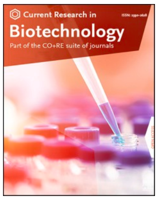
Current Research in Biotechnology 7 (2024) 100194

Contents lists available at ScienceDirect

Current Research in Biotechnology

journal homepage: www.elsevier.com/locate/crbiot



ChatGPT's scorecard after the performance in a series of tests conducted at the multi-country level: A pattern of responses of generative artificial intelligence or large language models

Manojit Bhattacharya [a,1], Soumen Pal [b], Srijan Chatterjee [c], Abdulrahman Alshammari [d], Thamer H. Albekairi [d], Supriya Jagga [e], Elijah Ige Ohimain [f], Hatem Zayed [g], Siddappa N. Byrareddy [h], Sang-Soo Lee [c], Zhi-Hong Wen [i], Govindasamy Agoramoorthy [j], Prosun Bhattacharya [k], Chiranjib Chakraborty [l,*,1]

[a] *Department of Zoology, Fakir Mohan University, Vyasa Vihar, Balasore 756020, Odisha, India*
[b] *School of Mechanical Engineering, Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India*
[c] *Institute for Skeletal Aging & Orthopaedic Surgery, Hallym University-Chuncheon Sacred Heart Hospital, Chuncheon-si 24252, Gangwon-do, Republic of Korea*
[d] *Department of Pharmacology and Toxicology, College of Pharmacy, King Saud University, Post Box 2455, Riyadh 11451, Saudi Arabia*
[e] *Division of Endocrinology, Brigham and Women's Hospital, Harvard Medical School, 25 Shattuck St, Boston, MA 02115, United States*
[f] *Microbiology Department, Niger Delta University, Wilberforce Island, Bayelsa State, Nigeria*
[g] *Department of Biomedical Sciences, College of Health and Sciences, Qatar University, QU Health, Doha, Qatar*
[h] *Department of Pharmacology and Experimental Neuroscience, University of Nebraska Medical Center, Omaha, NE 68198, United States*
[i] *Department of Marine Biotechnology and Resources, National Sun Yat-sen University, Kaohsiung 80424, Taiwan*
[j] *College of Pharmacy and Health Care, Tajen University, Yanpu, Pingtung 907, Taiwan*
[k] *Department of Sustainable Development, Environmental Science and Engineering, KTH Royal Institute of Technology, Teknikringen 10B, SE 10044 Stockholm, Sweden*
[l] *Department of Biotechnology, School of Life Science and Biotechnology, Adamas University, Kolkata, West Bengal 700126, India*

ARTICLE INFO

Keywords:
ChatGPT
Accuracy
Reproducibility
Plagiarism
Answer length

ABSTRACT

Recently, researchers have shown concern about the ChatGPT-derived answers. Here, we conducted a series of tests using ChatGPT by individual researcher at multi-country level to understand the pattern of its answer accuracy, reproducibility, answer length, plagiarism, and in-depth using two questionnaires (the first set with 15 MCQs and the second 15 KBQ). Among 15 MCQ-generated answers, 13 ± 70 were correct (Median : 82.5; Coefficient variance : 4.85), 3 ± 0.77 were incorrect (Median: 3, Coefficient variance: 25.81), and 1 to 10 were reproducible, and 11 to 15 were not. Among 15 KBQ, the length of each question (in words) is about 294.5 ± 97.60 (mean range varies from 138.7 to 438.09), and the mean similarity index (in words) is about 29.53 ± 11.40 (Coefficient variance: 38.62) for each question. The statistical models were also developed using analyzed parameters of answers. The study shows a pattern of ChatGPT-derive answers with correctness and incorrectness and urges for an error-free, next-generation LLM to avoid users' misguidance.

Introduction

Generative artificial intelligence (AI) provides an effective platform to the user. OpenAI (San Francisco, USA) recently developed ChatGPT, a generative AI-based chatbot powered by large language models (LLMs). This chatbot is gaining interest daily because it can respond to text queries. It has been noted that over one million users were using it within five days after its release (Chakraborty et al., 2023a; De Angelis et al., 2023). An LLM is a neural network trained with extensive human-generated text. It can predict the subsequent word (token) and provide a sequence of words (Shanahan et al., 2023). LLMs have recently incited considerable interest across the industrial domains and academics. However, it has been noted that all the recent AI-based chatbots use LLMs or multi-modal LLMs. Due to the response to text queries, scientists use chatbots as their research assistants for tasks like summarizing and organizing research literature, writing code, etc., in different science and technologies and medical research domains (Ali et al., 2023; Chakraborty et al., 2023b; Chakraborty et al., 2023a; Chatterjee et al.,

[*] Corresponding author.
[1] Authors contributed equally.

https://doi.org/10.1016/j.crbiot.2024.100194
Received 27 September 2023; Received in revised form 21 February 2024; Accepted 1 March 2024
Available online 2 March 2024

Table 1
Different MCQ questions were asked to ChatGPT to evaluate its performance.

Q. NO.	Questions	Type of Knowledge Based Questions
Q. No.1	Which of the following tumors stay confined to one region? a. Malignant tumor b. Benign tumor c. Metastatic tumor d. Solid tumor	Remembering
Q. No.2.	Which is the most common malignancy affecting men in the UK? a. Lung cancer b. Prostate cancer c. Renal cancer d. Bladder cancer	Remembering
Q. No.3.	Hereditary nonpolyposis colorectal cancer is also known as_________. a. Lynch syndrome b. Burkitt lymphoma c. Li-Fraumeni syndrome d. Cowden syndrome	Remembering
Q. No.4.	Intraductal papillary mucinous neoplasm (IPMN) is a common diagnostic method for which kind of cancer? a. Prostate cancer b. Renal cancer c. Breast cancer d. Pancreatic cancer	Remembering
Q. No.5.	Name the procedure which is conducted to examine the bladder more completely under anesthesia and remove tumors. a. Transurethral resection of a bladder tumor (TURBT) b. Intravesical Therapy c. Bladder Preservation Therapy d. Chemotherapy	Remembering
Q. No.6.	The rous sarcoma virus, the first tumor-inducing virus, contains four genes namely gag, pol, env, and v-src. What is the function of pol gene? a. Encodes the capsid protein of the virus b. Encodes the reverse transcriptase c. Encodes a viral envelope protein d. Encodes a protein kinase that inserts into the plasma membranes of infected cells	Remembering
Q. No.7.	Which of the following is a tumor suppressor gene? a. TP53 gene b. Myc gene c. Ras gene d. All of the above	Remembering
Q. No.8.	Which of the following is a non-surgical method used for removing the tumor? a. Metastasectomy b. Radical nephrectomy c. Both (a) and (b) d. None of the above	Remembering
Q. No.9.	Which of the following disease is not considered as cancer, but considered a form of precancer? a. Kaposi sarcoma b. Actin keratosis c. Melanoma d. None of the above	Remembering
Q. No.10.	Which of the following disease increase the risk of liver cancer? a. Porphyria cutaneatarda b. Wilson disease c. Tyrosinemia d. All of the above	Remembering
Q. No.11.	Name the FDA approved drug for merkel cell carcinoma. a. Lynparza plus abiraterone b. Padcev plus Keytruda	Analytical

(*continued on next page*)

Table 1 (*continued*)

Q. NO.	Questions	Type of Knowledge Based Questions
	c. Zynyz d. None of these	
Q. No.12.	The FDA has cleared an investigational new drug (IND) application for the bispecific autologous chimeric antigen receptor (CAR) T-cell therapy. Name the drug. b. Rylaze c. Aduhelm d. Zynyz e. IMPT-314	Analytical
Q. No.13.	A breast cancer patient is palbociclib resistant due to an increase in the level of the ABCB1 protein. At the same time, it was found that an alteration in ESR1 promoted in the same breast cancer patient causes resistance to alpelisib. Again, in that patient, the Akt induced docetaxel resistance was noted. Which therapeutic molecules should be included in the treatment plan in order to treat the breast cancer? a. Alpelisib b. Exemestane c. Docetaxel d. Palbociclib	Analytical
Q. No.14.	A child has Acute Lymphoblastic Leukemia (ALL). Cytarabine shows complete remission and relapse-free five years survival of 40–70 % in ALL. At the same time, it was noted that imatinib therapy in ALL considarated as a relapse-free survival rate at five years was 80–85 %. Similarly, at five years, daunorubicin's relapse-free survival rate was 55 % in ALL. Likewise, pegaspargase showed 90–94 % at year 5 in ALL. Which drug should be included in the treatment plan? a. Cytarabine b. Matinib c. Daunorubicin d. Pegaspargase	Analytical
Q. No.15.	The patient suffering from melanoma has shown resistance to Dacarbazine, Braftovi, Nivolumab, and Aldesleukin. The doctor administered a combination of Nivolumab and Relatlimab as a treatment cocktail, but no significant improvements were observed, indicating the patient's resistance to this particular combination. At this position which molecule should be incorporated in the treatment plan? a. Nivolumab and Relatlimab b. Nivolumab c. Relatlimab d. Aldesleukin	Analytical

2023; Hutson, 2022; Pal et al., 2023a; Patel and Lam, 2023).

ChatGPT is a sophisticated language model with numerous benefits and applications in the healthcare and medical industries. It can help healthcare practitioners with research, diagnosis, patient monitoring, and medical education, and other things (Dave et al., 2023). Establishing a strong connection with patients is valuable across various healthcare domains, but it may sometimes be essential for achieving optimal therapeutic results. ChatGPT can supplement human healthcare providers' care and improve patient outcomes by enhancing treatment adherence and delivering more convenient healthcare services (Homolak, 2023; Mbakwe et al., 2023). The practical applications of ChatGPT extend to improving patient care and treatment results by offering medical knowledge and enabling communication between healthcare providers and patients. In an academic context, ChatGPT can contribute to advancing knowledge, uncovering fresh research inquiries, and enhancing the accuracy of data analysis and interpretation. Consequently, when integrating ChatGPT-based interventions, researchers and healthcare providers need to consider these factors carefully (Ruksakulpiwat et al., 2023). When utilized in healthcare, ChatGPT's prediction skills remain be constrained. A transformer model, such as ChatGPT, learns patterns from training data and then utilizes that information to make predictions. Transformer models may sometimes produce erroneous predictions in medical applications because they are incentivized to discover patterns and make predictions. Their access to licensed healthcare experts, such as diabetes educators, may be problematic for people living in underserved or rural locations. They could, however, use ChatGPT as a trusted source of advice and expertise if they are unable to visit a physical healthcare center (Iftikhar, 2023; Mann, 2023). ChatGPT has demonstrated its significance to the advancement of the medical sphere in recent years. By offering rigorous and exact data analysis, it has contributed to the evolution of translational medicine and medication development. Furthermore, it has improved medical reporting, diagnostics, and treatment plans, ultimately boosting overall medical practice and patient experience. The model's performance in providing fundamental support in research and clinical settings has been impressive. It has enormous potential to alter the field of medicine and healthcare if more technological advancements are made in partnership with the medical industry (Ruksakulpiwat et al., 2023). ChatGPT can be helpful in medical education, research, and clinical management, but it can only partially replace human expertise and comprehension due to AI limitations. Nonetheless, rapid advances in information technology, machine learning, and AI are causing significant shifts in how we approach medical education and clinical administration. These advancements are taking place at such a rapid pace that they are predicted to revolutionize the field quickly (Khan et al., 2023).

However, several studies raised questions on ChatGPT about plagiarism, sources of biases, and the accuracy of responses. Alser and Waisberg have shown their concern about plagiarism and sources of biases (Alser and Waisberg, 2023). Bhattacharyya et al. experimented using ChatGPT to evaluate the accuracy of ChatGPT-produced references. They found inaccurate references in the ChatGPT-produced answer. They found 64 % wrong page numbers, 64 % incorrect volume, and 60 % inaccurate years of publication. They concluded that these three components (wrong page numbers, wrong volume, and inaccurate year of publication) were the most frequent errors (Bhattacharjee et al., 2023). Weng et al. conducted a test of ChatGPT using the questions of Taiwan's family medicine board exam, and they found that the accuracy rate of answering the question was not reasonable (Wang et al., 2023). Therefore, it is crucial to understand more details about the answering the pattern of ChatGPT in a test, such as accuracy, reproducibility, length of the answer, and plagiarism in answering the questions.

In this study, we evaluate ChatGPT's capability to answer during a test regarding the answer's accuracy, reproducibility, answer length, and similarity index (plagiarism). Here, we provided two categories of questions to understand ChatGPT's capability. First, we provided the multiple choice questions (MCQ) to understand the accuracy and reproducibility of the questions. Secondly, we provided broad and knowledge-based questions to understand the answer's length and plagiarism during the answering of questions. We quantitatively evaluated all parameters (accuracy, reproducibility, answer length, and similarity plagiarism). Similarly, we evaluate one qualitative parameter of ChatGPT's response while answering the questions, i.e., overall insights into answering all the knowledge-based questions.

Thus, the aim of the current research is to provide both quantitative and qualitative performance of an AI based bot during a test. Finally, our study provided results that help in gap analysis for ChatGPT's answer during a test. At the same time, the study will initiate a fundamental basis of discussion to develop a high-quality, next-generation ChatGPT or LLM model.

Table 2
Different knowledge-based questions were asked of ChatGPT to evaluate its performance.

