1 exam. This information is needed to assess potential content weaknesses prior to recommendation as a learning tool.

In this study, we expand upon the initial findings with GPT-3 and GPT-3.5 by using the newer model GPT-4 and reporting performance by subject and discipline. Primary objectives included determination of strict AI performance in the question sets. Secondary objectives involved identification of potential implications for GPT-4 in medical education.

Materials and Methods

Sample Question Acquisition

Authors received permission to source questions from AMBOSS with an active subscription [13]. AMBOSS is a comprehensive learning platform that offers a question bank for the USMLE STEP 1 examination [13]. In its custom session interface, users can select questions from a series of systems and disciplines to hone their studies to particular topics. For our purposes, we used this feature to extract questions from 18 systems and 12 disciplines.

Analogous systems and disciplines were merged and reclassified as demonstrated in Table 1. For example, “Reproductive System” contains questions drawn from “Female Reproductive System & Breast”, “Male Reproductive System”, and “Pregnancy, Childbirth & the Puerperium”. In the same manner, “Pharmacology” incorporates questions from both “Clinical Pharmacology” and “Pharmacology and Toxicology”. “Biochemistry” and “Nutrition” were also paired.

For each system and discipline, we selected 50 questions at random from the AMBOSS question bank. Systems and disciplines with less than 50 questions were omitted from our selection. Each set of questions were subjected to independent analysis, therefore we did not remove questions appearing in multiple systems and disciplines but rather employed repeat questions when applicable to the system/discipline assessed. Including these repeats, our analysis encompassed a total of n=1300 questions.
Table 1
Selected Systems and Disciplines.

<table>
<thead>
<tr>
<th>SYSTEMS</th>
<th>DISCIPLINES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral Health (BH)</td>
<td>Anatomy (AN)</td>
</tr>
<tr>
<td>Biostats, Epidemiology/Population Health &amp; Interpretation of the Medical Literature (BP)</td>
<td>Biochemistry, Nutrition (BN)</td>
</tr>
<tr>
<td>Blood &amp; Lymphoreticular System (BL)</td>
<td>Epidemiology, Biostatistics, and Medical Informatics (EB)</td>
</tr>
<tr>
<td>Cardiovascular System (CV)</td>
<td>Histology (HS)</td>
</tr>
<tr>
<td>Endocrine System (ES)</td>
<td>Microbiology &amp; Virology (MV)</td>
</tr>
<tr>
<td>Gastrointestinal System (GI)</td>
<td>Molecular &amp; Cell Biology (MC)</td>
</tr>
<tr>
<td>General Principles of Foundational Science (GP)</td>
<td>Pathology (PA)</td>
</tr>
<tr>
<td>Immune System (IS)</td>
<td>Pharmacology (PC)</td>
</tr>
<tr>
<td>Multisystem Processes &amp; Disorders (MS)</td>
<td>Physiology (PS)</td>
</tr>
<tr>
<td>Musculoskeletal System (MK)</td>
<td>Prevention, Health Promotion (PH)</td>
</tr>
<tr>
<td>Nervous System &amp; Special Senses (NS)</td>
<td></td>
</tr>
<tr>
<td>Renal &amp; Urinary Systems (RU)</td>
<td></td>
</tr>
<tr>
<td>Reproductive System (RP)</td>
<td></td>
</tr>
<tr>
<td>Respiratory System (RS)</td>
<td></td>
</tr>
<tr>
<td>Skin &amp; Subcutaneous Tissue (ST)</td>
<td></td>
</tr>
<tr>
<td>Social Sciences Literature (SS)</td>
<td></td>
</tr>
</tbody>
</table>

**Conversation Input**

Authors entered each question and answer choice into the ChatGPT command as an independent conversation. Following each input and output set, the conversation was deleted and a new conversation initiated to prevent influence from feedback knowledge in prior entries. Questions that contained media, such as histologic sections, gross anatomical specimens, genetic pedigrees, and biochemical pathways were excluded due to the inability of ChatGPT to process image files. Questions with tables that contained text and symbols, however, were included as demonstrated by the sample input and output shown in Fig. 1.

**Data Collection**

Answers for each question input were marked as either correct or incorrect following comparison to feedback after selecting the answer option within the AMBOSS question bank. In the event an output was
inconclusive, the question was voided and excluded. As AMBOSS provides student performance metrics for each question, the percent of students who chose the correct answer was documented as well for performance comparison on each question.

