




Development and Evaluation of a Diagnostic Exam for Undergraduate Biomedical Engineering Students Using GPT Language Model-Based Virtual Agents

Alberto Isaac Perez Sanpablo^{1,2}(✉) , María del Carmen Arquer Ruiz¹, Alicia Meneses Peñaloza³, Gerardo Rodríguez Reyes⁴, Ivett Quiñones Uriostegui², and Leonardo Eliú Anaya Campos²

¹ Biomedical Engineering Department, La Salle University, Cuauhtemoc, 06140 Mexico City, Mexico

² Laboratory of Motion Analysis and Rehabilitation Engineering, National Institute of Rehabilitation, Tlalpan, 14389 Mexico City, Mexico
albperez@inr.gob.mx

³ Pediatric Rehabilitation Department, National Institute of Rehabilitation, Tlalpan, 14389 Mexico City, Mexico

⁴ Technological Research Department, National Institute of Rehabilitation, Tlalpan, 14389 Mexico City, Mexico

Abstract. The creation of a diagnostic exam for biomedical engineering undergraduate students using virtual agents based on GPT language models is analyzed in this work. Thirty-nine eighth-semester students answered a 20-question exam generated by ChatGPT-3 covering the topics of acquisition, amplification, processing, and visualization of biomedical signals encompassing different levels of thinking according to the taxonomy of Bloom, including the application, analysis, and evaluation levels. Three academic experts assessed the quality of the questions based on clarity, relevance, level of thinking, and difficulty. Also, difficulty and discrimination indexes and Rasch analysis were calculated. Students obtained an average grade of 5.91, with a standard deviation of 1.39 points. Subject reliability was 0.599, and the p-value for the fit of the model of Rasch was 0.017. High correlations between some questions were observed. Based on their difficulty, a few questions could be considered irrelevant. The Wright competency map showed a good distribution on the ability scale with some redundancies and gaps. In conclusion, virtual agents have great potential to create diagnosis exams in biomedical engineering. However, it is necessary to consider their limitations and conduct a rigorous evaluation of the quality and reliability of the questions generated.

Keywords: ChatGPT-3 · education · biomedical engineering · teaching · artificial intelligence

1 Introduction

Artificial intelligence and virtual agents based on GPT language models have recently gained relevance in academia [1]. Texts on specific academic topics can be generated by these tools facilitating the teaching and learning process. Their use allows academics to focus on essential tasks through time savings [2]. These tools can simplify tasks for scholars and improve their efficiency by making summaries and translating information [3]; Virtual agents can help students learn by acting as tutors and answering their questions. For example, in microbiology, ChatGPT-3, a virtual agent based on GPT language models, has proven to be effective for automatically answering questions, providing students with accurate and relevant information [4].

However, scholars must know the limitations and potential biases of virtual agents. Although the information generated by these tools may seem authentic to a not fully trained reader, such as a student-in-training, it is essential to recognize that ChatGPT needs help interpreting and understanding the content profoundly, which can probably cause it to generate incorrect information [5]. Thus, there is concern about the reliability and possible biases of virtual agents since these tools are trained on large amounts of data and may reflect the tendencies present in that data. Therefore, it is essential to verify the facts of all virtual agent statements and be aware of the possibility of incorrect content. Another reported aspect is the tendency of virtual agents to generate information not based on actual events, known as “hallucinating” [6]. Therefore, it is necessary to exercise critical judgment when evaluating the text generated by virtual agents and use reliable sources to contrast the information.

Despite these considerations, the use of virtual agents has the potential to increase scholarly output. These tools can help in the creation of questionnaires. These agents have been used in medicine to organize material and generate and correct texts [7]. Thus, virtual agents have potential benefits for the academy, such as the generation of exams; however, their limited interpretation capacity, biases, and errors must be considered, so it is essential to verify the information and use a critical approach to evaluate its content. Therefore, this work aims to evaluate the use of virtual agents based on GPT language models to create diagnostic tests for biomedical engineering courses. This will be done by analyzing the quality of the questions generated.

There are multiple reports exploring the use of ChatGPT to answer quizzes. Although some authors propose the creation of questions, only one have tested this hypothesis by creating multiple-choice questions. The study compared various virtual agents to generate reasoning-based multiple-choice questions in medical physiology [8]. The authors found that virtual agents require improvement and that each tool had limitations, as Bing developed less valid and ChatGPT less complex questions. This work contributes to the field of biomedical engineering education by exploring the potential of virtual agents in creating personalized diagnostic tests at different cognitive levels. The work validates the quality of the exam and considers practical issues. The findings of the article could improve assessment methods in biomedical engineering education to improve the quality of education for students in this field.