Q. NO.	Questions	Type of Knowledge Based Questions
1.	Is cancer a genetic disease?	Remembering
2.	What are the early markers of hepatocellular carcinoma?	Remembering
3.	What is the worldwide statistics of cancer incidence, mortality and cured cases in the year 2022?	Quantitative
4.	What are the probable causes of cancer?	Descriptive
5.	Categorize the different types of cancer on the basis of the type of affected tissue.	Remembering
6.	Explain the process of metastasis.	Descriptive
7.	Migrating cancer cells can die from a variety of causes. Is this statement true or false? Explain.	Analytical
8.	What will be the probable death rate of colorectal cancer in this year?	Quantitative
9.	List down the probable treatments for cancer.	Remembering
10.	Can you suggest a treatment method for chronic myelogenous leukemia?	Analytical
11.	Explain the merits and demerits of chemotherapy.	Descriptive
12.	What is the mortality rate of prostate cancer in the last five years?	Quantitative
13.	How is systemic immunity linked with cancer?	Descriptive
14.	Which type of cancer occurs in the bone marrow and results in the formation of blood cells?	Remembering
15.	Two persons are having lung cancer which is diagnosed in the advanced stage. One of them is treated with Paclitaxel and the other is treated with Vincristine. Who has the greater chance of recovery?	Analytical

Method

We conducted an experiment utilizing ChatGPT, using two sets of questionnaires to understand its ability to perform in a test. We evaluated ChatGPT-generated answers.

The questionnaire, question type, and difficulty level

The experiment involved administering two questionnaire sets, each comprising 15 cancer-related questions. The first set of questionnaires contained 15 MCQ (n = 15) (Table 1).

Here, we used some remembering and analytical questions. The second set of questionnaires comprised 15 knowledge-based questions (KBQ) (n = 15) (Table 2).

We asked ChatGPT different questions about cancer-related topics. There is no such overlap between questions on particular topics in the sets of "MCQs" (Table 1) and "KBQs" (Table 2).

Here, we have used several knowledge-based questions, such as analytical, remembering, quantitative, and descriptive. Here, we used

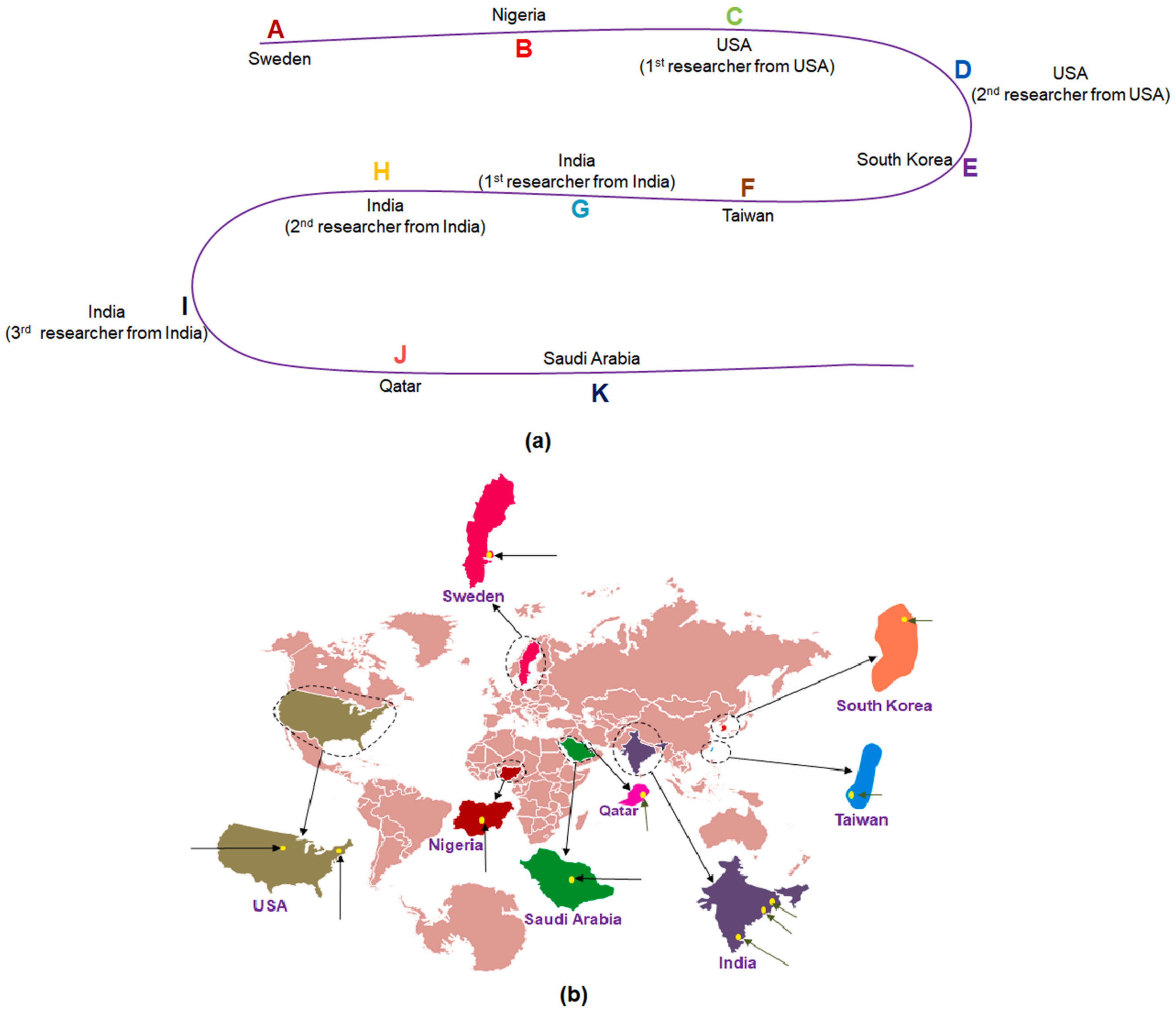


Fig. 1. Different countries with the ChatGPT test and their geographical locations worldwide. (a) Different countries were considered for the ChatGPT test using two sets of questionnaires (one MCQ set of questionnaires with 15 questions and one KBQ Set of questionnaires with 15 questions) in different time schedules within a month. (b) A world map comprised of the distribution of participatory researchers for the ChatGPT test. The figure informed us that participatory researchers are located in all geographic locations worldwide.

the “KBQs” as per the bloom taxonomy. The specificity of the test questions arises from a meticulous process of development. Attention has been directed toward subjective inquiries about cancer and objective queries regarding diseases to evaluate ChatGPT’s capabilities comprehensively. A systematic selection process has been employed to ensure representativeness, considering diverse content domains. Incorporating case studies has further enriched the evaluation, allowing ChatGPT to analyze and respond to real-world scenarios. The test questions have been selectively crafted through a two-step process. An extensive literature review was conducted to identify crucial themes in medicine, focusing on cancer and various diseases. Subsequently, questions were gathered from reputable sources encompassing healthcare professionals and AI tools collaborating to design inquiries encompassing both subjective aspects of cancer and objective investigations regarding diseases. This multidisciplinary approach guarantees that the questions represent the desired content domains and align with real-world medical challenges. At the same time, we have used Bloom’s taxonomy to frame the different questions with different knowledge levels (Cheng et al., 2021; Stringer et al., 2021). By incorporating this diverse set of questions that covers both subjective and objective aspects across medical domains, our objective was to create a robust assessment, thereby enhancing the depth and relevance of our study.

Data acquisition from different countries

The questionnaire sets were distributed to representatives from different countries across the globe (Fig. 1a). We selected the countries from all over the world covering all continents across the globe. The countries are distributed in all geographical locations worldwide, such as Europe, Africa, the USA, East Asia, and the Middle East (Fig. 1b). However, we have randomly chosen the countries. In an experiment, Fergus et al. used two different user accounts to understand the pattern of ChatGTP answer in terms of identical answers (Fergus et al., 2023). Similarly, here, we use different countries to understand the ChatGPT answers’ pattern.

We did not use the different prompts for all the questions of two sets of questionaries (set-1: MCQs (Table 1) and set-2: KBQs (Table 2)). We used the same prompt and format of two sets of questionaries for

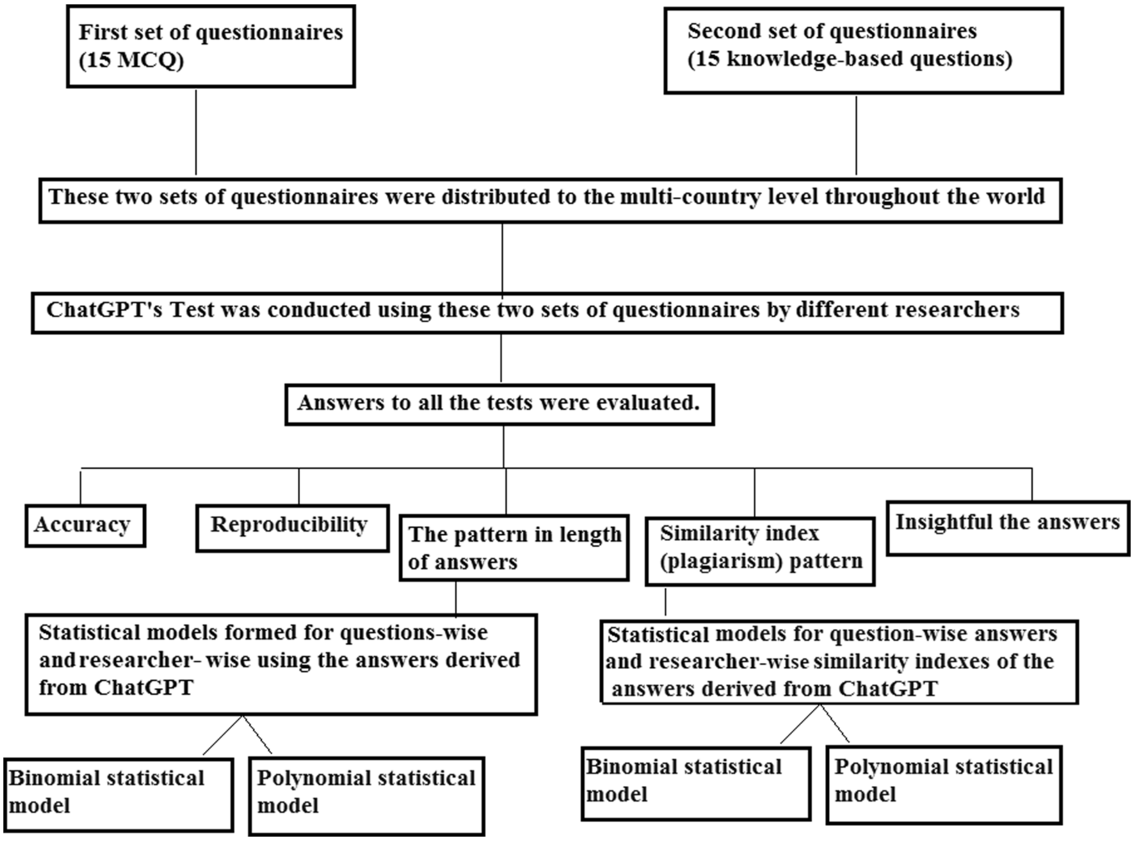

First set of questionnaires (15 MCQ)
Second set of questionnaires (15 knowledge-based questions)
These two sets of questionnaires were distributed to the multi-country level throughout the world
ChatGPT's Test was conducted using these two sets of questionnaires by different researchers
Answers to all the tests were evaluated.
Accuracy
Reproducibility
The pattern in length of answers
Similarity index (plagiarism) pattern
Insightful the answers
Statistical models formed for questions-wise and researcher- wise using the answers derived from ChatGPT
Statistical models for question-wise answers and researcher-wise similarity indexes of the answers derived from ChatGPT
Binomial statistical model
Polynomial statistical model
Binomial statistical model
Polynomial statistical model

Fig. 2. A schematic diagram presents the study's overall workflow, i.e., the ChatGPT's test and performance evaluation in the test.

different countries across the globe. No changes have been made in the two sets of questionnaires while asking the questions to ChatGPT, and, finally, its answers from all those countries were noted.

ChatGPT and test

All researchers used ChatGPT's (OpenAI; San Francisco, CA), free and open version (version 3.5), to perform the test. Researchers from different countries used the two sets of questions for the test. All questions from each set were asked to ChatGPT one after another. The researchers conducted the test at different times within a month (10th June to 10th July 2023). After the test of ChatGPT, the responses were collected in a particular format by representatives across the country, and they sent the responses to the corresponding author. Then the responses were accumulated, and further analysis were done. Statistical model was developed.

All representatives from different countries used the same questions in the ChatGPT. However, we framed the texts for effective communication with ChatGPT during question framing by the corresponding author. Now, the prepared questions were sent to representatives from different countries worldwide. After that, representatives from different countries kept the questions same. No alterations of questions were performed.

Analysis and statistical model

We collected the response and performed an exhaustive analysis: (i) We analyzed the accuracy of answering the questions and the reproducibility of ChatGPT using 15 MCQs. (ii) We analyzed the length and similarity index (plagiarism) of each answer obtained from 15 KBQs across different countries. The length of the question and the similarity index were measured in words. (iii) We evaluate the overall insights to answer all the KBQs. The first two points were analyzed quantitatively, and the third one was analyzed qualitatively. In the case of the third point, we followed the classroom assessment method of a teacher while evaluating the answer sheet (Anderson, 2023).

