

ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination

Andrew Mihalache, Ryan S. Huang, Marko M. Popovic & Rajeev H. Muni

To cite this article: Andrew Mihalache, Ryan S. Huang, Marko M. Popovic & Rajeev H. Muni (15 Oct 2023): ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination, Medical Teacher, DOI: [10.1080/0142159X.2023.2249588](https://doi.org/10.1080/0142159X.2023.2249588)

To link to this article: <https://doi.org/10.1080/0142159X.2023.2249588>



Published online: 15 Oct 2023.



Submit your article to this journal [↗](#)



Article views: 189



View related articles [↗](#)



View Crossmark data [↗](#)

ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination

Andrew Mihalache^a, Ryan S. Huang^a, Marko M. Popovic^b and Rajeev H. Muni^{b,c}

^aTemerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada; ^bDepartment of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Ontario, Canada; ^cDepartment of Ophthalmology, St. Michael's Hospital/Unity Health Toronto, Toronto, Ontario, Canada

ABSTRACT

Purpose: ChatGPT-4 is an upgraded version of an artificial intelligence chatbot. The performance of ChatGPT-4 on the United States Medical Licensing Examination (USMLE) has not been independently characterized. We aimed to assess the performance of ChatGPT-4 at responding to USMLE Step 1, Step 2CK, and Step 3 practice questions.

Method: Practice multiple-choice questions for the USMLE Step 1, Step 2CK, and Step 3 were compiled. Of 376 available questions, 319 (85%) were analyzed by ChatGPT-4 on March 21st, 2023. Our primary outcome was the performance of ChatGPT-4 for the practice USMLE Step 1, Step 2CK, and Step 3 examinations, measured as the proportion of multiple-choice questions answered correctly. Our secondary outcomes were the mean length of questions and responses provided by ChatGPT-4.

Results: ChatGPT-4 responded to 319 text-based multiple-choice questions from USMLE practice test material. ChatGPT-4 answered 82 of 93 (88%) questions correctly on USMLE Step 1, 91 of 106 (86%) on Step 2CK, and 108 of 120 (90%) on Step 3. ChatGPT-4 provided explanations for all questions. ChatGPT-4 spent 30.8 ± 11.8 s on average responding to practice questions for USMLE Step 1, 23.0 ± 9.4 s per question for Step 2CK, and 23.1 ± 8.3 s per question for Step 3. The mean length of practice USMLE multiple-choice questions that were answered correctly and incorrectly by ChatGPT-4 was similar (difference = 17.48 characters, SE = 59.75, 95%CI = [-100.09, 135.04], $t = 0.29$, $p = 0.77$). The mean length of ChatGPT-4's correct responses to practice questions was significantly shorter than the mean length of incorrect responses (difference = 79.58 characters, SE = 35.42, 95%CI = [9.89, 149.28], $t = 2.25$, $p = 0.03$).

Conclusions: ChatGPT-4 answered a remarkably high proportion of practice questions correctly for USMLE examinations. ChatGPT-4 performed substantially better at USMLE practice questions than previous models of the same AI chatbot.

KEYWORDS

artificial intelligence; natural language processing; United States medical licensing examination; chatgpt-4

Introduction

ChatGPT-4 is an artificial intelligence (AI) chatbot recently developed by OpenAI. ChatGPT-4 and similar systems have the potential to lead to great infrastructural changes in science, similar to those seen with the release of the internet (Sanderson 2023). This AI chatbot recognizes virtually any context and conversates naturally with users, making it the fastest-growing consumer application to-date, with over 100 million users (Milmo 2023). The new model of ChatGPT-4 can process over 25,000 words per prompt, allowing for substantially more comprehensive conversations compared to ChatGPT-3 (GPT-4). Nevertheless, there are major concerns surrounding the potential for this AI chatbot to be inappropriately used in education and research (Milmo 2023; O'Connor and ChatGPT 2023). There are also concerns that ChatGPT is capable of posing a threat to the integrity of conventional assessments, including online examinations and assignments (Stokel-Walker 2022; Susnjak 2022). Though ChatGPT-4 has been introduced for general use very recently, its predecessor, ChatGPT-3, has already been used in medical writing (Biswas 2023), simplifying reports (Jeblick et al. 2022),

Practice points

- ChatGPT-4 answered 88% questions correctly on USMLE Step 1, 86% on Step 2CK, and 90% on Step 3.
- ChatGPT-4 answered a remarkably high proportion of practice questions correctly for USMLE examinations.
- ChatGPT-4's medical knowledge corpus is advancing rapidly and it performed substantially better at USMLE practice questions than previous models of the same AI chatbot.

providing decision-making for medications (ChatGPT Generative Pre-trained Transformer and Zhavoronkov 2022), conducting literature reviews (Aydin and Karaarslan 2022), as well as writing abstracts and editorials (Gao et al. 2022; Else 2023; Cahan and Treutlein 2023).

The United States Medical Licensing Examination (USMLE) step examinations are needed for medical licensure in the United States and have been used in the selection of trainees into residency programs (Cohen et al.

