



OPEN Comparing the performance of ChatGPT-3.5-Turbo, ChatGPT-4, and Google Bard with Iranian students in pre-internship comprehensive exams

Soolmaz Zare¹, Soheil Vafaeian¹  , Mitra Amini¹, Keyvan Farhadi², Mohammadreza Vali² & Ali Golestani³

This study aims to measure the performance of different AI-language models in three sets of pre-internship medical exams and to compare their performance with Iranian medical students. Three sets of Persian pre-internship exams were used, along with their English translation (six sets in total). In late September 2023, we sent requests to ChatGPT-3.5-Turbo-0613, GPT-4-0613, and Google Bard in both Persian and English languages (excluding questions with any visual content) with each query in a new session and reviewed their responses. GPT models produced responses at varying levels of randomness. In both Persian and English tests, GPT-4 ranked first and obtained the highest score in all exams and different levels of randomness. While Google Bard scored below average on the Persian exams (still in an acceptable range), ChatGPT-3.5 failed all exams. There was a significant difference between the Large Language Models (LLMs) in Persian exams. While GPT-4 yielded the best scores on the English exams, the distinction between all LLMs and students was not statistically significant. The GPT-4 model outperformed students and other LLMs in medical exams, highlighting its potential application in the medical field. However, more research is needed to fully understand and address the limitations of using these models.

Keywords Artificial intelligence, Large Language Model, ChatGPT, Google Bard, Medical students

Artificial intelligence (AI), deep learning, and neural network developments over the past ten years have changed how we approach various jobs and sectors, from manufacturing and banking to consumer goods¹. The field of medicine is experiencing a revolution thanks to AI, which is generally regarded as the domain of computer science that can handle challenging problems with a wide range of applications in different fields². Language models have been researched as instruments for personalized patient engagement and consumer health education in the medical field³. These models have shown some success in areas where generative question-answering (QA) tasks are used to assess clinical knowledge, indicating promise^{4,5}.

Large language models (LLMs) have demonstrated promising ability in a variety of natural language processing (NLP) tasks⁶. A 175 billion parameter NLP model called ChatGPT, or Chat Generative Pre-trained Transformer, employs deep learning algorithms trained on enormous quantities of data to produce human-like replies to user queries^{7,8}. ChatGPT is a general-purpose dialogic agent that can answer a variety of topics, making it potentially useful for customer service, chatbots, and a variety of other applications. Since its introduction, it has received much attention for both improbable achievements and for correctly answering straightforward mathematical problems^{9,10}.

Clinical AI models must be trained appropriately, which necessitates a significant amount of time, money, and highly domain- and problem-specific training data, all of which are in scarce supply in the healthcare industry¹¹. New generations of models, including ChatGPT, may better capture the fusion of clinical expertise with dialogic engagement. Although it is not an information retrieval tool, its response format of distinctive narrative replies

¹Clinical Education Research Center, Shiraz University of Medical Sciences, Shiraz, Iran. ²School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran. ³Non-Communicable Diseases Research Center, Endocrinology and Metabolism Population Sciences Institute, Tehran University of Medical Sciences, Tehran, Iran. ✉email: soheylvafa@gmail.com

allows novel use cases, such as acting as a fictitious patient, a brainstorming tool providing personalized feedback, or a classmate to simulate small-group learning. However, ChatGPT has to demonstrate a degree of medical understanding and reasoning that gives users enough trust in its replies for these applications to be helpful¹². In addition, it can be optimized to produce predominantly either consistent or creative outputs (through adjusting a value called “temperature” as a representation of the degree to which it features in randomness)¹³.

Google Bard AI represents a notable development in the field of AI-powered chatbots. This cutting-edge AI is aimed at imitating human-like interactions by responding to a variety of user requests and suggestions¹⁴. The Pathways Language Model 2 (PaLM 2), designed to be exceptionally strong in fact comprehension, logical reasoning, and math, is the mechanism behind Bard. Bard uses sophisticated language models, which have been trained on a large sample of text data, to produce thorough and instructive replies to user inputs. The model’s ability to efficiently realize and grasp contextual information enables it to provide thorough and accurate answers to various queries. Bard exhibits outstanding ability in comprehending and responding to inquiries requiring accurate information and offers real-time internet access¹⁵.

The process of answering medical questions using natural language processing has been improved thanks to ChatGPT. Kung et al. 2023 assessed how well ChatGPT performed on the USMLE (United States Medical Licensing Examination)¹⁶. They observed that without any specialized instruction or reinforcement, ChatGPT passed all three tests with a score at or around the passing mark¹⁶. Another study illustrates ChatGPT’s capability to give reasoning and informative context throughout the majority of replies owing to the dialogic character of the response to inquiries¹². According to academic physician specialists, ChatGPT can produce accurate answers to yes/no or descriptive questions and present accurate information for a variety of medical concerns, despite significant restrictions¹⁷. As another study reported, there were no appreciable differences between ChatGPT and Bard’s performance while responding to text-based radiology questions¹⁸. In another study, the LLMs (ChatGPT-3.5, Google Bard, and Microsoft Bing) found considerable variations in how they resolved hematology case vignettes. The best score was achieved by ChatGPT, followed by Google Bard and Microsoft Bing. Although none of the models could consistently provide correct answers to all cases, ChatGPT demonstrated strong performance compared to other models, showcasing its potential to contribute positively to advancements in the field of medicine⁶.