The similarity index of all answers was checked with the Turnitin software, and evaluated the similarity percentage (Halgamuge, 2017). Using the similarity index (plagiarism) and the answer's length, we have developed statistical (binomial and polynomial) models of two parameters such as, question-wise and country-wise answering patterns. All the statistical models were developed using an open-source R software package. We also used PAST statistical software to depict some statistical graphs and plots (Hammer, 2001). Finally, a schematic diagram has been portrayed, showing the overall workflow of our work (Fig. 2).

Result

Accuracy

One critical question among the researchers is: how accurate and reliable the artificial intelligence-strengthened responses are. The accuracy of ChatGPT while answering the MCQ of the same question should be consistent with choosing the same answer. Using the first set of questionnaires of 15 MCQ, we assess the accuracy of the ChatGPT-generated answers. Among 15 MCQ, the mean correctness of the ChatGPT-generated answers was 13 ± 70 (Median : 82.5; Coefficient variance : 4.85). Similarly, the mean incorrectness of the answers was 3 ± 0.77 (Median: 3, Coefficient variance: 25.81). The correctness and incorrectness of 15 MCQs were analyzed across various counteries. The highest number of correct answers was observed at Nigeria, with 13 correct answers. Conversely, Saudi Arabia, had the highest number of incorrect answers, with 5 incorrect responses (Fig. 3a). Among 165 MCQ, 132 are correct, and 33 are incorrect. We found the ChatGPT-

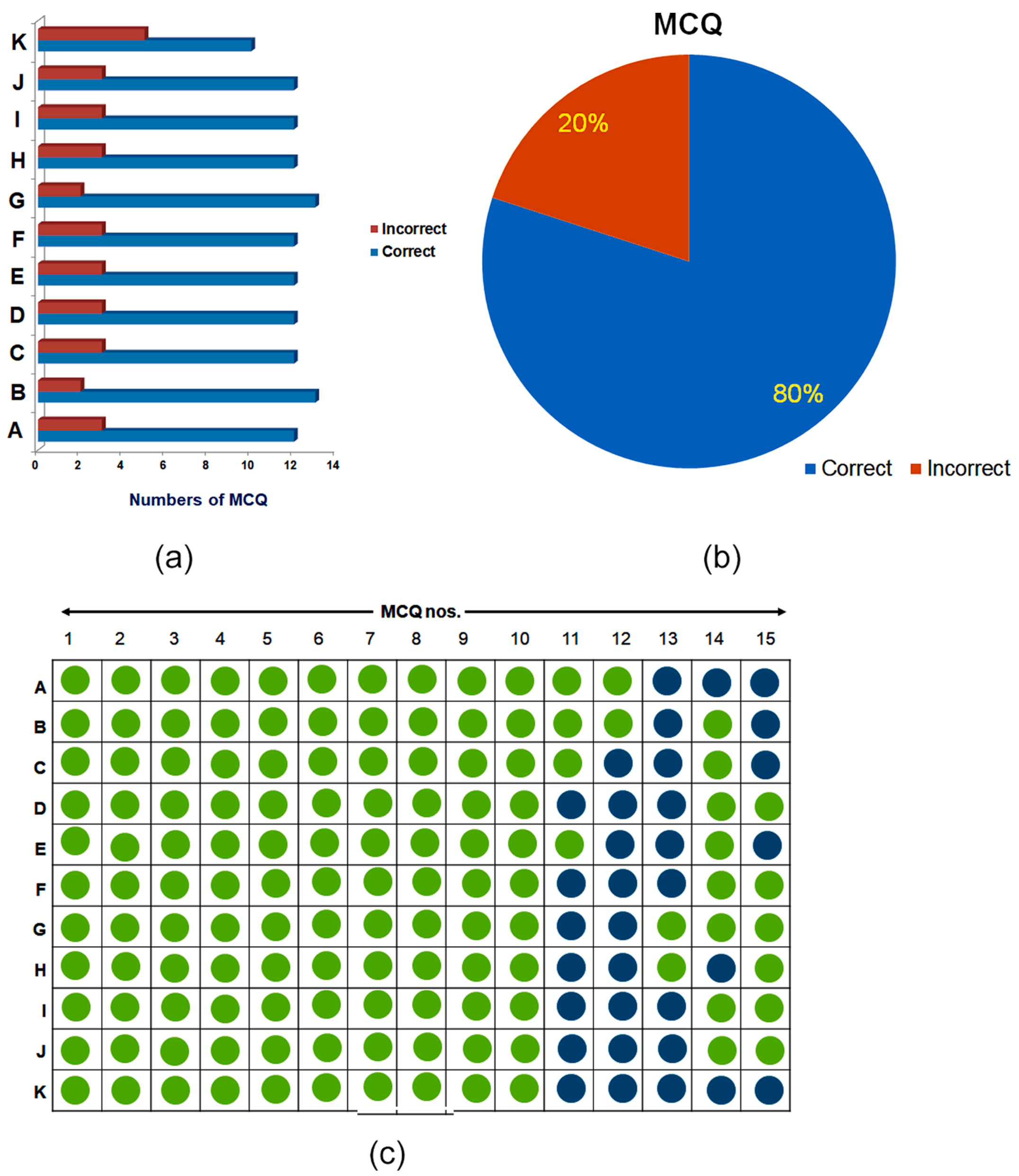


Fig. 3. ChatGPT's accuracy on the test with 15 MCQ. (a) The country-wise correct and incorrect answers of ChatGPT with the test in 15 MCQ. (b) The figure also depicts the percentage of correct and incorrect ChatGPT-derived answers in the test. (c) A diagram shows the ELISA plate template representation to show the reproducibility of ChatGPT-derived answers. It shows that the ChatGPT-generated answers to questions 1 to 10 in the MCQ are reproducible, and the answers to questions 11 to 15 in the MCQ are not.

generated answers were overall 80 % correct and 20 % were incorrect (Fig. 3b). The analysis period took one-and-half months (16th July to 31st August 2023).

Reproducibility

The scientific method is characterized by reproducibility and is considered one of the significant factors for today's research (Engineering, Medicine, 2019). Across the disciplines, the fundamental principle of scientific research is the independent verification of data. In general, it is essential for one researcher's capability to reproduce the findings using the scientific method for publication. Although, the researcher found reproducibility is a challenge in science (Baker, 2016). In this same line, one of the main questions of AI-generated results is reproducible or not, and researchers are arguing about the reproducibility of AI, which is a significant question today (Erik Gundersen, 2021). However, the reproducibility of ChatGPT while answering the same questions should be consistent. In this direction, we found that the ChatGPT-generated answers to questions 1 to 10 in the MCQ are reproducible. At the same time, we found ChatGPT-generated answers to questions 11 to 15 in the MCQ are not reproducible (Fig. 3c). In our MCQ questions, we used questions 1 to 10 as remembering questions, and questions 11 to 15 as analytical questions. It might be possible that AI does not show reproducibility among analytical questions.

The pattern in length of answers

The researchers are curious about the pattern in the length of the answers of ChatGPT. Here we tried to develop two types of answering patterns of ChatGPT in the length of answers: Question-wise answering pattern and country-wise answering pattern.

To evaluate the question-wise answering pattern, we calculated the

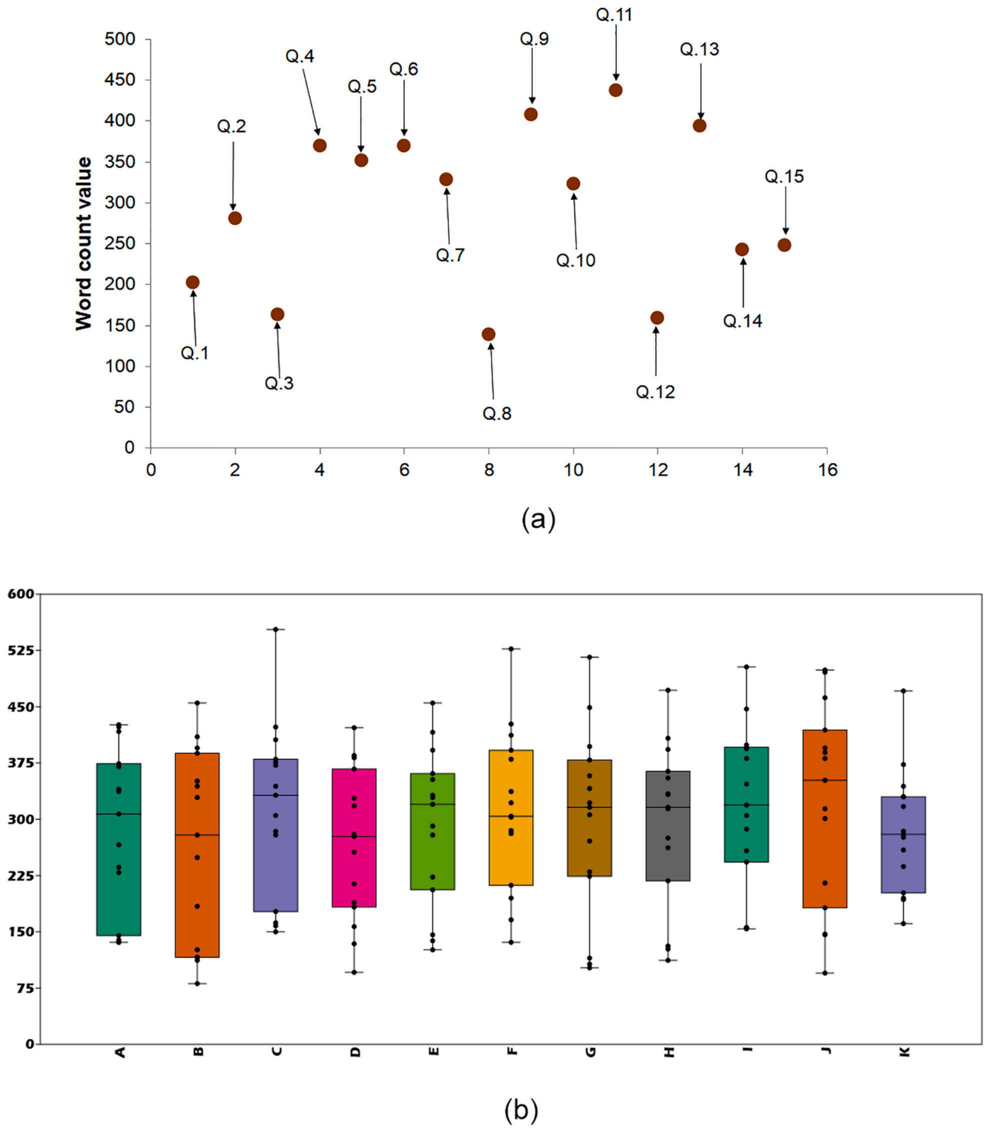


Fig. 4. ChatGPT's answer length on the test with 15 knowledge-based questions. (a) A scatter plot using the mean word count of all answers from different participatory researcher from several countries. The plot represented the mean word count of all 15 answers. Here we found the highest word count in the answer to question 11 and the lowest in the answer to question 8. (b) The country-wise answering pattern of the word count for all questions from ChatGPT has been represented through the Box-plot.

mean word count length of all answers to each question from ChatGPT from the researchers of different countries (n = 11). We found that the length of word count of each question is about 294.5 ± 97.60 (Coefficient variance: 33.14), which was obtained from the grand mean of 15 KBQs. The mean range for each of the 15 questions varies from 138.7 to 438.09 (Table S1). We developed a scatter plot using the mean of the word count of all 15 answers (Fig. 4a).

Similarly, we have tried to understand the country-wise answering pattern for all answers from ChatGPT. We found the broader range of word count patterns at India (G), which varies from 102 to 516 (Fig. 4b).