2020). A recent study found that the previous model of ChatGPT was able to perform at or near the threshold for passing the USMLE Step 1, Step 2CK, and Step 3 examinations without any training (Kung et al. 2023). Another preprint found that ChatGPT-3 answered over 60% sample questions correctly pertaining to subjects on the USMLE Step 1, surpassing the threshold considered for passing the examination (Gilson et al. 2022). In an internal evaluation by Open AI, the upgraded model of this AI chatbot, ChatGPT-4, was found to obtain a 75th percentile score on the Medical Knowledge Self-Assessment Program, marking a great improvement from ChatGPT-3, which scored in the 53rd percentile. (OpenAI 2023) ChatGPT-4 has also improved remarkably relative to ChatGPT-3 at the United States Biology Olympiad, achieving scores in the 99-100th percentile compared to the 31-33rd percentile with its previous model (OpenAI 2023). The exponential growth of this AI chatbot's performance extends to other disciplines as well, as ChatGPT-4 has been shown to outperform average test-takers on the sample questions from the Bar Examination (Katz et al. 2023).

There is currently a growing emphasis on the integration of AI training curricula for medical education (Hu et al. 2022). Students must be able to recognize both the benefits and limitations of AI chatbots such as ChatGPT, which are capable of producing inaccurate outputs (Azaria). Nevertheless, ChatGPT-4 may provide medical students with a unique platform to improve their clinical skills and medical knowledge with its ability to facilitate the brainstorming of ideas and simulate the dialogue of a patient (Gilson et al. 2022). Currently, the performance of the new release of ChatGPT-4 has not been characterized for the various USMLE Step examinations. In this study, we aim to assess the performance of ChatGPT-4 to answer practice questions for the USMLE Step 1, Step 2CK, and Step 3 examinations.

Methods

We used practice multiple-choice questions for the United States Medical Licensing Examination (USMLE) Step 1, Step 2CK, and Step 3 available from the Federation of State Medical Boards (FSMB) and National Board of Medical Examiners. The USMLE Step 1 booklet was last updated in June 2022 (USMLE Step 1 Sample Test Questions 2022), the Step 2CK booklet in March 2023 (Step 2 Clinical Knowledge (CK) Sample Test Questions 2023), and the Step 3 booklet in August 2022 (Step 3 Sample Questions August 2022 2022). We collected the following data: the question number from each USMLE Step examination practice question, the body of each question and its multiple-choice options exactly as they appeared in the practice question booklet, the length of questions in characters, the response generated by ChatGPT-4 to each question, the length of ChatGPT-4's responses, whether ChatGPT-4's response was correct or incorrect, whether ChatGPT-4's response provided additional insight in its response, the duration of ChatGPT-4's response in seconds, and whether ChatGPT-4's servers experienced high demand while responding to questions. If the AI chatbot selected none or all the above as its response to a multiple-choice question, this response was deemed incorrect if there were no such options in the

practice booklets. All of ChatGPT-4's responses were reviewed by two independent authors (A.M., R.S.H.) to determine the multiple-choice option that was selected. Our primary outcome was the performance of ChatGPT-4, measured as the proportion of multiple-choice questions answered correctly for official practice USMLE Step 1, Step 2CK, and Step 3 examinations. Our secondary outcomes were the mean length of questions and ChatGPT-4's responses in characters for questions that ChatGPT-4 answered correctly and incorrectly.

Given that ChatGPT-4 only accepts text inputs, we excluded all questions containing embedded figures or formatted tables from our analysis. No hints or tips were provided to ChatGPT-4 to simulate real examination conditions. A new ChatGPT Plus account was created with the initiation of our study to ensure no prior conversation history with the AI chatbot. Before each multiple-choice question was inputted into ChatGPT-4, the previous conversation history with the AI chatbot was cleared to ensure that its responses were not influenced by concurrent conversations. ChatGPT-4 was also refreshed prior to inputting every question to determine if OpenAI's servers were experiencing self-reported high demand. All multiple-choice questions were asked on March 21st, 2023, to gauge ChatGPT-4's performance at a single point in time. Each question was asked on two separate internet browsers (Mozilla Firefox and Google Chrome) to ensure ChatGPT-4 provided consistent responses, irrespective of internet browser.

Statistical analysis was performed on MedCalc using a chi-squared test (χ^2) to compare the proportion of questions answered correctly by ChatGPT-4 across practice USMLE Step 1, Step 2CK, and Step 3 examinations per methods outlined by Campbell (2007) and Richardson (2011) (Altman et al. 2000; Campbell 2007; Richardson 2011; Comparison of proportions calculator 2023). A t-test was used to compute the difference between observed means in the length of practice USMLE multiple-choice questions answered correctly and incorrectly by ChatGPT-4 (Altman 1990; Kirkwood and Sterne 2003; MedCalc 2023). We performed a one-way analysis of variance (ANOVA) with a Tukey honestly significant difference (HSD) post-hoc test to determine if the mean length of questions, mean length of ChatGPT-4's responses and mean duration of ChatGPT-4's responses varied across practice questions for USMLE Step 1, Step 2CK, and Step 3 examinations (Statistics). P-values were all two-tailed and a p-value less than 0.05 was considered statistically significant. We conducted our study using the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines. Research ethics approval was not required as our study used publicly accessible practice questions and did not involve research participants.