2 Methodology

The methodology used in this study was based on the ChatGPT-3 virtual agent due to its more equitable and unrestricted access compared to more recent models such as ChatGPT-4. The study aimed to create a 20-question multiple-choice diagnostic exam for biomedical engineering students in their final year before the biomedical measurements course. The biomedical measurements course focuses on designing instruments for electrophysiological signal measurements to support the clinical diagnosis of various pathologies. The course is part of the curriculum line of applied engineering. The questions evaluated topics related to the acquisition, amplification, processing, and visualization of biomedical signals, which consider different levels of thinking according to the taxonomy of Bloom, including the application, analysis, and evaluation levels. An example of a prompt for ChatGPT-3 to generate questions of analysis level is shown in Fig. 1.

Create a diagnostic exam for biomedical engineering undergraduate students with ten multiple-choice questions to evaluate Bloom's Cognitive Dimension of Analyze. Each question has three options; you must indicate the correct answer. Cover the topics of acquisition, amplification, processing, display of biosignals

Fig. 1. Example of a prompt for ChatGPT-3 to generate questions of analysis level.

The exam creation process began with a researcher reviewing the content and selecting twenty out of thirty questions generated by the virtual agent based on their clarity, relevance, level of thinking, and difficulty. Subsequently, the selected questions were captured in a question bank using the Moodle learning platform. With this bank of questions, the exam was created, where each question and its response options were presented individually and sequentially, with a maximum resolution time of 20 min.

The exam was applied to all students in the last year of a degree program in biomedical engineering, certified according to CACEI. The quality of each question was evaluated by an external sample of academic experts in the subject who had training in biomedical engineering and at least five years of teaching experience related to the topics to be evaluated. The consistency of the evaluators was calculated using the Intraclass correlation coefficient (ICC).

Academics were asked to rate the clarity, validity and relevance, level of thinking, and difficulty of each question and answer through a survey using a standardized form in Word. The document contained a supporting explanation for each domain to be assessed. Clarity and relevance were assessed using a dichotomous variable (yes/no), the thinking level was classified according to the levels of Bloom (application, analysis, or evaluation), and the difficulty was classified as low, medium, or high.

General descriptive statistics were performed for test subjects and scores. The difficulty and discrimination indexes were calculated for each question. The Rasch analysis was carried out in the free software Jamovi [9], where the reliability of the subjects, the p-value for the model fit, the correlation matrix of all the exam questions, and the skills map of Wright for the test were calculated. The difficulty index measures the difficulty of the question calculated by dividing the number of students who answered the question correctly by the total number of students. The discrimination index measures the ability

of questions to distinguish between high and low-achieving students. Subject reliability is a measure of the consistency of students in their responses. The question correlation matrix shows the correlation between all questions and may indicate that a pair of questions measure the same thing. The Wright Skill Map is a graphical representation of the relationship between student skills and the difficulty of test questions. The ability and error in estimating item ability measures how well the Rasch model estimates the skills of students. The fit of the answer to each question (infit) and the general fit of all the answers (outfit) to the model indicates how well the question fits the model.

3 Results

The 20-question diagnostic exam was applied to 39 eighth-semester biomedical engineering Spanish-speaking students (see Fig. 2).

| Question ID | Description |
|-------------|--|
| A01 | Question about types of biosignal acquisition systems. |
| A02 | Inquiry about the definition of gain in the context of an amplifier. |
| A03 | Question about digital signal processing filters. |
| A04 | Question about the purpose of signal averaging in biosignal processing. |
| A05 | Question about display methods for electrocardiogram signals. |
| A06 | Inquiry about biosignal amplifiers. |
| A07 | Question about the Nyquist frequency. |
| A08 | Exploration of methods of biosignal processing. |
| A09 | Question about the characteristics of biosignal acquisition systems. |
| A10 | Inquiry about types of biosignal processing techniques. |
| Ap01 | Question about the component responsible for converting analog signals to digital in a biosignal acquisition system. |
| Ap02 | Question about a method to reduce noise in biosignals. |
| Ap03 | Exploration of the purpose of amplification in biosignal processing. |
| Ap04 | Question about the primary function of a biosignal processing system. |
| Ap05 | Inquiry about a type of biosignal that requires a high sampling rate for accurate acquisition. |
| E01 | Question about the primary purpose of signal amplification in biomedical signal processing. |
| E02 | Exploration of disadvantages of using digital filters for signal processing. |
| E03 | Question about factors considered when selecting the appropriate sampling rate for a biosignal. |
| E04 | Inquiry about the type of electrode commonly used for recording electroencephalography (EEG) signals. |
| E05 | Question about the most critical parameter to consider when selecting a filter for biosignal processing. |

Fig. 2. Description of the 20 questions obtained by ChatGPT for the diagnostic exam. Analysis, application, and evaluation questions are coded as A, Ap, and E.