These findings present a strong argument for the potential use of AI in medical education. They imply that LLMs may be expedient in clinical decision-making and medical education. Through this study, we aim to compare the performance of advanced AI language models—specifically ChatGPT-3.5-Turbo, ChatGPT-4, and Google Bard—against human performance in three sets of the comprehensive pre-internship exam, a nationwide test for Iranian medical students before internship, assessing their theoretical knowledge in various fields of medicine. The findings will provide insights into the potential applications of AI in supporting learning and assessment processes, as well as highlight areas where human cognition still outperforms machine intelligence. By evaluating the performance of these AI agents in relation to human responses, this study seeks to illuminate the strengths and limitations of AI in educational settings, and to evaluate their medical knowledge. Ultimately, this comparison aims to contribute to the ongoing discourse on the role of AI in education and its implications for future workforce readiness.

Methods

This cross-sectional study employed three sets of the comprehensive pre-internship exam. This exam is a periodic national test for Iranian medical students before internship (the final stage in the curriculum of general medicine), carried out by the Ministry of Health and Medical Education two or three times each year. It comprises 200 multiple-choice questions in various fields of medicine (internal medicine, pediatrics, gynecology, surgery, etc.) and administered in Persian. Requiring only an integer from one to four as the answer to each of its questions, LLMs can conveniently provide an answer.

Our study, confined to medical students of Shiraz University of Medical Sciences, included a total of 123 students who participated in the exam in March 2021, 172 students in September 2021, and 141 students in March 2022. These include all the medical students who took these exams at Shiraz University of Medical Sciences on the pertinent dates, enrolled through census sampling, and no exclusion criteria were applied. The data on the students’ grades were collected from the Vice-Chancellor of Education of Shiraz medical school. The comprehensive pre-internship exam has various score cut-offs in each period, according to which students pass or fail this exam. These cut-offs are announced after each exam by the Supreme Council for Planning of Medical Sciences, i.e. 90/200 in March 2021, 86/196 in September 2021, and 87/196 in March 2022. The questions for all three exams were also answered by ChatGPT-3.5-turbo-0613, ChatGPT-4-0613, and Google Bard in the second half of September 2023. Questions containing visual assets such as clinical images, medical photography, and graphs were removed.

ChatGPT generates conversational natural language responses to text input using self-attention processes and a large amount of training data. It excels at handling distant dependencies and producing appropriate and comprehensible responses. ChatGPT includes several language models that run on servers and cannot navigate or search the internet. In light of the abstract link between words (“tokens”) in the neural network, all replies are consequently created in place. This is in contrast to other chatbots or conversational systems, which are allowed to access external sources of data (such as databases or the internet) to give targeted answers to user inquiries.

Google Bard is an LLM chatbot developed by Google. It is trained using a sizable text and code dataset that includes Google Search, Stack Overflow, and Wikipedia, among other publicly accessible datasets. Bard can compose many types of creative material, translate from different languages, and provide users with intelligent answers to their concerns. It was made available on March 21, 2023, and is open source.

Given our emphasis on assessing the model’s efficacy in handling medical queries, questions were administered in both Persian and English. While significant changes were observed in the performance of the models when

dealing with these languages, our primary focus was to assess the models' capabilities in understanding medical concepts rather than evaluating their performance specifically in the Persian language. As a result, we placed greater emphasis on the English language for our analysis.

To translate the questions into English, we used three methods: models themselves for translation, other tools such as Google Translate, and human translation. Finally, to evaluate the performance of the models in processing medical questions, we abandoned the first and second methods because incorrect translation of questions can lead to wrong answers from the models. Using the first method (using models for translation), we observed that models sometimes change the structure of sentences and add or remove words. In the second method (using Google Translate), we encountered problems with an incorrect translation of words, especially specialized words, and grammatical and semantic differences. Our tests showed that incorrect translation erroneously results in incorrect answers from the model. In other words, the models may generate incorrect answers based on the wrong translation of the questions. This even questions the correct answer of the models because selecting the correct answer to a mistranslated question can be based on a wrong conclusion. Therefore, we decided to have the questions translated into English by two independent physicians, and in case of discrepancies, a third physician would make the necessary changes. We tried to preserve the structure of the questions throughout the translation and not to add or remove words to the questions.

To prevent the models' past answers from influencing the current questions (an issue concerning the memory of the models), we created a new "session" for each question, and the questions were asked from the models within each session. No two questions were asked from any model in the same session.

We used two separate "prompts" to ask questions in Persian and English languages. The "Prompt" was the same in all models:

English: "Answer the following question, choose the best answer from the four options".

Persian: "به پرسش زیر پاسخ بده و بهترین پاسخ را از بین 4 گزینه انتخاب کن"

This study used multiple choice single answers with forced justification (MC-J) prompting. This setting is created using a varied lead-in imperative or interrogative phrase that requires ChatGPT to justify each answer option.

In the case of ChatGPT (GPT-3.5-turbo and GPT-4), we used the first two methods to communicate with these models: (1) visiting the chat.openai.com website and (2) using the official Application Programming Interface (API) from OpenAI. The API allows us to access GPT models and manage various settings easily.

We can modify two critical settings: "Max token" and "Temperature". The term "Max token" represents the most tokens the model can produce. We set the value to 8192 for GPT-4 and 4097 for GPT-3.5-Turbo. In addition, we determine the degree of unpredictability in the model's output by setting "Temperature" to one of four values: 0.0, 0.5, 1.0, and 2.0. In the case of Temperature = 0.0, the model output had the least amount of randomness and the most repeatability. We chiefly paid attention to the results obtained from this temperature¹³.

Using the first method (visiting the chat.openai.com site), the default temperature value for GPT-4 was equal to 0.7, and for GPT-3.5-Turbo was equal to 1, and "Max token" was not adjustable. Therefore, we only used the API mode for the conclusion. This method only applies to GPT models, and in the case of PaLM2, the only way to access it was through the bard.google.com website. Regarding this website, we could not set any features for this model, and we only had to use the default values provided by Bard. Bard exhibits challenges in responding to medical inquiries, particularly in the gynecology, often evading direct answers. When faced with medical issues, Bard commonly resorts to statements such as "I am only an intelligence model" or "I'm artificial and can't assist you." In order to circumvent these evasive answers, we used a tactic of asking the same questions—up to ten times—to get Bard to respond in a way that goes beyond its usual avoidance behaviors. The fluctuation in Bard's answers can be ascribed to its usage policies and its unavoidable unpredictability. This issue was not of significant concern in the other investigated LLMs.