**Statistical Analysis**

Statistical analysis was conducted with IBM SPSS 28. Accuracy percentages for each domain and the cumulative data set were determined. Unpaired chi-squared tests were utilized to compare accuracies of each domain with each other for all systems and disciplines. Student accuracy for the total questions were divided into four quartiles. Unpaired chi-squared tests were used to determine the difference in ChatGPT accuracy when compared to student accuracy.

**Results**

A total of 1,300 questions were analyzed for the study. Of that sample, 800 were analyzed in regard to systems (Fig. 2) and 500 in regard to disciplines (Fig. 3). ChatGPT answered 86% of the total questions accurately. Table 2 lists the obtained accuracies across all domains. There were no significant differences between the various systems \([x^2(15) = 20.38, p = .158]\) or disciplines \([x^2(9) = 11.17, p = .265]\).

Figure 4 depicts the accuracy for ChatGPT when compared to different quartiles of student accuracy. Each quartile demonstrated a significant difference in ChatGPT accuracy when compared to each other \([x^2(4) = 108.56, p < .001]\). Specifically, ChatGPT accuracy was significantly higher when comparing the higher student accuracy quartiles to the lower quartiles and similarly for each comparison thereafter \(p < .001\).
### Table 2
Accuracy comparisons between ChatGPT and student output based on each system and discipline domain

<table>
<thead>
<tr>
<th>System</th>
<th>AI Accuracy</th>
<th>Student Accuracy</th>
<th>Discipline</th>
<th>AI Accuracy</th>
<th>Student Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral Health</td>
<td>88%</td>
<td>70%</td>
<td>Anatomy</td>
<td>82%</td>
<td>59%</td>
</tr>
<tr>
<td>Biostats, Epidemiology/Population Health &amp; Interpretation of Medical Literature</td>
<td>80%</td>
<td>62%</td>
<td>Biochemistry, Nutrition</td>
<td>80%</td>
<td>58%</td>
</tr>
<tr>
<td>Blood &amp; Lymphoreticular System</td>
<td>84%</td>
<td>60%</td>
<td>Epidemiology, Biostatistics, and Medical Informatics</td>
<td>80%</td>
<td>61%</td>
</tr>
<tr>
<td>Cardiovascular System</td>
<td>90%</td>
<td>64%</td>
<td>Histology</td>
<td>90%</td>
<td>57%</td>
</tr>
<tr>
<td>Endocrine System</td>
<td>82%</td>
<td>61%</td>
<td>Microbiology &amp; Virology</td>
<td>90%</td>
<td>55%</td>
</tr>
<tr>
<td>Gastrointestinal System</td>
<td>82%</td>
<td>60%</td>
<td>Molecular &amp; Cell Biology</td>
<td>84%</td>
<td>52%</td>
</tr>
<tr>
<td>General Principles of Foundational Science</td>
<td>80%</td>
<td>50%</td>
<td>Pathology</td>
<td>84%</td>
<td>59%</td>
</tr>
<tr>
<td>Immune System</td>
<td>92%</td>
<td>56%</td>
<td>Pharmacology</td>
<td>92%</td>
<td>57%</td>
</tr>
<tr>
<td>Multisystem Processes &amp; Disorders</td>
<td>84%</td>
<td>54%</td>
<td>Physiology</td>
<td>86%</td>
<td>56%</td>
</tr>
<tr>
<td>Musculoskeletal System</td>
<td>94%</td>
<td>61%</td>
<td>Prevention, Health Promotion</td>
<td>96%</td>
<td>73%</td>
</tr>
<tr>
<td>Nervous System &amp; Special Senses</td>
<td>78%</td>
<td>59%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Renal &amp; Urinary Systems</td>
<td>82%</td>
<td>60%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reproductive System</td>
<td>96%</td>
<td>58%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory System</td>
<td>90%</td>
<td>59%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skin &amp; Subcutaneous Tissue</td>
<td>90%</td>
<td>59%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Sciences</td>
<td>92%</td>
<td>83%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Discussion

### AI Performance
We present the first analysis of ChatGPT-4 performance in different subjects and disciplines appearing on
the USMLE STEP 1 examination. The findings of our study validate the proficiency of ChatGPT-4, which
posted 86% accuracy in the 1300 questions assessed. This is a significant improvement from previous
studies that used the earlier ChatGPT-3 model on a similar question set, reporting accuracy of 44% [11].
This improvement corroborates the newer model’s demonstration of success in various other
standardized exams [8]. As per the NBME, students must achieve an approximate 60% to pass the
USMLE STEP 1 exam [15]. ChatGPT-4 exceeds that threshold by 26 percentage points and therefore
exhibits the competency level of a student who has completed at least two years of medical school.