The average grade obtained on the exam was 5.91, with a standard deviation of 1.39, in a total score range of 3.00 to 9.00 points. The exam was designed to assign half a point to each correct answer with a maximum total score of 10 points. The exam was written in English. Some students needed help with an English written exam. Students were helped by paraphrasing and using synonyms to clarify questions. All questions were classified as relevant by the three experts. Nine questions were related to the amplification topic, five to acquisition and processing, and one to display. Of the 20 questions, ten selected questions corresponded to the analysis level of thinking (A01 to A10) and five to the

application and evaluation levels (Ap01 to Ap05 and E01 to E05, see Fig. 3). There was no consistency among evaluators regarding the clarity and level of thinking of questions (ICC < 0.01, p-value = 0.48). However, in most of the questions, at least two experts agreed on the level of thinking (N = 15) and clarity (N = 14). On the other hand, a moderate agreement (ICC = 0.55, p-value < 0.01) was reached between experts for the difficulty of the questions.

| Question ID/ Experts | Clarity | | | Level of thinking | | | Difficulty | | |
|-------------------------|---------|-----|-----|-------------------|------------|-----------|------------|--------|--------|
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| A01 | Yes | Yes | Yes | Applying | Evaluating | Applying | low | low | low |
| A02 | Yes | Yes | Yes | Analyzing | Evaluating | Analyzing | low | low | medium |
| A03 | Yes | Yes | Yes | Applying | Evaluating | Applying | low | low | low |
| A04 | Yes | Yes | Yes | Analyzing | Evaluating | Analyzing | medium | medium | medium |
| A05 | Yes | No | Yes | Applying | Analyzing | Analyzing | low | high | medium |
| A06 | Yes | No | Yes | Applying | Analyzing | Applying | low | low | low |
| A07 | Yes | No | Yes | Applying | Evaluating | Applying | medium | medium | low |
| A08 | Yes | Yes | Yes | Analyzing | Evaluating | Applying | low | medium | low |
| A09 | Yes | Yes | Yes | Applying | Evaluating | Analyzing | low | medium | medium |
| A10 | Yes | No | Yes | Applying | Analyzing | Analyzing | low | high | medium |
| Ap01 | Yes | Yes | Yes | Applying | Analyzing | Applying | low | medium | low |
| Ap02 | Yes | No | Yes | Applying | Evaluating | Analyzing | low | low | medium |
| Ap03 | Yes | Yes | Yes | Analyzing | Evaluating | Applying | medium | medium | low |
| Ap04 | Yes | Yes | Yes | Analyzing | Analyzing | Analyzing | medium | high | medium |
| Ap05 | Yes | Yes | Yes | Analyzing | Analyzing | Analyzing | high | high | high |
| E01 | Yes | Yes | Yes | Applying | Analyzing | Applying | low | high | low |
| E02 | Yes | Yes | Yes | Analyzing | Analyzing | Analyzing | high | high | high |
| E03 | Yes | No | Yes | Analyzing | Analyzing | Analyzing | high | high | high |
| E04 | Yes | Yes | Yes | Applying | Evaluating | Applying | low | low | low |
| E05 | Yes | Yes | Yes | Analyzing | Analyzing | Analyzing | medium | high | high |

Fig. 3. Analysis by experts of the validity: clarity, level of thinking, and difficulty of each question. Analysis, application, and evaluation questions are coded as A, Ap, and E.

All students answered one knowledge application question correctly (Ap02) and was therefore omitted from the Rasch analysis. Rasch analysis revealed a subject reliability of 0.599 and a p-value for the model fit of 0.017, indicating acceptable consistency in student responses. The correlation matrix showed a few questions with a strong correlation between them (see Fig. 4). The highest correlation was found between knowledge application questions (A01 and A10), while another knowledge application question (A08) presented several high correlations (>0.3) with three questions of the same thinking level (A04, A05, and A06).