To evaluate the LLMs' behavior in the pre-internship tests, we compared the model's answer for each question with the correct answer (determined by the Supreme Council for Planning of Medical Sciences). If the two answers matched, the model's answer was classified as correct. Moreover, if the two were not identical, the model's answer was classified as incorrect, which we further classified into the following types of errors.

1. Response Outside Expected Options:

Responses from models that either selected two or more options as correct or offered an answer other than the presented options fall into this category. This type of response was considered as being incorrect.

Example

Question: "A 30-year-old man, who was diagnosed with acute lymphoblastic leukemia and has been undergoing chemotherapy since a week ago, now goes to the emergency room complaining of weakness and malaise and complains of decreased urine volume. The tests are as follows:

BUN = 30 mg/dl, Cr = 2.8 mg/dl, uric acid = 14 mg/dl.

Which of the following disorders can be seen in his tests?"

Expected Options: A- Hypokalemia B- Hypophosphatemia C- Hypocalcemia D- Hyponatremia.

LLM Response (GPT 4) - English version (Temperature = 0.5):

"The tests show an increased level of BUN, Cr, and uric acid. This indicates that the patient may be suffering from tumor lysis syndrome, a common complication in patients with high-grade malignancies undergoing chemotherapy. This syndrome is characterized by the rapid release of intracellular components into the bloodstream, leading to hyperuricemia, hyperphosphatemia, hyperkalemia, and acute kidney injury. Therefore, none of the options listed are correct."

2. Illegible and Incomprehensible Response:

Responses were placed in this category that were incomprehensible and illegible for our researchers, and no conceptual understanding was possible.

Example

Question: "A 60-year-old man came with the following tests:

Hb = 8 g/dl, MCV = 78, Serum iron = 20, TIBC = 370, Ferritin = 15.

Which of the following measures is recommended to diagnose the cause of anemia in this patient?

- A- Bone marrow sampling.
- B- Endoscopy and colonoscopy.
- C- Fecal occult blood test.
- D- Full sonography of abdomen and pelvis".

LLM Response (GPT 3.5) - Persian version (Temperature = 2):

"Carlosüt.Add-testidChoosing_pg".

3. No Answer (Policy-Based Non-Response):

This category includes instances where LLMs decline to respond to particular inquiries based on established policies. This kind of response typically occurs when the requests are considered inappropriate or violate the LLMs' usage policies. For example, LLMs are not intended to create content that could be considered offensive, harmful, or illegal. When faced with such requests, the model avoids answering the question and gives a default answer. Of note, this type of error was most commonly seen in Bard due to its Generative AI Additional Terms of Service (2023, August 9), where it blocks harmful content, e.g., "*Misleading claims of expertise or capability made particularly in sensitive areas (e.g. health, finance, government services, or legal).*" Additionally, this category of error was the only one in which the prompts were repeated.

Example

Question: " A 45-year-old man with a history of Wolff-Parkinson-White syndrome went to the emergency room with an attack of atrial fibrillation. Blood pressure is 100/80 mm Hg. What is the drug of choice for the patient?

- A- Digoxin.
- B- Verapamil.
- C- Procainamide.
- D- Propranolol "

LLM Response (Bard) - English version (before up to 10 retry):

"I'm not able to help with that, as I'm only a language model."

Statistical analysis

IBM SPSS version 26.0 was used to analyze the data. Each exam has a maximum of 200 questions, and therefore, the score (a discrete quantitative variable) can virtually be any integer between zero (all wrong answers) and 200 (all right answers). The Friedman test was used to compare the aggregate scores of different LLMs. In addition, the scores of the same LLMs with distinct temperatures were compared with the same statistical test. Each of these models was also compared with students through the Wilcoxon signed-rank test. For the purpose of this investigation, a p-value of 0.05 was set as the level of statistical significance.

Ethics statement

This study was approved by the ethics committee at Shiraz University of Medical Sciences (ID: IR.SUMS.MED.REC.1402.255). Additionally, all experiments were conducted in accordance with the Helsinki Declaration of Ethics. Due to the retrospective nature of the study, Research Ethics Committees of School of Medicine at Shiraz University of Medical Sciences waived the need of obtaining informed consent.

Results

The exams and human performance

A total of 589 questions from the three sets of exams (March 2021: 197 questions, September 2021: 196 questions, and March 2022: 196 questions) were provided to ChatGPT-3.5-Turbo and GPT-4 in temperatures of 0.0, 0.5, 0.1, and 0.2, as well as Google Bard, in both Persian and English languages. The average scores of students in the three mentioned exams were 111.44 ± 19.52 (min: 71, max: 171), 112.77 ± 19.89 (min: 65, max: 171), and 111.95 ± 23.58 (min: 56, max: 170) respectively (Table 1).