In our analysis, we found no significant difference in ChatGPT-4 accuracy across the 18 systems and 12
disciplines shown. These findings demonstrate that ChatGPT-4 is proficient in various preclinical
domains. Interestingly, the AI had no deficits in solving questions that required ethical considerations or
empathetic responses to patients. This was endorsed by a non-significant difference between
performance in the “Social Sciences” system and other content areas consisting of questions that
required pure factual recall and multi-step processing of physiologic processes. These findings are
compatible with other studies that have revered ChatGPT’s capabilities in empathetic communication
skills with patients and demonstrated its proficiency in the area [16].

In addition to assessments of pure AI performance, we compared answer output accuracy to medical
students answering the same questions. As student accuracy increased or decreased, we found ChatGPT
accuracy to trend in the same direction for subjects and disciplines. This could indicate similar logical
reasoning between AI and medical students or otherwise reflect variance in question difficulty between
systems and disciplines.

**Implications in Medical Education**

A recent topic of discussion is the potential implications for ChatGPT-4 in medical education [17, 18]. Its
high accuracy suggests that ChatGPT-4 could be used as a learning tool for medical students, who can
engage in dialogue with the model to test their knowledge and understanding of various medical subjects
and disciplines. As we have shown in our sample input and output, ChatGPT-4 demonstrates multi-step
processing and provides detailed, logical explanations to questions that can help students to understand
complex concepts. Interestingly, however, we found that the output can be leveraged further to provide
explanations at a desired level of education. For instance, ChatGPT-4 can be prompted to provide
simplified explanations or analogies while maintaining accurate reasoning to enhance learning. This is
demonstrated below with follow-up input to our sample question (Fig. 5).

The response provided demonstrates that GPT-4 can serve useful in providing alternative responses when
further understanding is needed. This is similar to the role of a tutor and can be beneficial when
explanations in a textbook or other learning resource are insufficient to the student. In essence, we show
that GPT-4 can be used to provide interactive feedback and present content at the level requested by the
student, which we believe could help to bridge knowledge gaps and aid students struggling with complex
concepts.
As a caveat, there is some concern that an increased dependence on AI could lead to less human interaction and decreased problem-solving skills. As with any learning tool it would be critical to balance AI-based learning with traditional methods to ensure a comprehensive education. In its current stage of development AI also has some limitations, such as lack of image support. For that reason, students who use the current version of AI are unable to use this resource to aid with interpretation of ECGs, histology, or radiographic images. Although this feature is currently non-existent we anticipate it will become available in the future, which will warrant further research.

Limitations of our study include possible bias in question selection for fields with a high proportion of image-based questions, as these were excluded. Furthermore, we were unable to assess more specific disciplines included within the AMBOSS question set due to a smaller sample size of questions within those domains. Application of ChatGPT-4 in its current state, however, offers promise as a supplemental resource for preclinical medical education. Future studies should aim to assess the practicality and efficacy of ChatGPT in supplementation with the current preclinical medical curriculum. Assessment of its accuracy on other medical school and specialty board exams may also be beneficial to understand the maximum potential of the software in its current stage of development.

Conclusion

The capability of GPT-4 to sort through large volumes of data in seconds allows it to provide immediate, interactive feedback, something that is invaluable to students preparing for high-stakes examinations. In preparation for exams such as USMLE STEP 1, ChatGPT-4 demonstrates proficiency in all subjects and disciplines assessed and could potentially serve a role in fundamentally changing our approach to preclinical education. In leveraging this technology, we stand on the brink of significant breakthroughs in our ability to enhance learning outcomes, especially in fields as critical and demanding as medicine.

Declarations

Data Availability

The dataset obtained during and analyzed during the current study is available from the corresponding author on reasonable request.

Acknowledgements

The authors wish to acknowledge AMBOSS for allowing our use of their comprehensive question set, without which this research would not have been possible.

Funding

No funding was received for conducting this study.

Conflict of Interest
The authors have no relevant financial or non-financial interests to disclose.

**Author Information**

All authors contributed to the study conception and design. Material preparation and data collection were performed by RG, BM. JC. Analyses were performed by RG. The first draft of the introduction was written by JC, methods and results by RG, discussion and conclusion by BM. All authors commented on previous versions of the manuscript. Revisions were completed by BM. All authors read and approved the final manuscript. Correspondence to RG.