The diagnostic exam questions presented an average difficulty index of 0.57, with a standard deviation of 0.27 (see Fig. 5). However, according to this index, most of the questions (12 in total) could be considered invalid or irrelevant because they were straightforward (>0.7) or very difficult (<0.3). On the other hand, the average ability of the students was -0.47, with a standard deviation of 1.66. The average infit and outfit values were close to 1, with standard deviations of 0.09 and 0.31, respectively. However, one analysis question (A10) was identified that presented a significant outfit deviation (2.040), suggesting the presence of unexpected answers, such as a low-ability student

| | A01 | E01 | E02 | E03 | E04 | E05 | A02 | A03 | A04 | A05 | A06 | A07 | A08 | A09 | A10 | Ap01 | Ap03 | Ap04 | Ap05 | |
|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|--|
| A01 | — | | | | | | | | | | | | | | | | | | | |
| E01 | 0.27 | — | | | | | | | | | | | | | | | | | | |
| E02 | -0.38 | 0.17 | — | | | | | | | | | | | | | | | | | |
| E03 | 0.04 | -0.03 | 0.05 | — | | | | | | | | | | | | | | | | |
| E04 | 0.26 | 0.13 | -0.08 | 0.04 | — | | | | | | | | | | | | | | | |
| E05 | -0.21 | -0.16 | 0.14 | 0.10 | -0.14 | — | | | | | | | | | | | | | | |
| A02 | 0.09 | -0.01 | -0.26 | 0.01 | -0.19 | 0.21 | — | | | | | | | | | | | | | |
| A03 | -0.11 | 0.01 | -0.10 | 0.09 | 0.03 | 0.21 | 0.01 | — | | | | | | | | | | | | |
| A04 | 0.14 | 0.07 | -0.05 | 0.06 | -0.20 | 0.17 | 0.35 | -0.21 | — | | | | | | | | | | | |
| A05 | -0.21 | -0.04 | 0.22 | 0.18 | -0.17 | 0.09 | -0.46 | 0.17 | -0.17 | — | | | | | | | | | | |
| A06 | 0.24 | 0.04 | -0.09 | -0.14 | 0.35 | 0.03 | 0.09 | -0.04 | 0.08 | -0.29 | — | | | | | | | | | |
| A07 | 0.12 | -0.13 | 0.11 | -0.23 | -0.01 | -0.45 | -0.06 | -0.23 | -0.24 | -0.06 | 0.25 | — | | | | | | | | |
| A08 | -0.25 | -0.21 | 0.25 | 0.06 | 0.05 | 0.33 | -0.07 | 0.24 | -0.37 | 0.32 | -0.36 | 0.02 | — | | | | | | | |
| A09 | -0.08 | -0.14 | 0.14 | 0.11 | -0.11 | 0.07 | -0.14 | 0.00 | -0.15 | 0.21 | -0.17 | -0.34 | -0.08 | — | | | | | | |
| A10 | -0.56 | -0.34 | 0.05 | 0.03 | -0.26 | 0.01 | 0.19 | -0.19 | -0.14 | 0.05 | -0.27 | -0.06 | 0.14 | 0.13 | — | | | | | |
| Ap01 | -0.10 | 0.32 | -0.29 | 0.02 | 0.31 | 0.17 | -0.21 | 0.54 | -0.13 | 0.15 | 0.27 | -0.13 | 0.10 | -0.21 | -0.41 | — | | | | |
| Ap03 | 0.30 | 0.54 | -0.23 | -0.08 | 0.00 | -0.11 | 0.25 | -0.19 | 0.29 | -0.14 | 0.25 | 0.08 | -0.30 | -0.13 | -0.12 | 0.02 | — | | | |
| Ap04 | 0.10 | 0.08 | -0.19 | 0.07 | -0.09 | -0.32 | -0.06 | -0.12 | -0.12 | -0.17 | 0.05 | 0.26 | -0.34 | 0.09 | -0.06 | 0.09 | 0.16 | — | | |
| Ap05 | -0.22 | -0.04 | 0.03 | -0.15 | -0.43 | -0.06 | -0.28 | -0.09 | 0.10 | -0.06 | -0.44 | -0.25 | -0.16 | 0.10 | 0.06 | -0.14 | -0.25 | -0.03 | — | |

Fig. 4. Correlation Matrix of Rasch analysis.

obtaining a high score on a difficult question. To evaluate the discrimination index, the sample of students was divided into a group with the worst performance (36%, $N = 14$) and a group with the best performance (33%, $N = 13$) based on the total scores obtained. The average discrimination index was 0.19, with a standard deviation of 0.14.

The competency map of Wright showed a good distribution throughout the competency scale, with more significant clustering around the mean (see Fig. 6). Redundancy was also observed in four questions of the same dimension (E02 and E03, A04 and A05, E01, and E04, A03 and A09), as well as in two pairs of questions of different dimensions (A02 and Ap05, A08 and Ap04). In addition, gaps in the scale were identified around extreme difficulty levels (± 2), which suggests that the questionnaire might not have sufficient capacity to discriminate students with extreme performance. No groupings of students were observed at the extremes that could be attributed to the ceiling or floor effects of the questionnaire.