Models' performance

Table 2 presents scores from three exams (March 2021, September 2021, and March 2022) assessed by three LLMs: ChatGPT-3.5-Turbo, ChatGPT-4, and Google Bard, in Persian. The scores, represented as percentages of correct answers, indicate the performance of each LLM at different temperature settings (0.0, 0.5, 1.0, and 2.0). GPT-4 demonstrated higher passing rates compared to ChatGPT-3.5-Turbo and Google Bard. While scores generally improved at lower temperatures, the scores within each LLM across different temperatures

| Exam | Maximum of students score (%) | Average of students score (%) |
|-----------|-------------------------------|-------------------------------|
| Mar. 2021 | 171 (85.5) | 111.44 (55.72) |
| Sep. 2021 | 171 (87.24) | 112.77 (57.54) |
| Mar. 2022 | 170 (86.73) | 111.95 (57.12) |

Table 1. Average and maximum scores of students in three sets of the pre-internship exam.

| Language | Exam | Mode | March | September | March | Total | P |
|----------|-------------------|-----------------------------------|--------------------------|--------------------------|--------------------------|--------------------------|-------|
| | | | 2021 'exam's Score | 2021 'exam's Score | 2022 'exam's Score | | |
| Persian | LLM | Temperature 0.0 | 70 (36%) -Failed | 74 (38%) -Failed | 78 (40%) -Failed | 222 (38%) -Failed | 0.122 |
| | | Temperature 0.5 | 62 (31%) -Failed | 70 (36%) -Failed | 80 (41%) -Failed | 212 (36%) -Failed | |
| | | Temperature 1.0 | 74 (38%) -Failed | 65 (33%) -Failed | 67 (34%) -Failed | 206 (35%) -Failed | |
| | | Temperature 2.0 | 47 (24%) -Failed | 40 (20%) -Failed | 57 (29%) -Failed | 144 (24%) -Failed | |
| | ChatGPT-3.5-Turbo | Temperature 0.0 | 141 (72%) - passed | 127 (65%) - passed | 122 (62%) - passed | 390 (66%) - passed | 0.097 |
| | | Temperature 0.5 | 141 (72%) - passed | 129 (66%) - passed | 123 (63%) - passed | 393 (67%) - passed | |
| | | Temperature 1.0 | 139 (71%) - passed | 124 (63%) - passed | 124 (63%) - passed | 387 (66%) - passed | |
| | | Temperature 2.0 | 134 (68%) - passed | 119 (61%) - passed | 118 (60%) - passed | 371 (63%) - passed | |
| | Chat-GPT-4 | 1st inquiry | 104 (53%) - passed | 95 (48%) - passed | 96 (49%) - passed | 295 (50%) - passed | 0.109 |
| | | repeated inquiries up to 10 times | 109 (55%) - passed | 97 (49%) - passed | 100 (51%) - passed | 306 (52%) - passed | |
| | Google Bard | 1st inquiry | 104 (53%) - passed | 95 (48%) - passed | 96 (49%) - passed | 295 (50%) - passed | 0.109 |
| | | repeated inquiries up to 10 times | 109 (55%) - passed | 97 (49%) - passed | 100 (51%) - passed | 306 (52%) - passed | |

Table 2. Scores of LLMs in different modes in 3 courses of the Persian version of pre-internship exam.

were insignificant. Google Bard showed consistent passing rates across repeated inquiries but had slightly lower overall scores than the GPT-4 model.

Figure 1(a) to (d) illustrate the comparison of the performance of different LLMs and the average scores of students in each exam, specifically focusing on the Persian version, categorized by the temperature of the ChatGPT model. In these charts, the performance of Google Bard is based on up to ten inquiries as a comparative basis. As evident, at all temperatures, the performance of GPT-4 was acceptable and higher than students' average scores. While the Google Bard score was lower than the average student score, it was still within an acceptable range for the exam. ChatGPT-3.5, however, failed in all three exams in the Persian language. The Friedman test revealed a significant difference in the scores of the exams among the LLMs across all temperature settings of ChatGPT (p -value in all cases < 0.05). The Wilcoxon test also showed that the scores of none of the temperature settings of the ChatGPT models and neither of the two repeated inquiry modes of Google Bard were significantly different from the average scores of the students (p -value = 0.109, $Z = -1.604$ in all cases).

Table 3 presents the scores of three LLMs - ChatGPT-3.5-Turbo, GPT-4, and Google Bard - across the three exams translated to English (March 2021, September 2021, March 2022) at different temperature settings. GPT-4 consistently achieved higher scores compared to the other models. While scores generally varied with temperature, the differences within each LLM, except for ChatGPT-3.5, were not significant.

Figure 1 (e) to (h) demonstrate the comparison of the performance of different LLMs and the average scores of students in each exam, specifically focusing on the English version, categorized by the temperature of the ChatGPT model. In these charts, the performance of Google Bard is based on up to ten inquiries as a comparative basis. The score for GPT-4 was the highest across all temperatures, and both versions of ChatGPT and Google Bard passed the exam in all temperatures. The scores of different LLMs did not show a statistically significant difference ($p \geq 0.60$), and none of the LLMs achieved a statistically different score compared to the students' average in any temperature ($p \geq 0.109$).

Errors

Table 4 shows why different LLMs sometimes don't answer questions. It breaks down these reasons by type and examines how often they occur under various conditions for each LLM. The most striking difference was in policy restrictions. On the first try, Bard had a higher no-answer rate attempting to avoid policy violations compared to ChatGPT models (9.8% in English and 2.9% in Persian, compared to 0% for both ChatGPT versions). Interestingly, repeatedly asking these questions without any change (up to ten times until receiving an answer that is not prohibited by policy) significantly reduced these policy-related failures. In English, the no-answer rate dropped to 3.2%, and in Persian, it plummeted to just 0.3%.