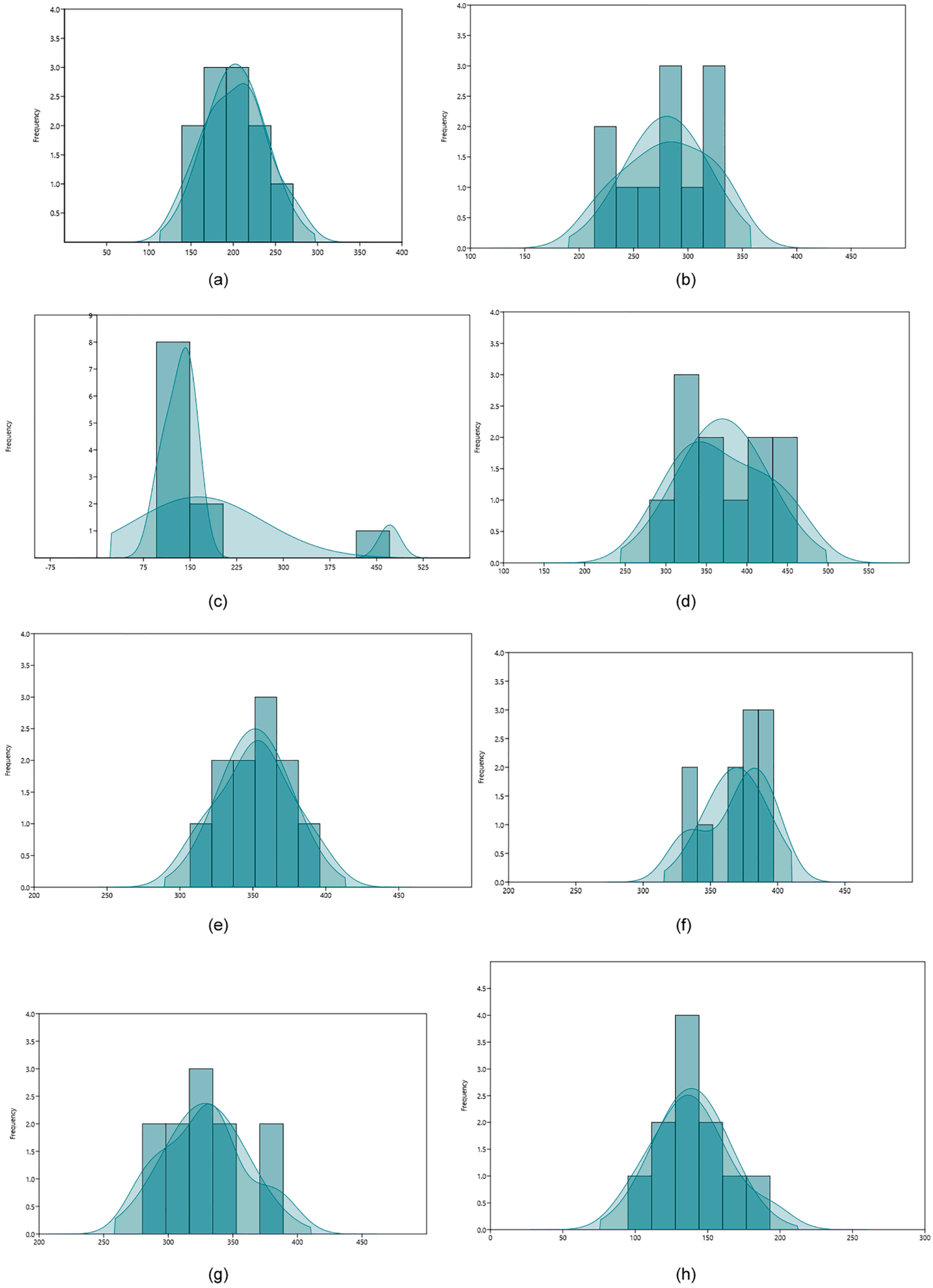
However, the minimum word count was found in Nigeria, in answer to question no 14, and the answer length is 81 words. At the same time, the maximum word count is found in USA (C), in the answer to question no 11, and the length of the answer is 553 words.

Similarly, we developed statistical models for questions and country-wise answers derived from ChatGPT. In both cases, we developed binomial and polynomial statistical models. The question-wise answers binomial models informed the word count pattern in each question's binominal distribution. These models informed us of the question-wise word count distribution pattern of ChatGPT-derived answers (Fig. 5a to Fig. 5o). At the same time, we developed a second-order polynomial plot of the word count of ChatGPT-derived answers of 15 questions (Fig. 6). Here, the second order polynomial equation indicates the R^2 value of 0.444.

Again, we developed the binomial models using the country-wise answers. These models informed country-wise word count distribution pattern of ChatGPT-derived answers (Fig. 7a to Fig. 7k). At the same time, we developed a second-order polynomial model for the country-wise word count of the answers (Fig. 8). Here, the second order polynomial equation indicates the R^2 value of 0.86.

Similarity index (plagiarism) pattern

Plagiarism is the act of using someone else's ideas, words, or work without giving them proper credit or obtaining permission and presenting it as one's own. It can involve various forms of intellectual property, including written or spoken words, ideas, images, and more (Habibzadeh, 2023).



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 5. A statistical model shows the question-wise binomial distribution pattern of the answer length. (a) Binomial distribution pattern answer length of question 1 (b) Binomial distribution pattern answer length of question 2 (c) Binomial distribution pattern answer length of question 3 (d) Binomial distribution pattern answer length of question 4 (e) Binomial distribution pattern answer length of question 5 (f) Binomial distribution pattern answer length of question 6 (g) Binomial distribution pattern answer length of question 7 (h) Binomial distribution pattern answer length of question 8 (i) Binomial distribution pattern answer length of question 9 (j) Binomial distribution pattern answer length of question 10 (k) Binomial distribution pattern answer length of question 11 (l) Binomial distribution pattern answer length of question 12 (m) Binomial distribution pattern answer length of question 13 (n) Binomial distribution pattern answer length of question 14 (o) Binomial distribution pattern answer length of question 15.

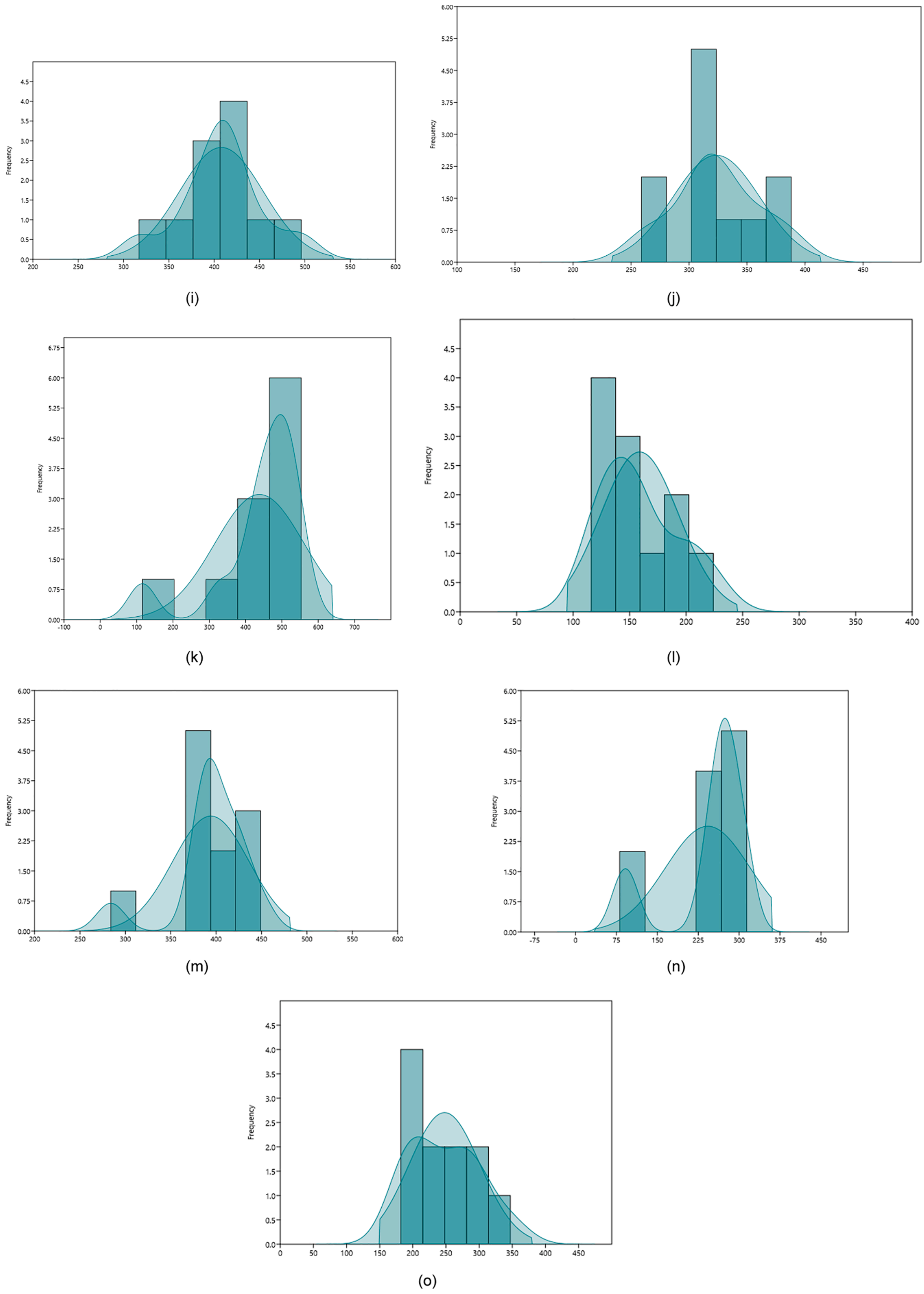


Fig. 5. (*continued*).

Presently, one significant concern about AI-generated answers is plagiarism. Several researchers have shown their concern about the plagiarism of ChatGPT-generated answers (Ventayen, 2023). Pal et al. also reported concern about ChatGPT plagiarism and urged for plagiarism-free ChatGPT answers (Pal et al., 2023b). But it is crucial to understand ChatGPT-generated answers' plagiarism pattern. Like word count, we attempted to assess the plagiarism pattern of ChatGPT-generated answers. We evaluated two patterns of plagiarism in those answers: question-wise plagiarism pattern and country-wise plagiarism pattern.

To evaluate the question-wise plagiarism pattern, we evaluated the mean similarity index of all answers to each question from ChatGPT

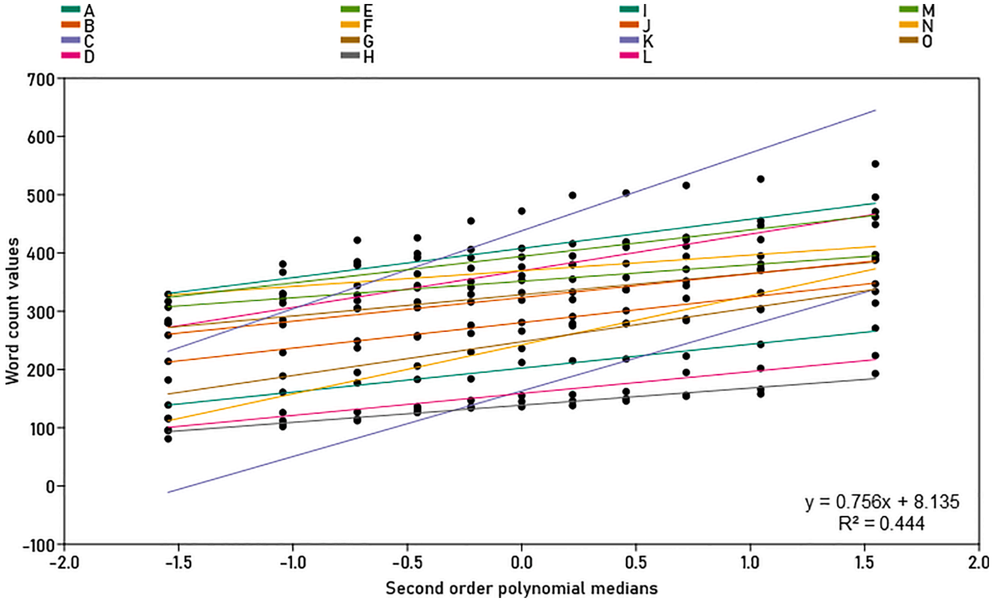


Fig. 6. A statistical model shows the question-wise, answer length's second-order polynomial statistical model. The second-order polynomial equation informs the $R^2 = 0.444$.

from the researchers of different countries (n = 11). We found the mean similarity index of each question is about 29.53 ± 11.40 (Coefficient variance: 38.62), and we evaluated it from the grand mean of the similarity index of 15 KBQs. The mean range for each of the 15 questions varies from 7.09 to 46.09 (Table S2). We developed a scatter plot using the mean of the similarity index of all 15 answers (Fig. 9a).

Correspondingly, the country-wise similarity index of answering patterns of all answers to each question was evaluated. We found a more comprehensive range of similarity index patterns in Saudi Arabia which varies from 7 % to 73 % (Fig. 9b). However, the minimum number of similarity indexes was observed in five counteries: USA (C) (Answer of Question 7), Taiwan (F) (Answer of Question 7), India (G) (Answer of Question 14), India (H) (Answer of Question 14), and Qatar (J) (Answer of Question 15). In all cases, the similarity index was noted to be 0 %. At the same time, the maximum similarity index was found in Saudi Arabia (K) in the answer to Question 9, which was 73 %.

Similarly, we developed statistical models for question-wise answers and country-wise similarity indexes of the answers derived from ChatGPT. In both cases, we developed binomial and polynomial statistical models. The question-wise answers binomial models informed the distribution pattern of the similarity index of the answer of each question. These models informed us of the question-wise word count distribution pattern of ChatGPT-derived answers (Fig. 10a to Fig. 10o). At the same time, we developed a second-order polynomial plot of the similarity index of ChatGPT-derived of all 15 answers (Fig. 11). Here, the second order polynomial equation indicates the R^2 value to be 0.78.

Again, we developed the binomial models using the country-wise similarity index of answers. These models informed country-wise the similarity index distribution pattern of ChatGPT-derived answers (Fig. 12a to Fig. 12k). At the same time, we developed a second-order polynomial model of the country-wise similarity index of answers (Fig. 13). Here, the second order polynomial equation indicates the R^2 value to be 0.46.

Insightful answers

Overall, we found that ChatGPT can provide insightful answers to the question and exhibits the critical capability of thinking skills. ChatGPT shows its capability and significant insight to view all the questions. However, the tool does not answer correctly in higher-order thinking questions. We included some higher-order thinking questions in the analytical questions of the MCQ, such as MCQ 13, MCQ 14, and MCQ 15. We found the ChatGPT is only correct sometimes while answering the questions (Table-1 and Fig. 3c). This AI model is trained up to September 2021. Therefore, ChatGPT-3.5 was unable to provide the information after September 2021. ChatGPT has provided some MCQ answers in the direction, such as MCQ 11 and MCQ 12.