Results

ChatGPT-4 was used to respond to 319 text-based multiple-choice questions from official practice USMLE Step 1, Step 2CK, and Step 3 booklets on March 21st, 2023. Out of 119 multiple-choice questions in the USMLE Step 1 practice booklet, 93 (78%) questions did not include figures or formatted tables and were included (Table 1). Similarly, 106 of

Table 1. Characteristics of questions from USMLE practice examinations and ChatGPT-4's responses on March 21, 2023. These Practice Items are owned, copyrighted by, and published with permission from the Federation of State Medical Boards and the National Board of Medical Examiners.

USMLE Examination	No. Available Questions	No. Text-Based Questions (%)	Question Length (Characters)	Response Length (Characters)	Response Time (seconds)	No. Questions Correct (%)
Step 1	119	93 (78%)	775.7 ± 269.8	584.7 ± 196.8	30.8 ± 11.8	82 (88%)
Step 2CK	120	106 (88%)	1087.1 ± 308.1	1087.1 ± 308.1	23.0 ± 9.4	91 (86%)
Step 3	137	120 (88%)	1204.4 ± 307.6	571.9 ± 198.0	23.1 ± 8.3	108 (90%)
Overall	376	319 (85%)	1040.4 ± 345.2	746.8 ± 339.2	25.3 ± 10.4	281 (88%)

120 (88%) practice USMLE Step 2CK and 120 of 137 (88%) Step 3 questions were included. The mean length of practice USMLE questions was 775.7 ± 269.8 characters for Step 1, 1087.1 ± 308.1 characters for Step 2CK, and 1204.4 ± 307.6 characters for Step 3. The mean length of practice USMLE questions was significantly shorter for Step 1 questions compared to Step 2CK and Step 3 questions ($F = 56.45$, $p < 0.00001$). The mean length of ChatGPT-4's responses to practice questions was 584.7 ± 196.8 characters for Step 1, 1087.1 ± 308.1 characters for Step 2CK, and 571.9 ± 198.0 characters for Step 3. The mean length of ChatGPT-4's responses to practice USMLE questions was significantly greater for Step 2CK compared to Step 1 and Step 3 questions ($F = 159.71$, $p < 0.00001$). The mean time taken for ChatGPT-4 to respond to practice USMLE questions was 30.8 ± 11.8 s for Step 1, 23.0 ± 9.4 s for Step 2CK, and 23.1 ± 8.3 s for Step 3 questions. ChatGPT-4's responses to practice USMLE Step 1 questions took a significantly longer time to compute compared to Step 2CK and Step 3 questions ($F = 21.01$, $p < 0.00001$). OpenAI did not report that ChatGPT-4 was experiencing high demand during its response to any multiple-choice question.

Across all practice booklets for the USMLE Step 1, Step 2CK, and Step 3 examinations, ChatGPT-4 answered 281 (88%) of multiple-choice questions correctly. ChatGPT-4 performed best on the practice USMLE Step 3 examination, responding to 108 (90%) multiple-choice questions correctly. ChatGPT-4 performed second best on the practice Step 1 examination, responding to 82 (88%) questions correctly, and poorest on the practice Step 2CK examination, responding to 91 (86%) questions correctly. The proportion of multiple-choice questions that ChatGPT-4 answered correctly was similar between the Step 1 and Step 2CK (difference = 2.32%, 95%CI = [-7.42%, 11.71%], $\chi^2 = 0.23$, $p = 0.63$), Step 1 and Step 3 (difference = 1.83%, 95%CI = [-6.56%, 10.97%], $\chi^2 = 0.18$, $p = 0.67$), as well as Step 2CK and Step 3 practice sets (difference = 4.15%, 95%CI = [-4.42%, 13.08%], $\chi^2 = 0.92$, $p = 0.34$). ChatGPT-4 provided explanations for all multiple-choice questions, and it was noted that the same multiple-choice response to all practice questions was provided when the study methods were performed on Google Chrome compared to Mozilla Firefox. A summary of questions answered correctly and incorrectly by ChatGPT-4 can be found in Table 2.

The mean length of practice questions that were answered correctly versus incorrectly by ChatGPT-4 was similar (difference = 17.48 characters, SE = 59.75, 95%CI = [-100.09, 135.04], $t = 0.29$, $p = 0.77$). The mean length of ChatGPT-4's correct responses to practice questions was significantly shorter than the mean length of ChatGPT-4's incorrect responses (difference = 79.58 characters, SE = 35.42, 95%CI = [9.89, 149.28], $t = 2.25$, $p = 0.03$).