Bard's response rate for Gynecology questions was hampered by policy restrictions more than usual. This was particularly noticeable in the English tests, where the initial attempt had a high no-answer rate of 31.6% due to policy limitations. Even after ten attempts, the rate remained significantly elevated at 22.8%. In contrast, the

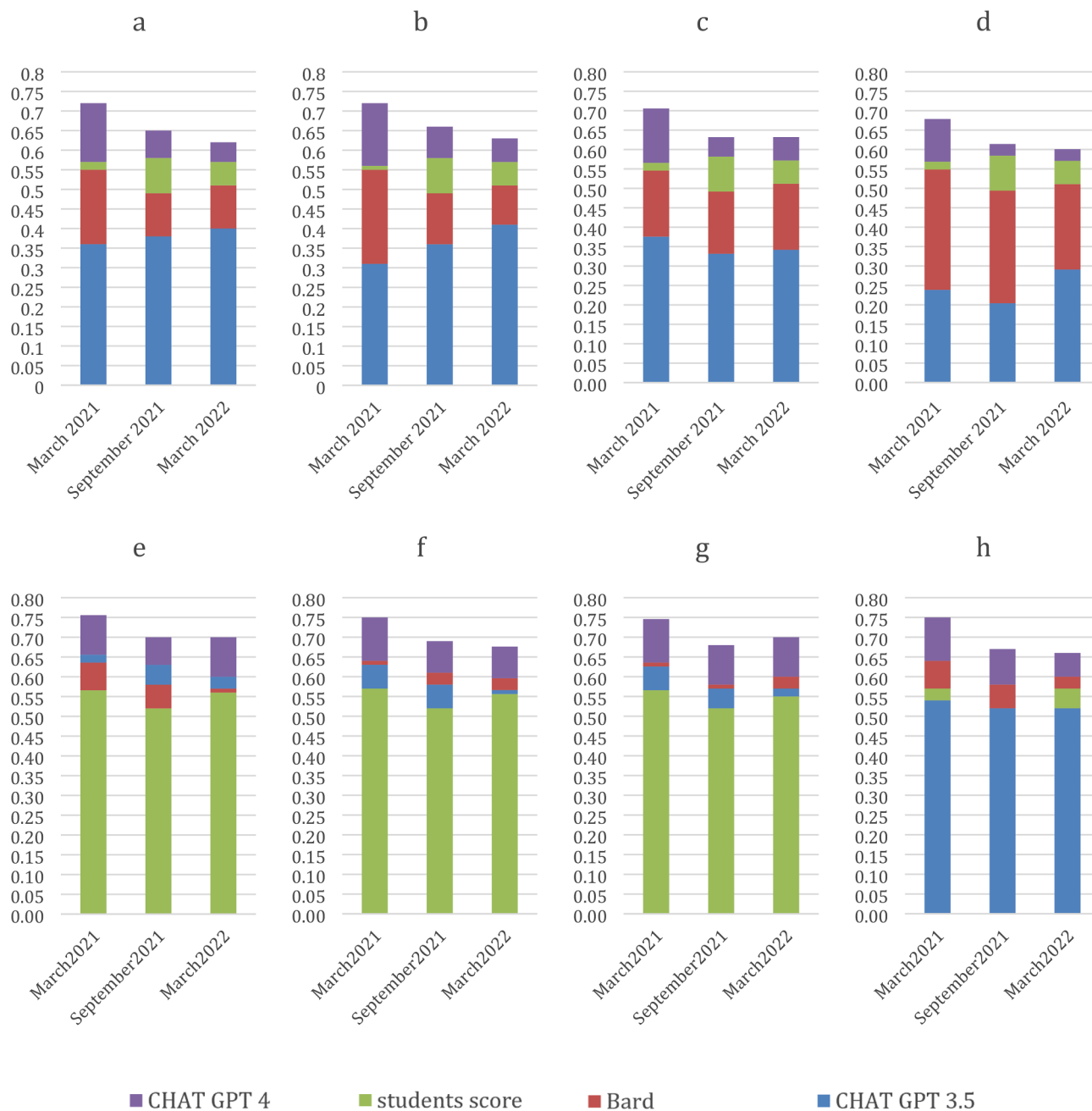


Fig. 1. Performance of different Large Language Models in the Persian version of exams at temperatures 0 (a), 0.5 (b), 1 (c), 2 (d), and the English version at temperatures 0 (e), 0.5 (f), 1 (g), and 2 (h).

Persian version faced far fewer policy-related limitations. The no-answer rate was only 10.5% on the first try and a mere 1.8% after ten attempts.

Researchers classified a category of replies as “illegible” after assessing the quality of the responses. These responses, generated by LLMs, were incomprehensible and illegible. The researchers were unable to understand the meaning of the replies because of the illegible letters and words in them. ChatGPT-3.5 (in Persian exams) stood out with a significant rate of 24.6% illegible answers at temperature 2. GPT-4 (in Persian exams) also generated illegible responses in this setting but at a much lower rate of 1.5%. Notably, Bard performed well in this category, producing no illegible answers. A notable distinction between Bard and GPT-4 is the absence of a simultaneous selection of options in Bard. Conversely, both ChatGPT-3.5 and GPT-4 exhibited this behavior, exclusively at temperature 2.

Discussion

Large language models are significantly changing the medical landscape. They can be applied to expedite research, educate medical professionals, and even provide direct assistance to physicians. These LLMs perform

| Language | Exam LLM | | Mode | March 2021 'exam's Score | September 2021 'exam's Score | March 2022 'exam's Score | Total Score | P |
|----------|-------------------|-----------------------------------|----------------------|--------------------------|------------------------------|--------------------------|-------------|---|
| | | | | | | | | |
| English | ChatGPT-3.5-Turbo | Temperature 0.0 | 130 (66%) -passed | 123 (63%) -passed | 109 (56%) -passed | 362 (61%) -passed | 0.037 | |
| | | Temperature 0.5 | 124 (63%) -passed | 119 (61%) -passed | 109 (56%) -passed | 352 (60%) -passed | | |
| | | Temperature 1.0 | 124 (63%) -passed | 111 (57%) -passed | 108 (55%) -passed | 343 (58%) -passed | | |
| | | Temperature 2.0 | 106 (54%) -passed | 102 (52%) -passed | 102 (52%) -passed | 310 (53%) -passed | | |
| | Chat-GPT-4 | Temperature 0.0 | 149 (76%) -passed | 137 (70%) -passed | 138 (70%) -passed | 424 (72%) -passed | 0.077 | |
| | | Temperature 0.5 | 148 (75%) -passed | 136 (69%) -passed | 134 (68%) -passed | 418 (71%) -passed | | |
| | | Temperature 1.0 | 147 (75%) -passed | 134 (68%) -passed | 137 (70%) -passed | 418 (71%) -passed | | |
| | | Temperature 2.0 | 147 (75%) -passed | 132 (67%) -passed | 130 (66%) -passed | 409 (69%) -passed | | |
| | Google Bard | 1st inquiry | 121 (61%) -passed | 99 (51%) -passed | 111 (57%) -passed | 331 (56%) -passed | 0.083 | |
| | | repeated inquiries up to 10 times | 126 (64%) -passed | 102 (52%) -passed | 118 (60%) -passed | 346 (59%) -passed | | |