Discussion

Recently, AI-generated responses, especially ChatGPT, have been able to provide different answers to students. Therefore, it is a valuable tool for human learners today. Several scientists illustrated that ChatGPT provides valuable medical information to students (Editorials, 2023). At the same time, several tests were being conducted using ChatGPT, including the law to medical entrance tests (Giannos and Delardas, 2023; Gilson et al., 2023). Although ChatGPT provides valuable medical information, our question is: how accurate the pieces of information were? In this direction, we design a critical test using two sets of questions, one with MCQ and another with knowledge-based questions. Here, the study provides evidence about ChatGPT's ability to perform during a test. The study model also evaluated the overall view of AI-driven tests. It will provide a novel insight into the answering pattern of the AI model. Our findings show the ChatGPT, an AI-driven system, comprises four major quantitative components in generating answers: (1) accuracy, (2) reproducibility, (3) similarity index, and (4) length of the answer. Along with these four quantitative components, we measured one qualitative parameter: the insights of ChatGPT answers. By comprehensively evaluating these factors, we aim to determine the future prospects of utilizing ChatGPT in various medical applications.

Among the ChatGPT-generated answers, we found that the overall correctness is 13 ± 70 (80 %). Several researchers have tried to conduct a test using ChatGPT and evaluate the correctness property of the tool. Kung et al. assessed the achievement of ChatGPT on the USMLE (United States Medical Licensing Exam). They found that ChatGPT can provide valid clinical insights and exhibited comprehensible reasoning aptitude in the test (Kung et al., 2023). Humar et al. used ChatGPT for the plastic surgery in-service exam and measured the performance with 1129 questions. They found that the tool provided 57 % correct answers to the questions. The study used all MCQ (Humar et al., 2023). Similarly,

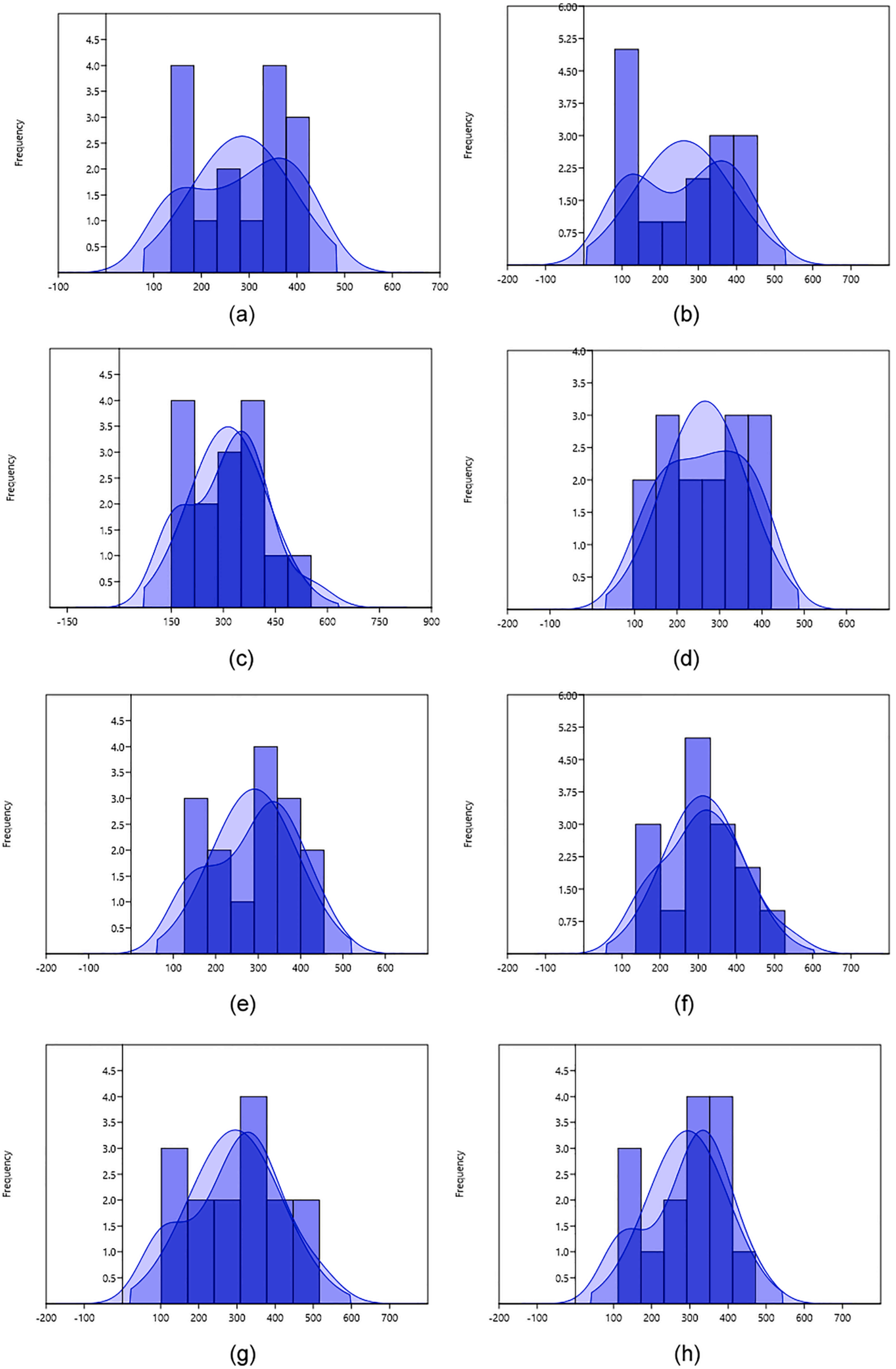
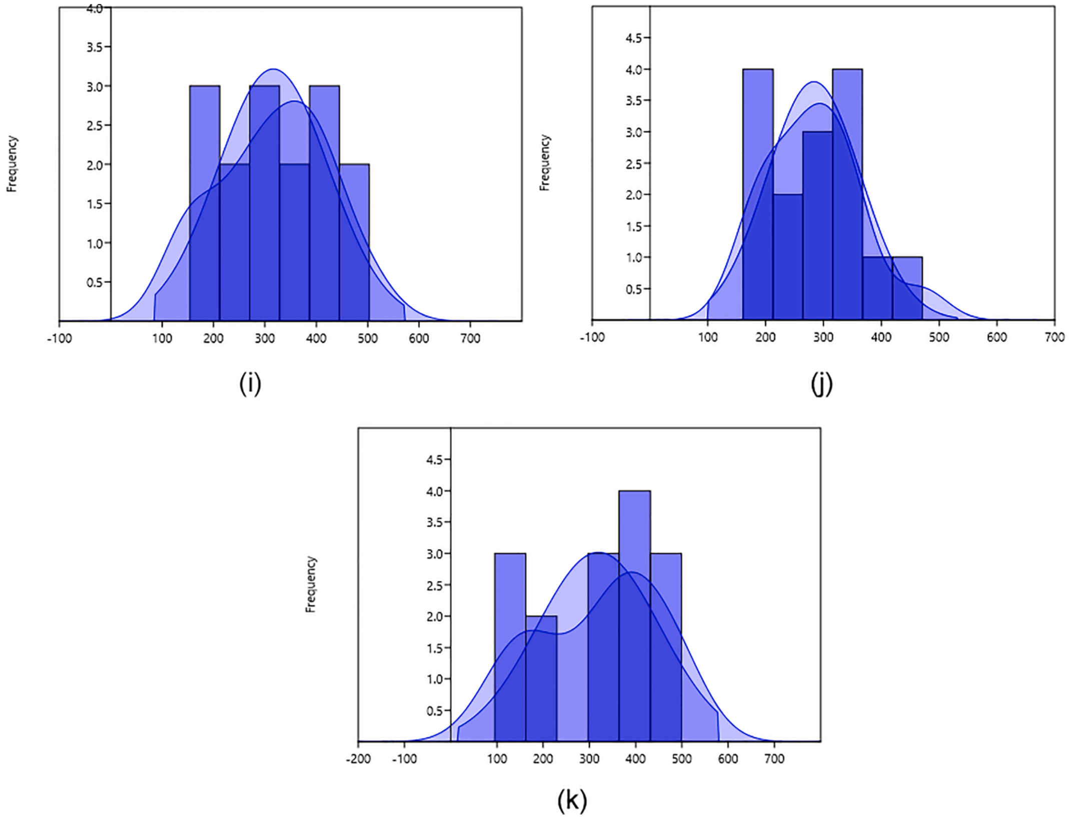


Fig. 7. A statistical model shows the country-wise binomial distribution pattern of the similarity index. (a) Binomial distribution pattern answer length from Sweden (b) Binomial distribution pattern answer length from Nigeria (c) Binomial distribution pattern answer length from USA (1st researchers) (d) Binomial distribution pattern answer length from USA (2nd researchers) (e) Binomial distribution pattern answer length from South Korea (f) Binomial distribution pattern answer length from Taiwan (g) Binomial distribution pattern answer length from India (1st researchers) (h) Binomial distribution pattern answer length from India (2nd researchers) (i) Binomial distribution pattern answer length from India (3rd researchers) (j) Binomial distribution pattern answer length from Qatar (k) Binomial distribution pattern answer length from Saudi Arabia.



(i)

(j)

(k)

Fig. 7. (*continued*).

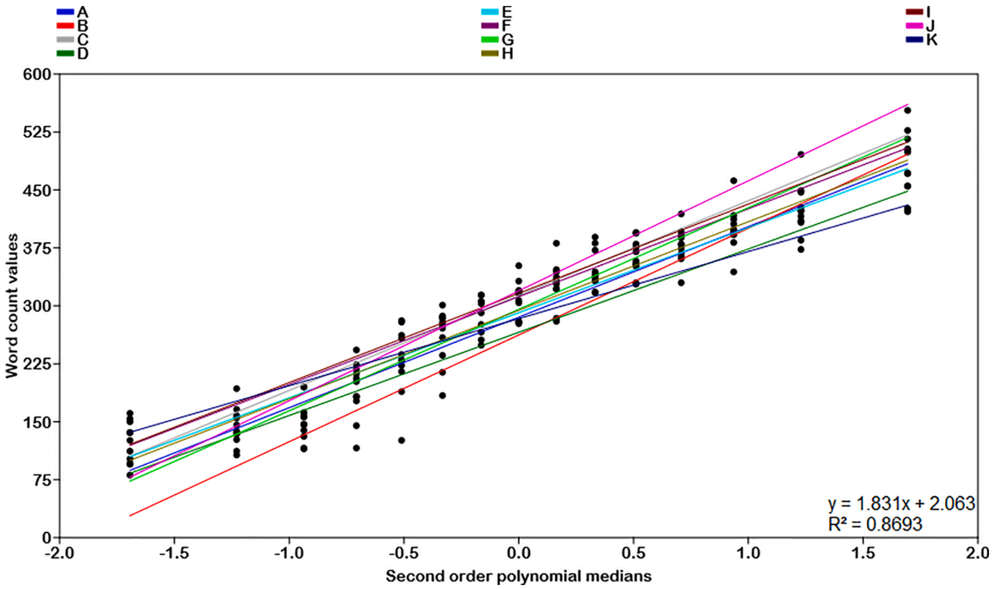


Fig. 8. A statistical model shows the country-wise answer length's second-order polynomial statistical model. The second-order polynomial equation informs the R^2 = 0.8693.

Giannos and Delardas also evaluated ChatGPT's performance in some standard admission tests in UK, such as TSA, LNAT, TMUA, and BMAT. They found that the correct responses were significantly lower than incorrect ones in some tests, such as TMUA paper 1, paper 2, and BMAT section 2. However, no significant differences were observed in the case of LNAT papers 1, 2, TSA section 1, and BMAT section 1 (Giannos and Delardas, 2023).