Discussion

ChatGPT-4 is a highly powerful natural language processing tool that may have the capacity to alter practices in education and medical sciences (Zhai 2022). This novel AI chatbot has brought conversational AI systems into the public domain. ChatGPT-4 provides efficient responses and requires minimal professional knowledge to be used for learning (Zhai 2022). The applicability of ChatGPT-4 to medical licensing examination questions was very high. Indeed, our study found that ChatGPT-4 responded to 88% of practice questions correctly for the USMLE Step 1, 86% for Step 2CK, and 90% for Step 3. If one were to allocate an equal amount of time per question, the USMLE Step 1, Step 2CK, and Step 3 examination allow for approximately 90 s per question (Step 1 Exam Content | USMLE; Step 2 CK Exam Content | USMLE; Step 3 Exam Content | USMLE). Nevertheless, ChatGPT-4 as used in our study spent 30.8 ± 11.8 s on average responding to practice questions for USMLE Step 1, 23.0 ± 9.4 s per question for Step 2CK, and 23.1 ± 8.3 s per question for Step 3.

The performance of natural language processing platforms like ChatGPT-4 continues to grow exponentially and rapidly. An early model of this AI chatbot achieved a 46% accuracy on USMLE sample questions in 2022 (Liévin et al. 2022). A recent study in February 2023 found that the previous model of ChatGPT achieved scores ranging from 36.1% to 61.3% across practice multiple-choice USMLE examinations (Kung et al. 2022). Our study, the first on ChatGPT-4 in this context, found that the AI chatbot's performance improved remarkably relative to its previous models. Another preprint in December 2022 found that the correct responses from ChatGPT-3 to practice questions for USMLE Step 1 and Step 2 examinations contained a significantly greater amount of external information compared to incorrect responses (Gilson et al. 2022). Hence, Gilson et al. (2022) suggested that the ability of ChatGPT-3 to answer a question correctly may depend on its ability to prompt data within its knowledge corpus (Gilson et al. 2022). Our current study found that ChatGPT-4's responses provided explanations to all practice multiple-choice questions for USMLE Step examinations, irrespective of whether the response was correct or incorrect. Contrary to the findings by Gilson et al. we found that the mean length of ChatGPT-4's correct responses was significantly shorter than the mean length of its incorrect responses.

The performance of the AI chatbot ChatGPT has also been remarkably high in specialty-specific examinations. In previous investigations, we prompted the chatbot with 125 practice questions for ophthalmology board certification, where ChatGPT-3.5 answered 46% correctly in January 2023 and ChatGPT-4 answered 84% correctly in March 2023 (Mihalache, Popovic, et al. 2023; Mihalache, Huang, et al.

Table 2. Sample multiple-choice questions and ChatGPT-4's output on March 21, 2023.