**Statements and Declarations**

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

**References**

3. GPT-4 is OpenAI’s most advanced system, producing safer and more useful responses. OpenAI. https://openai.com/gpt-4


13. The smarter way to learn and practice medicine. AMBOSS. https://www.amboss.com


15. United States Medical Licensing Examination, Examination Results and Scoring. https://www.usmle.org/bulletin-information/scoring-and-score-reporting


Figures
Sample Input:
A healthy 30-year-old woman participates in a study of the relationship between cardiovascular function and posture. Placed in a room with a comfortable temperature (72°F) and humidity (40%), she moves from a standing to a supine position. Which of the following sets of changes is most likely in the reflex control of this woman’s cardiovascular system?

<table>
<thead>
<tr>
<th>Carotid Sinus Baroreceptor Activity</th>
<th>Atrial Stretch Receptor Activity</th>
<th>Vagal Efferent Activity to the Heart</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. ↑</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>B. ↑</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td>C. ↑</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>D. ↑</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>E. ↓</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>F. ↓</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td>G. ↓</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>H. ↓</td>
<td>↓</td>
<td>↓</td>
</tr>
</tbody>
</table>

Sample Output:
Transitioning from standing to a supine position reduces the gravitational stress on the body’s circulatory system, especially in relation to the work of the heart in pumping blood upwards against gravity. This leads to certain changes in the reflex control of the cardiovascular system.

1. **Carotid Sinus Baroreceptor Activity**: The baroreceptors, which are located in the carotid sinus and the aortic arch, sense changes in blood pressure. When moving from standing to supine, blood that was previously pooled in the legs due to gravity is now more easily circulated, causing an increase in venous return to the heart. This results in increased cardiac output and blood pressure. Baroreceptors are stretch-sensitive and respond to this increase in pressure by increasing their activity - so we can expect an increase (↑) in carotid sinus baroreceptor activity.

2. **Atrial Stretch Receptor Activity**: The increase in venous return to the heart also results in the atria of the heart being stretched more than usual. Atrial stretch receptors would respond to this increased stretch by increasing their activity - so again, we can expect an increase (↑) in atrial stretch receptor activity.

3. **Vagal Efferent Activity to the Heart**: The increased activity of the baroreceptors and atrial stretch receptors would lead to a reflex response aimed at maintaining homeostasis, i.e., keeping the blood pressure in a normal range. This is achieved by increasing parasympathetic (vagal) activity to the heart, which leads to a reduction in heart rate and contractility. In other words, the increased pressure sensed by the baroreceptors results in an increase (↑) in vagal efferent activity to the heart to reduce the pressure back to normal.

So, the correct answer would be A. ↑↑↑: an increase in carotid sinus baroreceptor activity, an increase in atrial stretch receptor activity, and an increase in vagal efferent activity to the heart.

**Figure 1**
Sample input and output for demonstration of conversation with ChatGPT. Sample question obtained from USMLE Step 1 Orientation [14]
Figure 2

ChatGPT accuracy based on the various system domains analyzed
Figure 3

ChatGPT accuracy based on the various discipline domains analyzed
Figure 4

ChatGPT accuracy compared to quartiles of student correct percentages
Sample Input:
Explain this concept with an analogy.

Sample Output:
Sure, here's an analogy that might help:

Imagine a water pump (the heart) in a garden, with a hose (the blood vessels) leading up to a hanging flower pot (the brain). The Carotid Sinus Baroreceptor is like a pressure gauge on the hose near the pot, measuring the pressure of the water (blood pressure). The Atrial Stretch Receptor is like a volume meter in the water pump, measuring how much water (blood) it's pumping. The Vagal Efferent Activity is like the speed control on the water pump, it can slow down or speed up the pump.

When the flower pot is on the ground (standing), the pump needs to work harder to push the water up. But, when you hang the flower pot at the same level as the pump (lying down), the pump doesn't have to work as hard because it doesn't have to pump water against gravity.

The pressure gauge (Carotid Sinus Baroreceptor) will read higher because the water pressure has increased. The volume meter (Atrial Stretch Receptor) will also read higher because more water is getting back to the pump. To avoid flooding the garden, you turn down the speed of the pump (increase Vagal Efferent Activity) to balance the higher pressure and volume.

So, when you move from standing to lying down, both the pressure gauge and the volume meter would read higher (↑), and you'd also turn down the speed of the pump (↓). So, the answer is A.

Figure 5

Sample follow-up input and output for demonstration of conversation with ChatGPT