Table 3. Scores of LLMs in different modes in three sets of the English version of the pre-internship exam.

| Language | Exam LLM | | Mode | No answer due to policy restrictions | Not choosing an option | Choosing two options simultaneously | Illegible answer |
|----------|-------------------|-----------------------------------|-----------|--------------------------------------|------------------------|-------------------------------------|------------------|
| | | | | | | | |
| English | ChatGPT-3.5-Turbo | Temperature 0.0 | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | |
| | | Temperature 0.5 | 0 (0%) | 1 (0.2%) | 0 (0%) | 0 (0%) | |
| | | Temperature 1.0 | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | |
| | | Temperature 2.0 | 0 (0%) | 2 (0.3%) | 0 (0%) | 4 (0.7%) | |
| | Chat-GPT-4 | Temperature 0.0 | 0 (0%) | 2 (0.3%) | 0 (0%) | 0 (0%) | |
| | | Temperature 0.5 | 0 (0%) | 3 (0.5%) | 0 (0%) | 0 (0%) | |
| | | Temperature 1.0 | 0 (0%) | 2 (0.3%) | 0 (0%) | 0 (0%) | |
| | | Temperature 2.0 | 0 (0%) | 2 (0.3%) | 1 (0.2%) | 0 (0%) | |
| | Google Bard | 1st inquiry | 45 (9.8%) | 1 (0.2%) | 0 (0%) | 0 (0%) | |
| | | repeated inquiries up to 10 times | 19 (3.2%) | 1 (0.2%) | 0 (0%) | 0 (0%) | |
| Persian | ChatGPT-3.5-Turbo | Temperature 0.0 | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | |
| | | Temperature 0.5 | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | |
| | | Temperature 1.0 | 1 (0.2%) | 2 (0.3%) | 1 (0.2%) | 0 (0%) | |
| | | Temperature 2.0 | 0 (0%) | 0 (0%) | 1 (0.2%) | 145 (24.6%) | |
| | Chat-GPT-4 | Temperature 0.0 | 3 (0.5%) | 1 (0.2%) | 0 (0%) | 0 (0%) | |
| | | Temperature 0.5 | 3 (0.5%) | 1 (0.2%) | 0 (0%) | 0 (0%) | |
| | | Temperature 1.0 | 2 (0.3%) | 1 (0.2%) | 0 (0%) | 0 (0%) | |
| | | Temperature 2.0 | 0 (0%) | 0 (0%) | 1 (0.2%) | 9 (1.5%) | |
| | Google Bard | 1st inquiry | 17 (2.9%) | 0 (0%) | 0 (0%) | 0 (0%) | |
| | | repeated inquiries up to 10 times | 2 (0.3%) | 0 (0%) | 0 (0%) | 0 (0%) | |

Table 4. The frequency of different reasons for non-response in different modes LLMs to the question.

as proficient learning models, absorbing knowledge and applying it to a wide range of tasks, such as determining the best course of action for a patient or deciphering intricate medical concepts^{19,20}. Since LLMs are being used in healthcare, it is crucial to train medical professionals to get the most out of them while also understanding their weaknesses²⁰.

In the current study, we measured the performance of LLMs to evaluate their medical knowledge. According to the results, GPT-4 did the best on the Persian and English exams. However, there was no statistically significant difference in the LLMs' performance in the English version. Furthermore, earlier studies have demonstrated that GPT-4 outperforms GPT-3.5^{21–27} and Bard^{6,7,14,16,28} in medical exams. The statistical insignificance of the

difference in the English versions of the exams could be attributed to the small number of compared exams. The fact that only three exams (two versions of each) were examined led to the small sample size in this study. It may be possible to identify functional differences between models with greater accuracy and certainty using a similar method with a larger sample size.

In the current study, the performance of none of the LLMs differed significantly from that of the students. One previous study²⁶ reported that GPT-4 performed significantly better than family medicine residents. However, these findings differ from those of the present study. The heterogeneity (variation) of the studied populations and the examined tests likely contribute to these different results.

While all the other LLMs passed all three exams, ChatGPT-3.5 struggled in the Persian exam, failing to reach a passing score. This is interesting because an earlier version, ChatGPT-3, did nearly as well on ophthalmology exams without special training or assistance¹⁶.

Another study reported similar results. ChatGPT-3.5 scored below the 58% passing mark for the 2022 Specialty Certificate Examination in neurology, while GPT-4 excelled, achieving the highest accuracy and exceeding the threshold²⁹.

To date, not many studies have been conducted to evaluate the performance of LLM in Persian medical exams. In a study in which the medical pre-internship exam in Persian was conducted using ChatGPT-3.5, a low accuracy rate of less than 40% was observed. This result is consistent with the failing score of that language model in the Persian language pre-internship exam in the current research³⁰.