USMLE Examination	Outcome	Question	ChatGPT-4's Response
Step 1	Correct	<p>A previously healthy 33-year-old woman is brought to the emergency department by the Secret Service for stalking the president of the USA for 2 months. She claims to be married to the president's twin brother and states that the president just had his twin kidnapped to avoid competition. She speaks rapidly and is difficult to interrupt. Her associations are often loose. She says, 'I haven't slept for days, but I won't even try to sleep until my husband is rescued. God has been instructing me to take over the White House. I can't wait to be reunited with my husband. I hear his voice telling me what to do'. When asked about drug use, she says she uses only natural substances. She refuses to permit blood or urine tests, saying, 'I don't have time to wait for the results'. Which of the following is the most likely diagnosis?</p> <p>(A) Bipolar disorder, manic, with psychotic features (B) Brief psychotic disorder (C) Delusional disorder (D) Psychotic disorder due to general medical condition (E) Schizophrenia</p>	<p>The patient exhibits symptoms such as rapid speech, difficulty interrupting, loose associations, lack of sleep, grandiose delusions, and auditory hallucinations. These symptoms suggest a manic episode with psychotic features.</p> <p>The most likely diagnosis is: (A) Bipolar disorder, manic, with psychotic features The other options do not fit as well with the patient's presentation: (B) Brief psychotic disorder typically lasts less than 1 month, and the patient has been exhibiting symptoms for 2 months. (C) Delusional disorder involves non-bizarre delusions without other psychotic symptoms, such as hallucinations. (D) Psychotic disorder due to a general medical condition cannot be determined without further medical evaluation, and the patient's presentation is more consistent with a manic episode with psychotic features. (E) Schizophrenia is characterized by persistent psychotic symptoms and a decline in functioning, which is not evident in this patient's presentation.</p>
	Incorrect	<p>A 65-year-old woman comes to the physician for a follow-up examination after blood pressure measurements were 175/105 mm Hg and 185/110 mm Hg 1 and 3 weeks ago, respectively. She has well-controlled type 2 diabetes mellitus. Her blood pressure now is 175/110 mm Hg. Physical examination shows no other abnormalities. Antihypertensive therapy is started, but her blood pressure remains elevated at her next visit 3 weeks later. Laboratory studies show increased plasma renin activity; the erythrocyte sedimentation rate and serum electrolytes are within the reference ranges. Angiography shows a high-grade stenosis of the proximal right renal artery; the left renal artery appears normal. Which of the following is the most likely diagnosis?</p> <p>(A) Atherosclerosis (B) Congenital renal artery hypoplasia (C) Fibromuscular dysplasia (D) Takayasu arteritis (E) Temporal arteritis</p>	<p>(C) Fibromuscular dysplasia The patient's presentation of resistant hypertension, increased plasma renin activity, and angiography showing high-grade stenosis of the proximal right renal artery in the absence of other inflammatory markers or abnormal laboratory results suggest fibromuscular dysplasia. This is a non-atherosclerotic, non-inflammatory vascular disease that can cause narrowing of medium-sized arteries, most commonly the renal arteries, leading to renovascular hypertension.</p>
Step 2CK	Correct	<p>An 87-year-old woman, gravida 2, para 2, is brought to the emergency department from a skilled nursing care facility because of profuse vaginal bleeding for 2 days. At the age of 41 years, she underwent a cold knife cone biopsy for cervical dysplasia. Annual follow-up Pap smears until the age of 65 years showed no abnormalities. She is in moderate distress. Temperature is 36.1 °C (97.0 °F), pulse is 84/min, respirations are 30/min, and blood pressure is 140/90 mm Hg. Abdominal examination shows mild distention and tenderness; no masses are palpated. Pelvic examination shows a large verrucous lesion on the perineum partially covering the anal verge. Speculum examination shows atrophic changes on the vagina and cervix. Which of the following is the most likely explanation for these findings?</p> <p>(A) Cervical carcinoma (B) External hemorrhoids (C) Rectal carcinoma (D) Rectal polyps (E) Vulvar carcinoma</p>	<p>(E) Vulvar carcinoma The presence of a large verrucous lesion on the perineum partially covering the anal verge in an 87-year-old woman with profuse vaginal bleeding suggests vulvar carcinoma. Vulvar carcinoma can present with a variety of symptoms, including bleeding, itching, pain, or a visible lesion. In this case, the patient's age and presentation point towards vulvar carcinoma as the most likely explanation for her findings.</p>
	Incorrect	<p>A 47-year-old woman comes to the office because of a 2-month history of progressive weakness and numbness of her right leg. She has a 3-year history of type 2 diabetes mellitus treated with metformin. She has worked as a stock person in a storeroom for 12 years; she often moves and lifts heavy objects and sometimes crawls into tight spaces.</p> <p>On neurologic examination, muscle strength is 5+/5 in the hip extensor, thigh abductor, hamstring, and gastrocnemius muscles on the right and intact on the left. Muscle strength is intact in the iliopsoas, thigh adductor, quadriceps, and tibial anterior muscles bilaterally. Deep tendon reflexes are 1+ in the right ankle and 2+ in the left ankle. Sensation to light touch is mildly decreased over the lateral and inferior aspects of the right foot. Sensation otherwise is intact. Which of the following is the most likely diagnosis?</p>	<p>I am not a doctor, but I can provide you with some general information about these conditions. Based on the information provided, the most likely diagnosis for this 47-year-old woman with a history of type 2 diabetes and progressive weakness and numbness in her right leg could be: (A) Common peroneal neuropathy Common peroneal neuropathy, also known as fibular neuropathy, affects the common peroneal nerve, which branches from the sciatic nerve and provides motor and sensory functions to the lower leg. The symptoms described, such as weakness in hip extensor, thigh abductor, hamstring, and gastrocnemius muscles on the right side, and decreased sensation in the lateral and inferior aspects of the right foot, are consistent with common peroneal neuropathy. People with diabetes are at higher risk of developing neuropathies, and the</p>

(continued)

Table 2. Continued.