According to an observational study by Bhattacharyya et al., the authenticity and accuracy of references in medical articles produced by ChatGPT are a source of concern. Among the 115 references generated by ChatGPT, a mere 7 % were deemed authentic and accurate, while 47 % were found to be fabricated, and 46 % were authentic but lacked accuracy (Bhattacharyya et al., 2023). In a separate study, ChatGPT- 4.0 achieved a diagnostic accuracy rate of 57.86 % overall when solving diagnostic quizzes from the esteemed "Case of the Month" section of the American Journal of Neuroradiology (Suthar et al., 2023). In a different

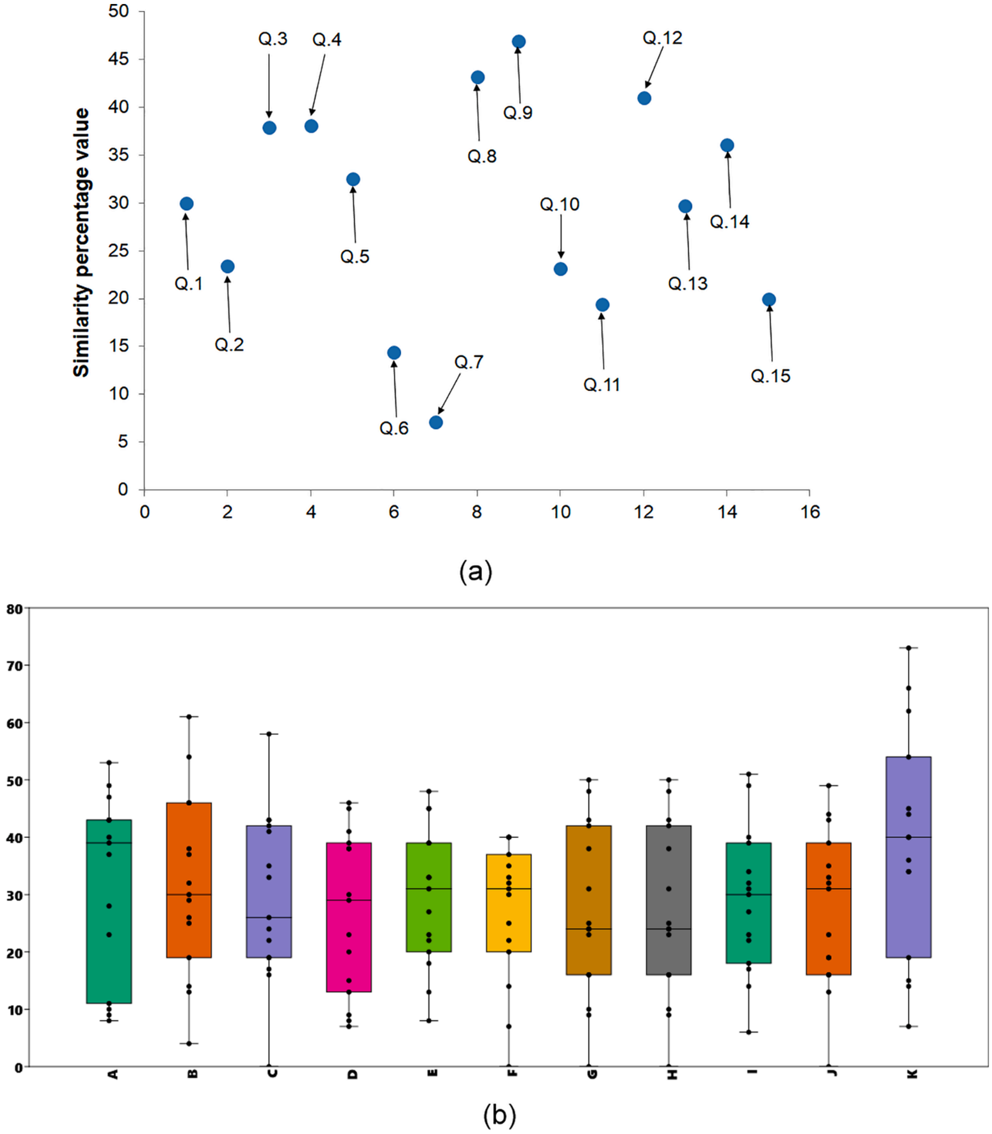


Fig. 9. ChatGPT's similarity index on the test with 15 knowledge-based questions. (a) A scatter plot using the mean similarity index of all answers from different participatory researcher from several countries. The plot represented the mean similarity index of all 15 answers. Here we found the highest similarity index in the answer to question 9 and the lowest in the answer to question 7. (b) The country-wise similarity-index pattern of ChatGPT derived answers for all questions, which has been represented through the Box-plot.

investigation, Horiuchi et al. revealed that ChatGPT achieved a diagnostic accuracy rate of 50 % in neuroradiology cases (Horiuchi et al., 2023). In an evaluation of ChatGPT's diagnostic accuracy and management recommendations for uveitis, it was found that ophthalmologists surpassed ChatGPT's likely diagnoses (Rojas-Carabali et al., 2023).

Nevertheless, ChatGPT 4.0 significantly improved comprehension of intricate surgical clinical information, achieving an impressive overall accuracy rate of 76.4 % on the Korean General Surgery Board Exam (Oh et al., 2023). In an interesting observational study, Zhu et al. demonstrated that ChatGPT can complete the BLS exam with an overall accuracy rate of 84 %. If questions were answered incorrectly, they were subsequently transformed into open-ended questions, resulting in an increased accuracy rate of 96 % and 92.1 % for the BLS (essential life support) and ACLS (advanced cardiovascular life support) exams, respectively (Zhu et al., 2023). Furthermore, Kaneda et al. assessed the performance and acceptance of responses generated by ChatGPT-3.5 and GPT-4 to Japanese childcare-related questions, and the correct answer rates were reported as 30.3 % for GPT-3.5 and 47.7 % for GPT-4 respectively (Kaneda et al., 2023). So, the accuracy rates of ChatGPT vary depending on the specific domain and task at hand, and its performance needs improvement for immediate clinical application use in patient care.

However, our study has measured the performance of ChatGPT quite comprehensively. Along with the correctness of the answer, we measured another three components of the ChatGPT-derived answers: reproducibility, similarity index, and length of the answers. In our study, we also used one set of knowledge-based questions along with the MCQ and tried to measure all those four parameters of ChatGPT-generated answers. Therefore, our study is very comprehensive.

Reproducibility is a significant component of science. Researchers attempt to understand the reproducibility of scientific methods. Our study has shown that the AI-derived model's answer is partly reproducible. The result shows the answers to questions no 1 to 10 are reproducible. On the other hand, the answers to questions no 11 to 15 are not reproducible. The AI is trained with the datasets properly when answering questions 1 to 10. Therefore, it might provide reproducible answers. On the other hand, the answers to questions no 11 to 15 are not reproducible because AI is not trained. In this case, ChatGPT informed

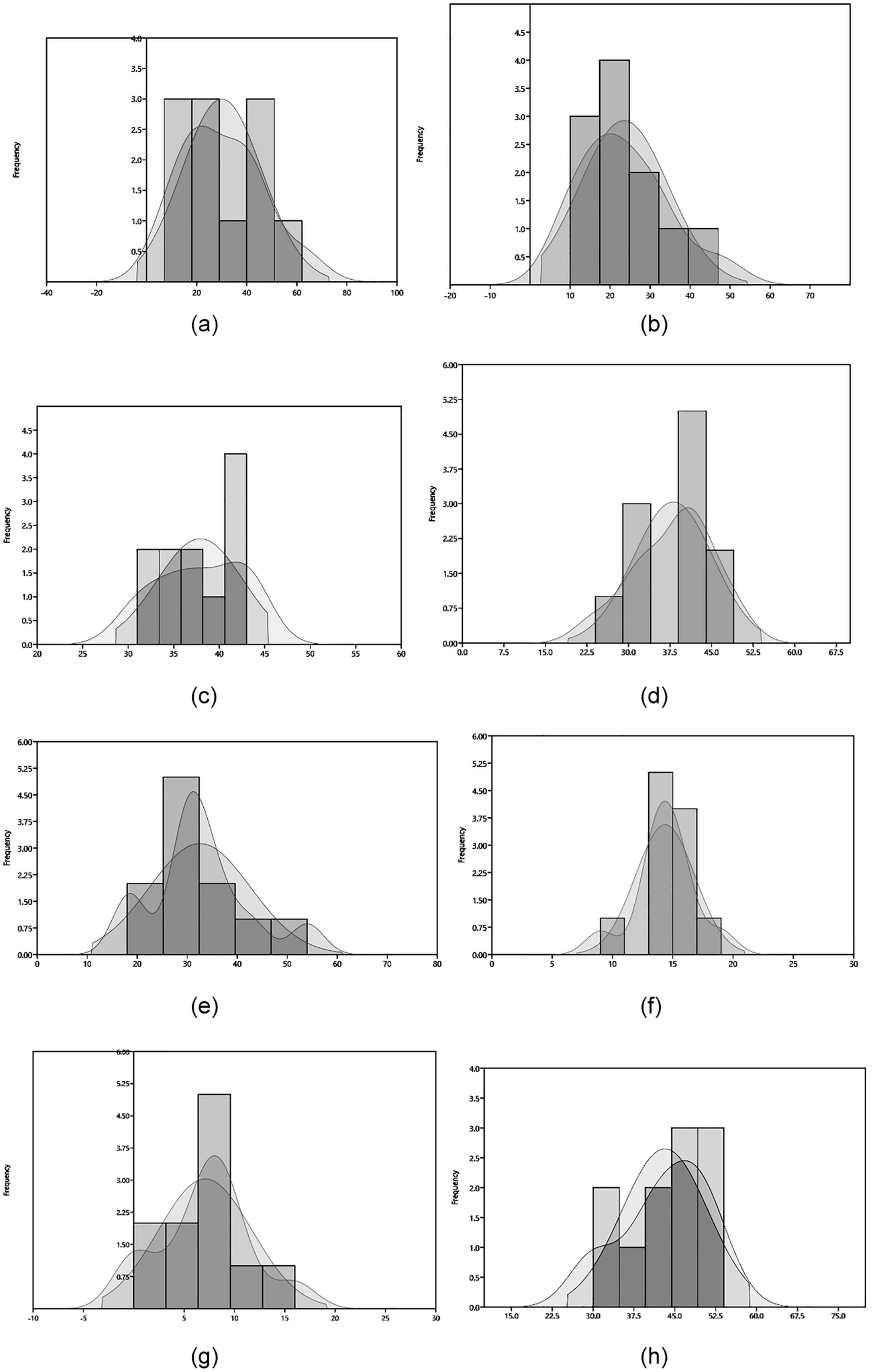
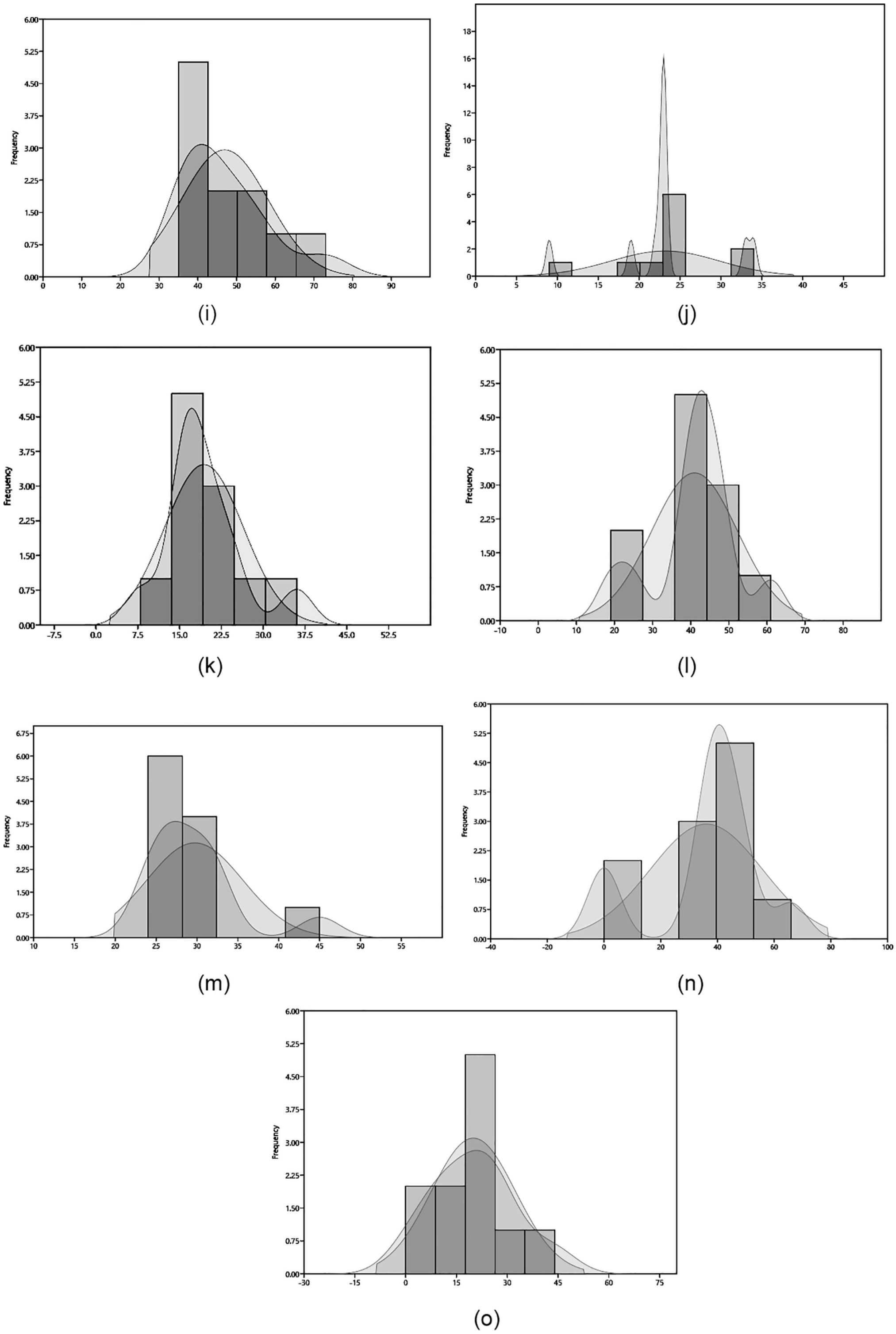


Fig. 10. A statistical model shows the question-wise binomial distribution pattern of the similarity index. (a) Binomial distribution pattern of similarity index of question 1 (b) Binomial distribution pattern of similarity index of question 2 (c) Binomial distribution pattern of similarity index of question 3 (d) Binomial distribution pattern of similarity index of question 4 (e) Binomial distribution pattern of similarity index of question 5 (f) Binomial distribution pattern of similarity index of question 6 (g) Binomial distribution pattern of similarity index of question 7 (h) Binomial distribution pattern of similarity index of question 8 (i) Binomial distribution pattern of similarity index of question 9 (j) Binomial distribution pattern of similarity index of question 10 (k) Binomial distribution pattern of similarity index of question 11 (l) Binomial distribution pattern of similarity index of question 12 (m) Binomial distribution pattern of similarity index of question 13 (n) Binomial distribution pattern of similarity index of question 14 (o) Binomial distribution pattern of similarity index of question 15.