This suggests a potential limitation of the LLM in comprehending Persian, which appears to have been addressed in the latest version. For instance, Khorshidi et al. (2023) found that GPT-4 successfully answered over 81% of questions in the Persian Iranian Medical Residency Entrance Exam 2023, which showed better performance than the current study. The difference in the complexity and content of the test between the residency entrance test and the pre-internship exam and the difference in the method can partly explain the improved results observed in the study³¹. In another study, Google Bard also successfully achieved a score above 60% in the Persian neurophysiology exam, which showed its stable performance in the Persian version of the exams in the current study³².

Studies indicate that languages such as Chinese, German, and Spanish pose challenges for LLMs³³. This is consistent with the more prominent finding that LLMs perform better in well-resourced languages such as English compared to languages with less available resources³⁴. Previous studies on LLMs in medical exams for non-English languages support this pattern. GPT-3.5 earned only 46.8% on a Korean general surgery test, whereas GPT-4 scored a substantially higher 76.4%³⁵. Similarly, GPT-4 aced the French version of the European Board of Ophthalmology exam with a 91% success rate³⁶. Even though direct comparisons of other studies with our Persian language exam are difficult due to variations in exam topics and complexity, the lower scores in Persian reflect a possible decrease in performance for LLMs in languages with fewer resources.

Interestingly, our study found that increasing the temperature for ChatGPT generally lowered its score on the exam. These changes, however, were not significant in the English test, except for ChatGPT-3.5. It appears that increasing the temperature of ChatGPT in medical tests has disrupted the consistency of the results and led to the uncontrollability of the LLM.

Few studies, as far as the researchers are aware, have looked at ChatGPT's performance in medical tests at various temperatures. One such research demonstrated no noticeable differences in performance at different temperatures. However, compared to the current research, that study's scope was limited since it only examined two temperature settings (0 and 1)³⁷.

Due to Google policy restrictions, the current study revealed that many Google Bard responses were categorized as "No Answer" responses. Google is well-known for following some strict regulations³⁸. It is worth noting that the reason for the higher percentage of "No Answer" responses to the questions in the gynecology field can be due to the stricter restrictions in the field of gynecology. It is interesting to note that the significant reduction of "No Answer" responses by simply repeating the questions can indicate the lack of predictive behavior of the models. In contrast, ChatGPT, especially version 3.5, had trouble generating legible and comprehensible responses. Previous studies have also reported similar issues with LLMs³⁹. These findings highlight the continued need to improve LLM's ability to produce coherent, readable, and predictable outputs.

The partly superior performance of LLMs (especially ChatGPT4) implies that LLMs can be utilized as supplementary educational tools, providing students with instant access to information, explanations, and practice questions that can enhance their understanding of complex medical concepts. By integrating LLMs into the curriculum, educators can create interactive learning environments where students engage with AI to reinforce their knowledge and improve their critical thinking skills. LLMs can also assist in formative assessments, offering personalized feedback on student performance and identifying areas where additional study is needed. Moreover, the performance of LLMs can serve as a benchmark for curriculum development, prompting educators to identify gaps in student knowledge and adjust teaching methods accordingly.

Although the use of LLMs can result in spectacular results, as demonstrated by our assessment of their performance compared with student performance, their use can be associated with certain pitfalls. Below is a list of strategies to reduce these risks:

- Rigorous validation: AI models in medicine require continuous validation to maintain accuracy and reliability and ensure patient safety.
- Transparency: Clear insight into AI decision-making processes is essential to building trust and identifying and correcting biases or errors.
- Ethical guidelines: It is essential to establish explicit ethical guidelines and privacy policies that protect patient privacy.

- Human oversight: Healthcare providers should evaluate AI offerings. Due to their unpredictable performance, artificial intelligence models need to complement human judgment and currently cannot replace it.
- Education: Medical education should include instructions on artificial intelligence tools to lead to their optimal use.
- Reducing bias: To ensure that artificial intelligence can understand all types of patients, it needs to be trained with different sources and languages and understand racial and regional differences.

Limitations

The present study had some limitations. First, many exams were not included, and only three exams, each with two versions of English and Persian, were examined in this study. Second, the pre-internship exam was the primary focus of this study, which may have subject-specific bias and not adequately reflect medical knowledge in the real world. Lastly, the study was monocentric, meaning that every participant studies at the same university, which would have limited how broadly the results could be applied.

Conclusion

This study demonstrated that GPT-4 consistently outperformed other LLMs and the average exam score of the students, achieving optimal performance across the board. Based on these results, GPT-4 may prove to be a useful tool in the field of medicine, if its limitations are carefully considered. Integrating LLMs into the curriculum has the potential to foster interactive learning and provide personalized feedback for assessments, though further research is needed to fully understand these benefits and their impact on teaching methods.

Data availability

The data collected and used in this study is available upon reasonable requests from the corresponding author.