USMLE Examination	Outcome	Question	ChatGPT-4's Response
		(A) Common peroneal neuropathy (B) Femoral neuropathy (C) L2 radiculopathy (D) L4 radiculopathy (E) Obturator neuropathy (F) S1 radiculopathy	<p>patient's occupation involving heavy lifting and crawling into tight spaces might contribute to nerve compression or injury.</p> <p>The other options seem less likely based on the provided information:</p> <p>(B) Femoral neuropathy: This condition would typically affect the quadriceps muscles and the medial aspect of the leg, which are not described in the scenario.</p> <p>(C) L2 radiculopathy: This would cause weakness in hip flexion and thigh adduction, which are not mentioned in the scenario.</p> <p>(D) L4 radiculopathy: This would generally affect the quadriceps and tibialis anterior muscles, which are intact in the patient.</p> <p>(E) Obturator neuropathy: This condition would affect thigh adduction, which is intact in the patient.</p> <p>(F) S1 radiculopathy: This would affect the gastrocnemius muscle and cause a decreased Achilles reflex, but the presentation doesn't fully match the patient's symptoms.</p> <p>It is essential to consult a healthcare professional for a proper evaluation and diagnosis.</p>
Step 3	Correct	<p>A study is being conducted to assess mesothelioma in shipyard workers. A large shipyard firm has provided the asbestos exposure records of all employees during the past 50 years. The health insurer for the workers has provided claims data that documents all chest x-rays and diagnoses of mesothelioma among current workers and retirees. The study enrolled shipyard workers who were diagnosed with mesothelioma and shipyard workers who were not diagnosed with mesothelioma. All subjects in the study had to have chest x-rays. Which of the following is the best rationale for selecting a comparison group that had chest x-rays?</p> <p>(A) Address confounding (B) Demonstrate causality (C) Minimize ascertainment bias (D) Reduce recall bias</p>	<p>(C) Minimize ascertainment bias</p> <p>Selecting a comparison group that had chest x-rays helps minimize ascertainment bias in the study.</p> <p>Ascertainment bias occurs when the detection or diagnosis of a disease is not consistent between the groups being compared. By ensuring that both the mesothelioma group and the comparison group had chest x-rays, the study minimizes the potential for biased detection of mesothelioma cases between the two groups.</p>
	Incorrect	<p>A 67-year-old man, who has been recovering in the hospital following surgical repair of a contained ruptured abdominal aortic aneurysm 10 days ago, suddenly develops severe abdominal pain. He does not have chest pain or shortness of breath.</p> <p>The patient's postoperative course had been uncomplicated until 4 days ago, when he developed a low-grade fever. He has been unable to tolerate a full liquid diet during the past 4 days. Medical history is remarkable for hypertension, hypercholesterolemia, and a myocardial infarction 6 years ago. Regular medications include metoprolol, lovastatin, and an 81-mg aspirin daily. He has smoked one pack of cigarettes daily for the past 40 years, and he typically drinks three to five beers daily. Vital signs now are temperature 38.1 °C (100.6 °F), pulse 68/min and regular, respirations 16/min, and blood pressure 150/90 mm Hg. Physical examination shows bilateral cataracts. Auscultation of the chest discloses diffuse mild wheezes and a grade 1/6 soft systolic ejection murmur at the cardiac apex. Palpation of the abdomen discloses mild to moderate mid and right upper quadrant abdominal tenderness. There is no guarding, rebound, or palpable mass. Rectal examination shows no abnormalities. Peripheral pulses are normal. Neurologic examination discloses no abnormalities.</p> <p>Serum electrolyte concentrations are within the reference ranges. Leukocyte count is 15,800/mm³ with mild left shift.</p> <p>Urinalysis shows no abnormalities. ECG shows an old inferior myocardial infarction. Chest x-ray shows scarring at the left and right cardiac bases; no masses or infiltrates are present. Which of the following is the most appropriate additional diagnostic study?</p> <p>(A) Arteriography of the abdomen (B) Blood cultures (C) MRI of the abdomen (D) Ultrasonography of the abdomen (E) No additional studies are needed</p>	<p>Given the patient's symptoms, recent surgical history, low-grade fever, and elevated leukocyte count, the most appropriate additional diagnostic study is:</p> <p>(B) Blood cultures</p> <p>The patient has recently undergone surgery for a contained ruptured abdominal aortic aneurysm and is now presenting with abdominal pain, low-grade fever, and an elevated leukocyte count. These findings raise suspicion for a possible postoperative infection. Blood cultures should be obtained to evaluate for potential bacteremia and guide appropriate antibiotic therapy. Other diagnostic studies may be considered later based on the results of blood cultures and the patient's clinical course.</p>

2023). Another study in March 2023 found that the large language models ChatGPT-4 and Bing Chat performed similar to human test takers across 250 ophthalmology board-style questions, all obtaining scores of over 70% (Cai et al. 2023). ChatGPT-4 has also notably outperformed ChatGPT-3.5 on multiple-choice sample questions from specialty certificate examination (SCE) question banks on dermatology and neurology, while exceeding thresholds for passing (Passby et al. 2023; Giannos 2023). Across 84 dermatology SCE practice questions, ChatGPT-3.5 answered 63.1% correctly and ChatGPT-4 answered 90.5% correctly (Passby et al. 2023). Moreover, across 69 neurology SCE practice questions, different releases of ChatGPT-3.5 answered 42–57% correctly and ChatGPT-4 answered 64% correctly (Giannos 2023).

Our present study was limited for a few reasons. The practice questions that served as the focus of this investigation were publicly and freely available, and it is possible that ChatGPT-4 may perform differently on the official USMLE examinations. Moreover, our analysis omitted practice questions that contained embedded figures or formatted tables, limiting the generalizability of the observed results. Given that ChatGPT-4 continuously updates its responses with user feedback, repeating the same methodology in our study may yield different results. Nonetheless, we attempted to limit these effects by asking ChatGPT-4 all multiple-choice questions within the same day and confirming that the chatbot provided equivalent responses to all questions on two separate internet browsers. The timed duration of ChatGPT-4's responses in our study may have been influenced by delays in response time and the status of OpenAI's servers at the time of use. A ChatGPT Plus account was used in the study, which is intended to perform well at high volumes of site traffic. ChatGPT-4 also requires cookies to operate, and its responses may be influenced by previous conversations, however we cleared all conversation history and refreshed the AI chatbot prior to inputting each question. Our findings should not be generalized to other disciplines and must be interpreted in the context of their date, as the performance of ChatGPT-4 will likely continue to improve with time.