(i)

(j)

(k)

(l)

(m)

(n)

(o)

Fig. 10. (*continued*).

that it is trained up to September 2021. Therefore, we found that AI model training is an essential factor, like ChatGPT. However, Heil and colleagues stated that all should create reproducible machine learning methods in life science research to meet the standard of the experiments, and analyses should be trustworthy (Heil et al., 2021). We urge researcher for an advanced, error-free ChatGPT or LLM, and the result of the advanced LLM should be reproducible (Chakraborty et al., 2023b; Chakraborty et al., 2023c). However, the error-free and reproducible AI model would be the gold standard method for designing future experiments for life science research. Similarly, Gundersen argues the terminology reproducibility. He discusses the challenges of reproducibility in AI-generated results. Finally, he highlights the four reproducibility types and three degrees of reproducibility (Erik Gundersen, 2021).

Our study shows the concern about the quality of ChatGPT-generated answers regarding accuracy, reproducibility, and plagiarism. The training datasets are essential for all AI models. Therefore, the quality of training datasets for ChatGPT should be improved. It has been noted that

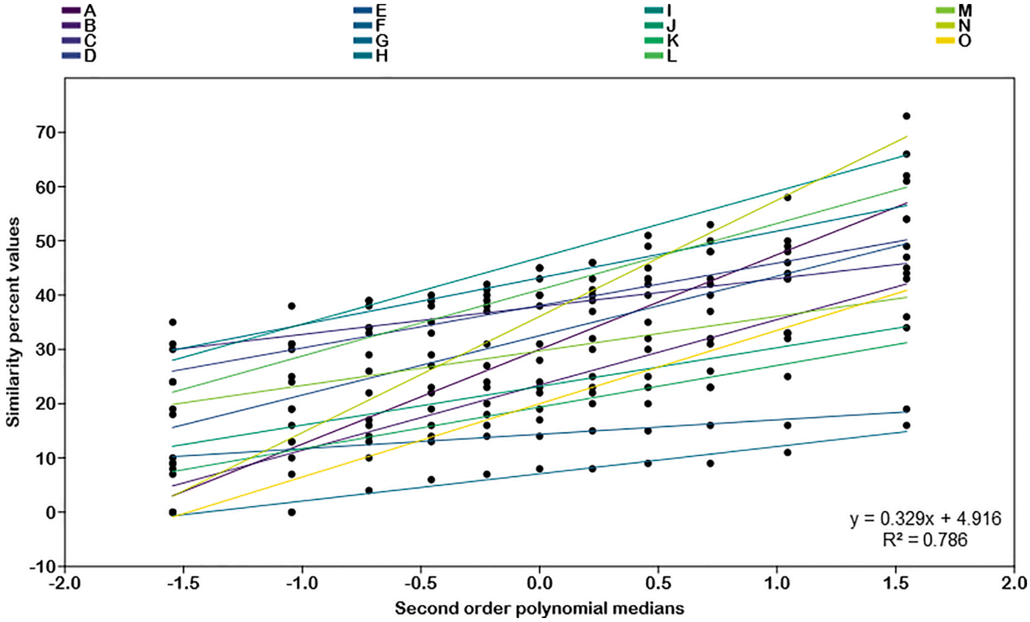


Fig. 11. A second-order polynomial statistical model developed using the question-wise answer's similarity index. The second-order polynomial equation informs the R2 = 0.786.

ChatGPT model version 3.5 was trained on the web data until June 2021. So, AI-strengthen ChatGPT should be trained with the latest datasets to provide the latest information to users. At the same time, ChatGPT should also be trained with the broad dataset.

However, there is a difference between "poor quality data" and "broad data." The solution at this point could be either to train a larger model or to fine-tune a "global" model like GPT-3. This process involves further training an already pre-trained model on a smaller, domain-specific dataset, producing better and more consistent results for the particular area of interest. Pal et al. have urged a next-generation, domain-specific LLM or ChatGPT (Pal et al., 2023a). Domain-specific LLM or ChatGPT might provide a more accurate answer with more information, as they are trained in domain-specific datasets. Many strategies should be explored regarding reproducibility, like a new algorithm for the Chatbot, which might help to produce a reproducible answer. However, plagiarism became an issue once the ChatGPT-derived text was used in the publications. Allthough, ChatGPT-derived text might greatly help non-native english-speaking countries for the publications (Osama et al., 2023; Hwang et al., 2023). Henceforth, there is a controversy about using ChatGPT-derived text in a publication.

Our study has some significant limitations. We used a relatively small input size (two sets of questionnaire, each set containing 15 questions) for these four components of analyses and the range of analyses. At the same time, after recording the answers from eleven researchers from different countries, we found differences in the responses to the same question with identical prompts. Several other researchers reported that the ChatGPT did not produce identical answers to the same question with identical prompts. In one experiment, Fergus et al. tried to understand whether ChatGPT provided identical answers to the same question with identical prompts. To understand the answer pattern of ChatGPT, they used two different user accounts (M.O. and M.B.) and found that ChatGPT did not produce identical answers to identical question prompts (Fergus et al., 2023). Similarly, Cheung et al. found that ChatGPT delivers different answers even with the same prompt (Cheung et al., 2023).

Different answers to the same question with identical prompts are generated due to the probabilistic nature of the large language model (LLM). ChatGPT is an LLM; it provides different answers to the same question with identical prompts. However, choosing the eleven researchers from different locations is a random phenomenon. There is no connection between choosing the countries and the probabilistic nature of the LLM. The geographical location might not influence the probabilistic nature of LLM. Therefore, geographical location has no role in different answers to the same question with identical prompts. However, several researchers tried to describe the reason for the different answers to the same question with identical prompts. Chang et al. describe the LLM generates and assembles tokens through the probabilistic for Blackbox of LLM (Chang et al., 2023).

Similarly, we previously illustrated the Blackbox nature of AI or ChatGPT. However, AI is a complicated process, and understanding the AI's Blackbox is challenging (Chakraborty et al., 2023c). Therefore, prompt engineering might minimize the difference between the text of answers to the same question when different people ask it. However, more studies are needed in this direction.

Conclusion

The performance of this AI-based ChatGPT model has quickly been popularized in different fields. In medical science, it is used in diversified applications such as medical education, clinical trial, writing doctor's discharge summaries, and several others. Researchers are also using it in research report writing. However, our study raises concern about the AI-based ChatGPT model and its various properties. AI-strengthened ChatGPT has billions of users with widespread applications. The number of users of ChatGPT is increasing day by day. It has been noted that the provided answer of the present version of the ChatGPT (i.e., GPT 3.5) has shown a similarity with other texts. Therefore, there is an urgent need for a novel algorithm for the chatbot, which will be used for plagiarism-free scientific script writing. Pal et al. have urged in this direction (Pal et al., 2023b). Similarly, the incorrect information of ChatGPT might mislead users. Therefore, it is urgently necessary to develop error-free LLMs-derived chatbots like ChatGPT to avoid user misguidance as the applications of LLMs-derived chatbots are increasing very quickly. Our study indicated an urgent need for greater awareness of AI-strengthened LLMs like ChatGPT. Moreover, appropriate guidelines and regulations are needed for all stakeholders to use LLMs.

At the same time, topics should be included in real-world learning in the education across the students for its safe, responsible, and proper use. We urge all stakeholders, such as governments, countries'

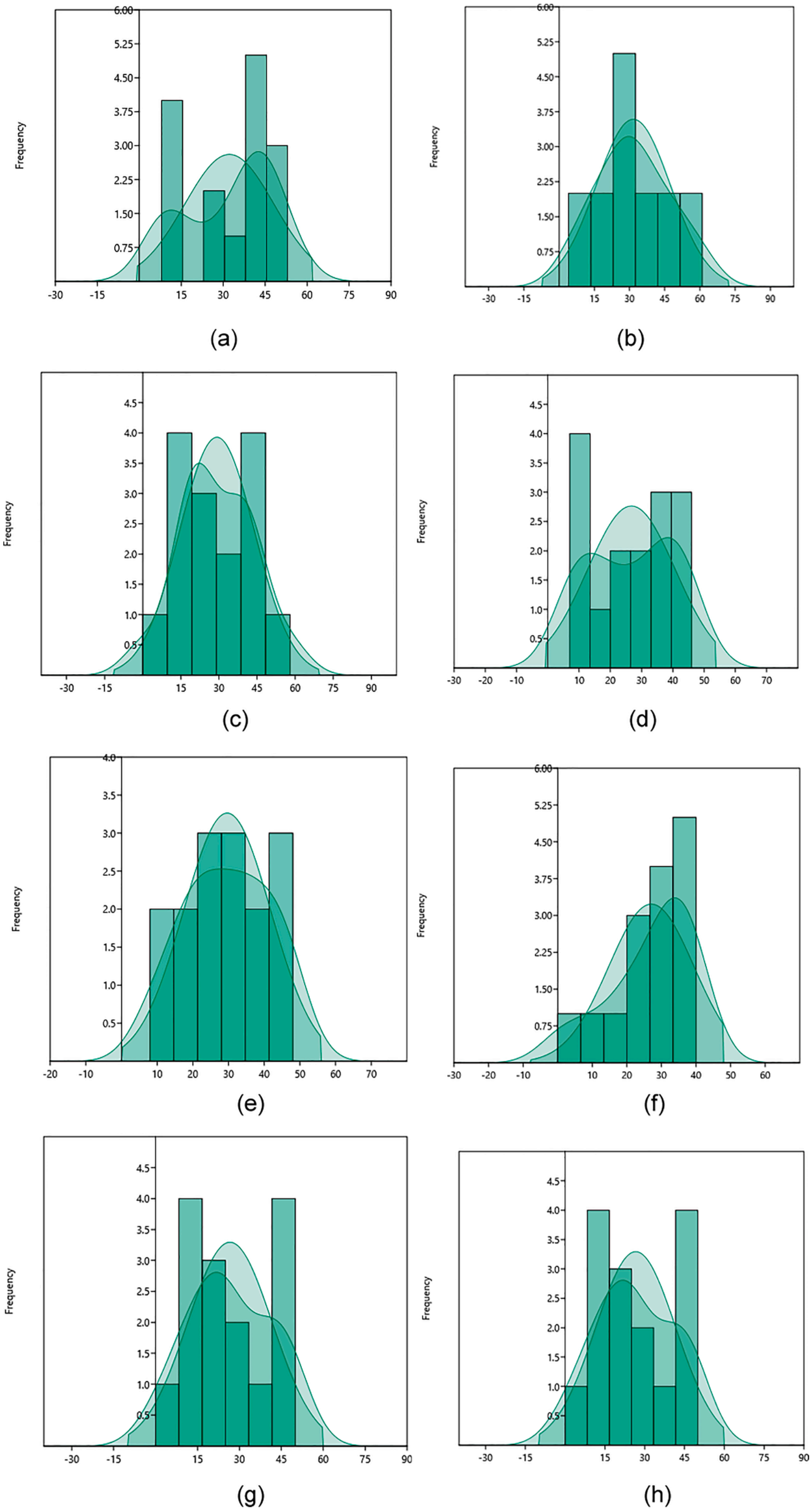
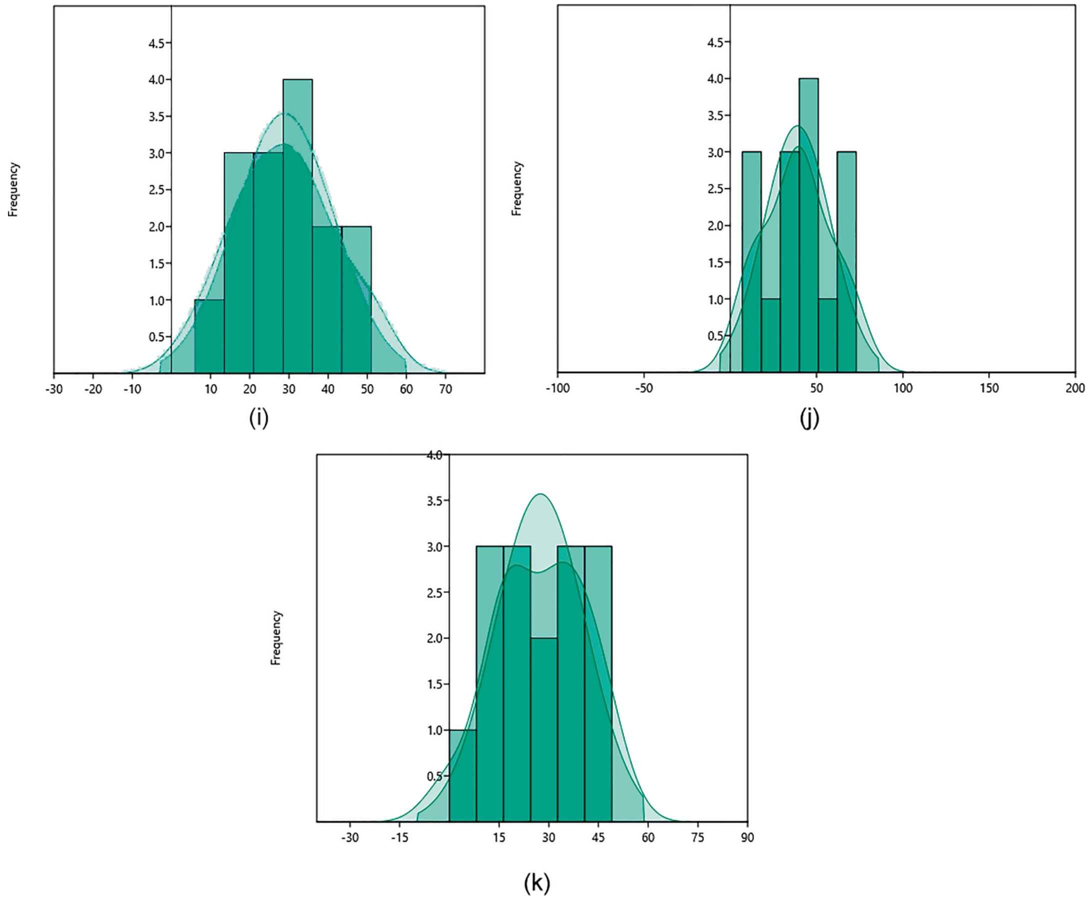