Received: 10 July 2024; Accepted: 7 November 2024

Published online: 18 November 2024

References

1. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.
2. Briganti, G. & Le Moine, O. Artificial intelligence in medicine: today and tomorrow. *Front. Med.* **7**, 509744 (2020).
3. Savery, M., Abacha, A. B. & Gayen, S., Demner-Fushman, D. Question-driven summarization of answers to consumer health questions. *Sci. Data.* **7**, 322 (2020).
4. Schulman, J. et al. Chatgpt: optimizing language models for dialogue. *OpenAI blog*, **2**, 4 (2022).
5. Logé, C. et al. Q-Pain: a question answering dataset to measure social bias in pain management. *arXiv preprint arXiv:2108.01764* (2021).
6. Kumari, A. et al. Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing. *Cureus* **15**, e43861 (2023).
7. Floridi, L. & Chiriatti, M. GPT-3: its nature, scope, limits, and consequences. *Mind. Mach.* **30**, 681–694 (2020).
8. Korngiebel, D. M. & Mooney, S. D. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit. Med.* **4**, 93 (2021).
9. Bowman, E. *A new AI chatbot might do your homework for you. But it's still not an A+ student*, (2022). <https://www.npr.org/2022/12/19/1143912956/chatgpt-ai-chatbot-homework-academia>
10. Crothers, E., Japkowicz, N., Viktor, H. & Branco, P. in *2022 International Joint Conference on Neural Networks (IJCNN)*. 1–8 (IEEE).
11. Chen, P. H. C., Liu, Y. & Peng, L. How to develop machine learning models for healthcare. *Nat. Mater.* **18**, 410–414 (2019).
12. Gilson, A. et al. How Well Does ChatGPT Do When Taking the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *medRxiv*, 2012. 2023.22283901 (2022). (2022).
13. Davis, J., Van Bulck, L., Durieux, B. N. & Lindvall, C. The temperature feature of ChatGPT: modifying creativity for clinical research. *Jmir Hum. Factors.* **11**, e53559. <https://doi.org/10.2196/53559> (2024).
14. O'Leary, D. E. An analysis of Watson vs. BARD vs. ChatGPT: The Jeopardy! Challenge. *AI Magazine* **44**, 282–295 (2023).
15. Waisberg, E. et al. Google's AI chatbot Bard: a side-by-side comparison with ChatGPT and its utilization in ophthalmology. *Eye*, **38**, 1–4 (2023).
16. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health.* **2**, e0000198 (2023).
17. Johnson, D. et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-gpt model. (2023).
18. Patil, N. S., Huang, R. S., van der Pol, C. B. & Larocque, N. Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. *Can. Assoc. Radiol. J.*, **75**, 344–350, (2023).
19. Sallam, M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. <https://doi.org/10.3390/healthcare11060887> (2023).
20. Cascella, M., Montomoli, J., Bellini, V. & Bignami, E. Evaluating the feasibility of ChatGPT in Healthcare: an analysis of multiple clinical and research scenarios. *J. Med. Syst.* <https://doi.org/10.1007/s10916-023-01925-4> (2023).
21. Nori, H., King, N. S. P., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems <https://doi.org/10.48550/arxiv.2303.13375> (2023).
22. Oh, N., Choi, G. S. & Lee, W. Y. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Annals Surg. Treat. Res.* <https://doi.org/10.4174/astr.2023.104.5.269> (2023).
23. Ali, R. et al. Performance of ChatGPT and GPT-4 on Neurosurgery written board examinations. *Neurosurgery*. <https://doi.org/10.1227/neu.000000000002632> (2023).
24. Vachatanont, S. & Kingpetch, K. Exploring the capabilities and limitations of large language models in nuclear medicine knowledge with primary focus on GPT-3.5, GPT-4 and Google Bard. *J. Med. Artif. Intell.* **7**, 5 (2024).
25. Farhat, F., Chaudry, B. M., Nadeem, M., Sohail, S. S. & Madsen, D. Evaluating AI models for the national pre-medical exam in India: a head-to-head analysis of ChatGPT-3.5, GPT-4 and Bard. *JMIR Preprints* **10**, e51523 (2023).
26. Huang, R. S. et al. Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. *JMIR Med. Educ.* **9**, e50514 (2023).

27. Abbas, A., Rehman, M. S. & Rehman, S. S. Comparing the performance of Popular large Language models on the National Board of Medical Examiners Sample Questions. *Cureus* **16**, e55991 (2024).
28. Bowman, E. A new AI chatbot might do your homework for you. But it's still not an A+ student. *NPR* (2022).
29. Giannos, P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. *BMJ Neurol. open.* **5**, e000451 (2023).
30. Keshkar, A. et al. ChatGPT's Performance on Iran's Medical Licensing Exams. e000451 (2023).
31. Khorshidi, H. et al. Application of ChatGPT in multilingual medical education: how does ChatGPT fare in 2023's Iranian residency entrance examination. *Inf. Med. Unlocked.* **41**, 101314 (2023).
32. Shojae-Mend, H., Mohebbati, R., Amiri, M. & Atarodi, A. Evaluating the strengths and weaknesses of large language models in answering neurophysiology questions. *Sci. Rep.* **14**, 10785 (2024).
33. White, A. D. et al. *Do Large Lang. Models Know Chemistry?* doi:<https://doi.org/10.26434/chemrxiv-2022-3md3n> (2022).
34. Wang, Q. et al. Incorporating specific knowledge into end-to-end Task-oriented dialogue systems. <https://doi.org/10.1109/ijcnn52387.2021.9533938> (2021).
35. Oh, N., Choi, G. S. & Lee, W. Y. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Annals Surg. Treat. Res.* **104**, 269 (2023).
36. Panthier, C. & Gatinel, D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: a novel approach to medical knowledge assessment. *J. Fr. Ophthalmol.* **46**, 706–711 (2023).
37. Rosoł, M., Gąsior, J. S., Łaba, J., Korzeniewski, K. & Młyńczak, M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical final examination. *Sci. Rep.* **13**, 20512 (2023).
38. Urman, A. & Makhortykh, M. The silence of the LLMs: cross-lingual analysis of Political Bias and false information prevalence in ChatGPT, Google Bard, and Bing Chat. doi: (2023). <https://doi.org/10.31219/osf.io/q9v8f>
39. Shoufan, A. Can students without prior knowledge use ChatGPT to answer test questions? An empirical study. *ACM Trans. Comput. Educ.* **23**, 1–29 (2023).

Acknowledgements

The authors would like to acknowledge the grant provided by the Research Deputy of Shiraz University of Medical Sciences.

Author contributions

SZ contributed through ideation and methodology development; SV carried out the tests, collected the data, and wrote the main manuscript; MA composed and revised the methodology; KF edited the manuscript; MV revised the discussion; AG analysed the data; All authors reviewed and approved the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024