In conclusion, we found that ChatGPT-4 responded to 88% of text-based questions correctly in the official practice question sets for the USMLE Step 1, Step 2CK, and Step 3 examinations. It is essential that medical professionals and medical students recognize the exponential growth of natural language processing models like ChatGPT-4 in medicine, evaluate its applicability in medical education, and scrutinize its limitations. We encourage future research to evaluate the knowledge of ChatGPT in various subspecialties and its utility as a companion study tool for medical licensing examination preparation.

Disclosure statement

The views expressed herein are those of the authors and do not necessarily reflect the position of the Federation of State Medical Boards or National Board of Medical Examiners. Information reported in this manuscript has not been previously presented at a conference. Data were collected from the artificial intelligence chatbot ChatGPT developed by OpenAI. As corresponding author, Rajeev H. Muni had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding

MMP: Financial support (to institution) – PSI Foundation, Fighting Blindness Canada. RHM: Consultant – Alcon, Apellis, AbbVie, Bayer, Bausch Health, Roche; Financial Support (to institution) – Alcon, AbbVie, Bayer, Novartis, Roche.

Notes on contributors

Andrew Mihalache is a MD candidate at the University of Toronto in Toronto, Ontario under the Temerty Faculty of Medicine.

Ryan S. Huang is a MD candidate at the University of Toronto in Toronto, Ontario, under the Temerty Faculty of Medicine.

Marko M. Popovic is the Chief Ophthalmology Resident in the Department of Ophthalmology and Vision Sciences at the University of Toronto and has completed a Master of Public Health at the Harvard T.H. Chan School of Public Health.

Rajeev H. Muni is a staff vitreoretinal surgeon at St. Michael's Hospital in Toronto, Ontario, Associate Professor and Vice-Chair of Clinical Research in the Department of Ophthalmology and Vision Sciences at the University of Toronto.

Data availability statement

The data that support the findings of this study may be requested at andrew.mihalache@mail.utoronto.ca, with support from the principal investigator RHM.

References

- Altman D, Machin D, Bryant T, Gardner M. 2000. Statistics with confidence 2nd ed. [Internet]. [accessed 2023 Jan 21]. <http://books.google.com/books?hl=en&lr=&id=Rfdg1MFx7mcC&oi=fnd&pg=PR11&dq=Statistics+With+Confidence&ots=7N6OkQBqd7&sig=JaOmVHLTmTt8P1alBysMBMU5kes>.
- Altman DG. 1990. Practical statistics for medical research. *Pract Stat Med Res* [Internet]. [accessed 2023 Jan 21]. https://www.medcalc.org/calc/comparison_of_means.php.
- Aydın Ö, Karaarslan E. 2022. OpenAI ChatGPT generated literature review: digital twin in healthcare. *SSRN J*. 2:22–31. doi:10.2139/ssrn.4308687.
- Azaria A. ChatGPT usage and limitations doi:10.13140/RG.2.2.26616.11526.
- Biswas S. 2023. ChatGPT and the future of medical writing. *Radiology* [Internet]. 307[(2):e223312. doi:10.1148/radiol.223312.
- Cahan P, Treutlein B. 2023. A conversation with ChatGPT on the role of computational systems biology in stem cell research. *Stem Cell Rep* [Internet]. 18(1):1–2. doi:10.1016/j.stemcr.2022.12.009.
- Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, Alabiad C. 2023. Performance of generative large language models on ophthalmology board style questions. *Am J Ophthalmol*. 254:141–149. doi:10.1016/j.ajo.2023.05.024.
- Campbell I. 2007. Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Stat Med*. 26(19):3661–3675. doi:10.1002/sim.2832.
- ChatGPT Generative Pre-trained Transformer, Zhavoronkov A. 2022. Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective. *Oncoscience* [Internet]. 9:82–84. doi:10.18632/oncoscience.571.
- Cohen ER, Goldstein JL, Schroedl CJ, Parlapiano N, McGaghie WC, Wayne DB. 2020. Are USMLE scores valid measures for chief resident selection? *J Grad Med Educ*. 12(4):441–446. doi:10.4300/JGME-D-19-00782.1.
- Comparison of proportions calculator. 2023. [accessed 2023 Jan 21]. https://www.medcalc.net/statisticaltests/comparison_of_proportions.php.
- Else H. 2023. Abstracts written by ChatGPT fool scientists. *Nature* [Internet]. 613(7944):423. accessed 2023 Jan 22] doi:10.1038/d41586-023-00056-7.
- Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, Pearson AT. 2022. Comparing scientific abstracts generated by ChatGPT to