Fig. 12. A statistical model shows the university-wise binomial distribution pattern of the similarity index. (a) Binomial distribution pattern of similarity index Sweden (b) Binomial distribution pattern of similarity index from Nigeria (c) Binomial distribution pattern of similarity index from USA (1st researchers) (d) Binomial distribution pattern of similarity index from USA (2nd researchers) (e) Binomial distribution pattern of similarity index from South Korea (f) Binomial distribution pattern of similarity index from Taiwan (g) Binomial distribution pattern of similarity index from India (1st researchers) (h) Binomial distribution pattern of similarity index from India (2nd researchers) (i) Binomial distribution pattern of similarity index from India (3rd researchers) (j) Binomial distribution pattern of similarity index from Qatar (k) Binomial distribution pattern of similarity index from Saudi Arabia.



(i)

(j)

(k)

Fig. 12. (*continued*).

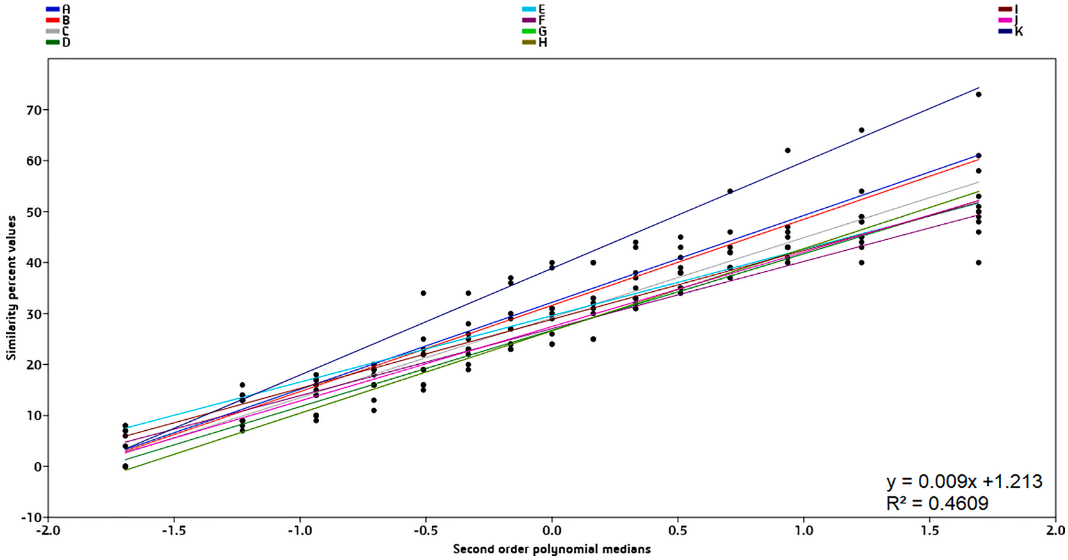


Fig. 13. A second-order polynomial statistical model developed using the country wise answer's similarity index. The second-order polynomial equation informs the $R^2 = 0.4609$.

policymakers, lawyers, healthcare professionals, ethicists, computer professionals, and scientists, to involve immediately and identify the solutions to move forward.

CRediT authorship contribution statement

Manojit Bhattacharya: Validation. **Soumen Pal:** Validation. **Srijan Chatterjee:** Validation. **Abdulrahman Alshammari:** Validation. **Thamer H. Albekairi:** Validation. **Supriya Jagga:** Validation. **Elijah Ige Ohimain:** Validation. **Hatem Zayed:** Validation. **Siddappa N.**

Byrareddy: Validation. **Sang-Soo Lee:** Validation. **Zhi-Hong Wen:** Validation. **Govindasamy Agoramoorthy:** Validation. **Prosun Bhattacharya:** Validation. **Chiranjib Chakraborty:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

Authors are thankful to the Researchers Supporting Project number (RSP2024R491), King Saud University, Riyadh, Saudi Arabia.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crbiot.2024.100194.

References

Ali, S.R., Dobbs, T.D., et al., 2023. Using ChatGPT to write patient clinic letters. Lancet Digit Health 5 (4), e179–e181.

Alser, M., Waisberg, E., 2023. "Concerns with the usage of ChatGPT in Academia and Medicine: A viewpoint." Am. J. Med. Open 100036.

Anderson, L.W., (2003). Classroom assessment: Enhancing the quality of teacher decision making. Routledge. ISBN 1135657602, 9781135657604 (200 pp).

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533 (7604), 452–454. https://doi.org/10.1038/533452a.

Bhattacharyya, M., Miller, V.M., et al., 2023. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Cureus 15 (5), e39238.

Chakraborty, C., Bhattacharya, M., et al., 2023c. ChatGPT indicates the path and initiates the research to open up the black box of artificial intelligence. Int. J. Surg. 109 (12), 4367–4368. https://doi.org/10.1097/JS9.0000000000000701.

Chakraborty, C., Bhattacharya, M., et al., 2023b. Need an AI-enabled, next-generation, advanced ChatGPT or large language models (LLMs) for error-free and accurate medical information. Ann. Biomed. Eng. 52 (2), 134–135.

Chakraborty, C., Pal, S., et al., 2023a. Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Front. Artif. Intell. 6, 1237704.

Chang, Y., Wang, X., Wang, J., et al., 2023. A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. https://doi.org/10.48550/arXiv.2307.03109.

Chatterjee, S., Bhattacharya, M., et al., 2023. Can artificial intelligence-strengthened ChatGPT or other large language models transform nucleic acid research? Mol. Therapy-Nucleic Acids 33, 205–207.

Cheng, Y., Cai, Y., et al., 2021. A cognitive level evaluation method based on a deep neural network for online learning: from a bloom's taxonomy of cognition objectives perspective. Front. Psychol. 12, 661235 https://doi.org/10.3389/fpsyg.2021.661235.

Cheung, B.H.H., Lau, G.K.K., et al., 2023. "ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023;18(8): e0290691. doi:10.1371/journal.pone.0290691.

Dave, T., Athaluri, S.A., et al., 2023. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front. Artif. Intell. 6, 1169595.

De Angelis, L., Baglivo, F., et al., 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front. Public Health 11, 1166120.

Editorials, 2023. Will ChatGPT transform healthcare? Nat. Med. 29 (3), 505–506.

Engineering, Medicine, 2019. "Replicability." Reproducibility and Replicability in Science, Washington (DC): National Academies Press (US). ISBN: 978-0-309-48619-4 (268 pp).

Erik Gundersen, O., 2021. The fundamental principles of reproducibility. Philos. Trans. A Math. Phys. Eng. Sci. 379 (2197).

Fergus, S., Botha, M., Ostovar, M., 2023. Evaluating academic answers generated using ChatGPT. J. Chem. Educ. 100 (4), 1672–1675.

Giannos, P., Delardas, O., 2023. Performance of ChatGPT on UK Standardized Admission Tests: Insights From the BMAT, TMUA, LNAT, and TSA Examinations. JMIR Med. Educ. 9.

Gilson, A., Safranek, C.W., et al., 2023. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med. Educ. 9.

Habibzadeh, F., 2023. Plagiarism: A Bird's Eye View. J. Korean Med. Sci. 38 (45), e373.

Halgamuge, M.N., 2017. The use and analysis of anti-plagiarism software: Turnitin tool for formative assessment and feedback. Comput. Appl. Eng. Educ. 25 (6), 895–909.

Hammer, O., 2001. PAST: Paleontological statistics software package for education and data analysis. Palaeontol. Electron. 4, 9.

Heil, B.J., Hoffman, M.M., et al., 2021. Reproducibility standards for machine learning in the life sciences. Nat. Methods 18 (10), 1132–1135.

Homolak, J., 2023. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat. Med. J. 64 (1), 1–3.

Horiuchi, D., Tatekawa, H., et al., 2023. Ueda D. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. DOI: 10.1007/s00234-023-03252-4.

Humar, P., Asaad, M., et al., 2023. "ChatGPT is Equivalent to First Year Plastic Surgery Residents: Evaluation of ChatGPT on the Plastic Surgery In-Service Exam." Aesthet Surg J.

Hutson, M., 2022. Could AI help you to write your next paper? Nature 611 (7934), 192–193.

Hwang, S.I., Lim, J.S., et al., 2023. Is ChatGPT a "Fire of Prometheus" for Non-Native English-Speaking Researchers in Academic Writing? Korean J. Radiol. 24 (10), 952–959. https://doi.org/10.3348/kjr.2023.0773.

Iftikhar, L., 2023. Docgpt: Impact of chatgpt-3 on health services as a virtual doctor. EC Paediatrics 12 (1), 45–55.

Kaneda, Y., Namba, M., et al., 2023. Artificial Intelligence in Childcare: Assessing the Performance and Acceptance of ChatGPT Responses. Cureus. 15 (8), e44484.

Khan, R.A., Jawaid, M., et al., 2023. ChatGPT - Reshaping medical education and clinical management. Pak. J. Med. Sci. 39 (2), 605–607.

Kung, T.H., Cheatham, M., et al., 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2 (2), e0000198.

Mann, D.L., 2023. Artificial Intelligence Discusses the Role of Artificial Intelligence in Translational Medicine: A JACC: Basic to Translational Science Interview With ChatGPT. JACC Basic Transl. Sci. 8 (2), 221–223.

Mbakwe, A.B., Lourentzou, I., et al., 2023. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2 (2).

Oh, N., Choi, G.S., et al., 2023. 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. Ann Surg Treat Res. 104 (5), 269–273. https://doi.org/10.4174/astr.2023.104.5.269.

Osama, M., Afridi, S., et al., 2023. ChatGPT: Transcending Language Limitations in Scientific Research Using Artificial Intelligence. J. Coll. Physicians Surg. Pak. 33 (10), 1198–1200. https://doi.org/10.29271/jcpsp.2023.10.1198.

Pal, S., Bhattacharya, M., et al., 2023a. A domain-specific next-generation large language model (LLM) or ChatGPT is required for biomedical engineering and research. Ann. Biomed. Eng. 52 (3), 451–454.

Pal, S., Bhattacharya, M., et al., 2023b. AI-enabled ChatGPT or LLM: A new algorithm is required for plagiarism free scientific writing. Int. J. Surg. https://doi.org/10.1097/JS9.0000000000000939.

Patel, S.B., Lam, K., 2023. ChatGPT: the future of discharge summaries? Lancet Digit Health 5 (3), e107–e108.

Rojas-Carabali, W., Cifuentes-González, C., et al., 2023. Evaluating the diagnostic accuracy and management recommendations of ChatGPT in Uveitis. Ocul. Immunol. Inflamm. 1–6 https://doi.org/10.1080/09273948.2023.2253471.

Ruksakulpiwat, S., Kumar, A., et al., 2023. Using ChatGPT in medical Research: Current Status and Future Directions. J. Multidiscip. Healthc. 16, 1513–1520.

Shanahan, M., McDonell, K., et al., 2023. Role play with large language models. Nature 623 (7987), 493–498.

Stringer, J.K., Santen, S.A., et al., 2021. Examining bloom's taxonomy in multiple choice questions: students' approach to questions. Med Sci Educ. 31 (4), 1311–1317.

Suthar, P.P., Kounsal, A., et al., 2023. Artificial Intelligence (AI) in Radiology: A Deep Dive Into ChatGPT 4.0's Accuracy with the American Journal of Neuroradiology's (AJNR) "Case of the Month". Cureus. 15 (8), e43958.

Ventayen, R. J. M, 2023. OpenAI ChatGPT Generated Results: Similarity Index of Artificial Intelligence-Based Contents (January 21, 2023). Advances in Intelligent Systems and Computing, Available at SSRN: https://ssrn.com/abstract=4332664 or https://doi.org/10.2139/ssrn.4332664.

Weng, T.L., Wang, Y.M., et al., 2023. ChatGPT failed Taiwan's Family Medicine Board Exam. J. Chin. Med. Assoc. 86 (8), 762–766.

Zhu, L., Mou, W., et al., 2023. ChatGPT can pass the AHA exams: Open-ended questions outperform multiple-choice format. Resuscitation 188, 109783. https://doi.org/10.1016/j.resuscitation.2023.109783.