- real abstracts with detectors and blinded human reviewers. *npj Digit.* [Internet]. [accessed 2023 Aug 17]: doi:10.1101/2022.12.23.521610.
- Giannos P. 2023. Evaluating the limits of AI in medical specialisation: chatGPT's performance on the UK Neurology Specialty Certificate Examination. *BMJ Neurol Open.* 5(1):e000451. doi:10.1136/bmjno-2023-000451.
- Gilson A, Safranek C, Huang T, Socrates V, CL, Taylor RA, Chartash D. 2022. How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment. *medRxiv* [Internet]. [accessed 2023 Jan 18]. doi:10.1101/2022.12.23.22283901.
- GPT-4. [accessed 2023 Mar 20]. <https://openai.com/product/gpt-4>.
- Hu R, Fan KY, Pandey P, Hu Z, Yau O, Teng M, Wang P, Li A, Ashraf M, Singla R. 2022. Insights from teaching artificial intelligence to medical students in Canada. *Commun Med.* 2(1):1–5. doi:10.1038/s43856-022-00125-4.
- Jeblick K, Schachtner B, Dextl J, Mittermeier A, Stüber AT, Topalis J, Weber T, Wesp P, Sabel B, Ricke J, et al. 2022. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports [Internet]. [accessed 2023 Jan 22] doi:10.48550/arxiv.2212.14882.
- Katz DM, Bommarito MJ, Gao S, Arredondo P. 2023. GPT-4 passes the bar exam. *SSRN J.* doi:10.2139/ssrn.4389233.
- Kirkwood BB, Sterne J. 2003. Essential medical statistics [Internet].:395–412. [accessed 2023 Jan 21]. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Essential+Medical+Statistics#0>.
- Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, et al. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2(2):e0000198. doi:10.1371/JOURNAL.PDIG.0000198.
- Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, et al. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health.* [2(2):e0000198. doi:10.1371/JOURNAL.PDIG.0000198.
- Liévin V, Hother CE, Winther O. 2022. Can large language models reason about medical questions? [Internet]. [accessed 2023 Mar 22]. <http://arxiv.org/abs/2207.08143>.
- MedCalc. 2023. MedCalc's comparison of means calculator [Internet]. [accessed 2023 Jan 21]. https://www.medcalc.org/calc/comparison_of_means.php.
- Mihalache A, Huang RS, Popovic MM, Muni RH. 2023. Performance of an upgraded artificial intelligence Chatbot for ophthalmic knowledge assessment. *JAMA Ophthalmol.* 141: 798–800. doi:10.1001/JAMAOPHTHALMOL.2023.2754.
- Mihalache A, Popovic MM, Muni RH. 2023. Performance of an artificial intelligence Chatbot in an ophthalmic knowledge assessment. *JAMA Ophthalmol.* 141(6):589–597. doi:10.1001/jamaophthalmol.2023.1144.
- Milmo D. 2023. ChatGPT reaches 100 million users two months after launch | Chatbots | The Guardian. *Guard* [Internet]. [accessed 2023 Mar 20]. <https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app>.
- O'Connor, S, ChatGPT. 2023. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? *Nurse Educ Pract.* 66:103537. doi:10.1016/j.nepr.2022.103537.
- OpenAI. 2023. GPT-4 technical report [Internet]. [accessed 2023 Mar 20]. <http://arxiv.org/abs/2303.08774>.
- Passby L, Jenko N, Wernham A. 2023. Performance of ChatGPT on dermatology specialty certificate examination multiple choice questions. *Clin Exp Dermatol.* doi:10.1093/ced/llad197.
- Richardson JTE. 2011. The analysis of 2×2 contingency tables—Yet again. *Stat Med.* 30(8):890. doi:10.1002/sim.4116.
- Sanderson K. 2023. GPT-4 is here: what scientists think. *Nature* [Internet]. 615(7954):773–773. doi:10.1038/d41586-023-00816-5.
- Statistics SS. One-way ANOVA calculator, including Tukey HSD [Internet]. <https://www.socscistatistics.com/tests/anova/default2.aspx>.
- Step 1 Exam Content | USMLE. [accessed 2023 Mar 22]. <https://www.usmle.org/step-exams/step-1/step-1-exam-content>.
- Step 2 CK Exam Content | USMLE. [accessed 2023 Mar 22]. <https://www.usmle.org/step-exams/step-2-ck/step-2-ck-exam-content>.
- Step 2 Clinical Knowledge (CK) SAMPLE TEST QUESTIONS. 2023.
- Step 3 Exam Content | USMLE. [accessed 2023 Mar 22]. <https://www.usmle.org/step-exams/step-3/step-3-exam-content>.
- Step 3 Sample Questions August 2022. 2022.
- Stokel-Walker C. 2022. AI bot ChatGPT writes smart essays—should academics worry? *Nature.* doi:10.1038/d41586-022-04397-7.
- Susnjak T. 2022. ChatGPT: the end of online exam integrity? [Internet]. [accessed 2023 Jan 20] doi:10.48550/arxiv.2212.09292.
- USMLE Step 1 Sample Test Questions. 2022.
- Zhai X. 2022. ChatGPT user experience: implications for education. *SSRN J.* [accessed 2023 Jan 20] doi:10.2139/ssrn.4312418.