4 Discussion

The questions generated by ChatGPT followed a normal difficulty index distribution with an average close to the ideal value of 0.5. Thirty percent of the questions had a discrimination index above the excellent value of 0.3. Subject confidence was slightly less than the ideal value 0.8. The p-value for model fit was lower than the excellent value of 0.05. Some correlations between the questions were higher than the ideal value of 0.3. The person-question map of Wright followed the perfect smooth curve with no outliers. Only one question had an outfit value above the ideal range of 1.3. A previous investigation showed that ChatGPT generated easy questions for medical physiology based on the assessment of experts using a 3-point scale [8]. This result is lower than the median rate of difficulty assigned by our experts using a similar scale. Our results showed no consistency among evaluators regarding the clarity and level of thinking of questions. Training experts and the supporting explanation included in the document for evaluation could improve those aspects. For future work, index values for Moodle could also be integrated for analysis.

| Item | Difficulty Index | Ability | Ability S.E. | Infit | Outfit | Discrimination Index |
|------|------------------|---------|--------------|-------|--------|----------------------|
| A01 | 0.97 | -3.86 | 1.02 | 1.02 | 0.95 | 0.00 |
| E01 | 0.77 | -1.33 | 0.40 | 0.92 | 0.84 | 0.26 |
| E02 | 0.21 | 1.49 | 0.41 | 0.97 | 1.10 | 0.21 |
| E03 | 0.21 | 1.49 | 0.41 | 0.89 | 0.81 | 0.31 |
| E04 | 0.74 | -1.18 | 0.38 | 1.05 | 1.05 | 0.05 |
| E05 | 0.62 | -0.53 | 0.35 | 0.90 | 0.86 | 0.41 |
| A02 | 0.33 | 0.77 | 0.36 | 0.98 | 1.02 | 0.31 |
| A03 | 0.87 | -2.09 | 0.49 | 0.95 | 0.82 | 0.15 |
| A04 | 0.54 | -0.18 | 0.34 | 0.98 | 0.98 | 0.36 |
| A05 | 0.56 | -0.29 | 0.34 | 0.99 | 0.97 | 0.26 |
| A06 | 0.72 | -1.04 | 0.37 | 0.98 | 0.94 | 0.26 |
| A07 | 0.44 | 0.28 | 0.34 | 1.12 | 1.16 | 0.10 |
| A08 | 0.41 | 0.40 | 0.34 | 0.99 | 1.00 | 0.15 |
| A09 | 0.87 | -2.09 | 0.49 | 1.04 | 0.99 | 0.10 |
| A10 | 0.08 | 2.69 | 0.61 | 1.06 | 2.04 | -0.05 |
| Ap01 | 0.95 | -3.13 | 0.74 | 0.91 | 0.50 | 0.05 |
| Ap03 | 0.79 | -1.49 | 0.41 | 0.89 | 0.75 | 0.36 |
| Ap04 | 0.41 | 0.40 | 0.34 | 1.02 | 1.01 | 0.31 |
| Ap05 | 0.33 | 0.77 | 0.36 | 1.28 | 1.41 | -0.05 |

Fig. 5. Analysis of the difficulty index, ability measurements, standard errors of ability estimation, infit and outfit values, and the discrimination index for each question in the diagnostic exam. The questions are categorized by their respective code, including ‘A’ for analysis, ‘E’ for evaluation, and ‘Ap’ for application questions.

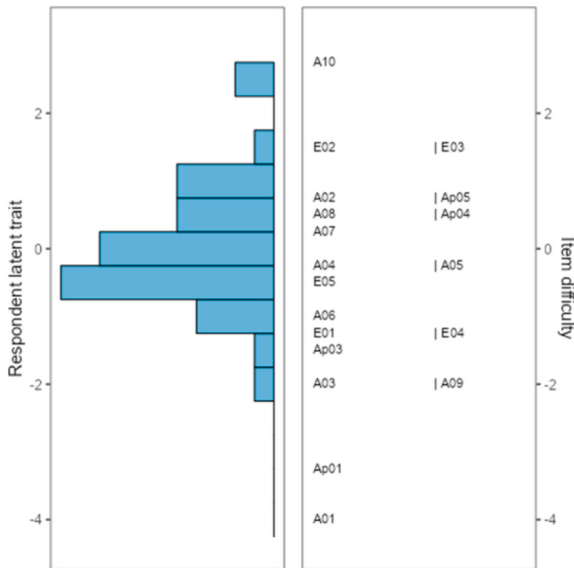


Fig. 6. Person-question map of Wright shows the relative position of questions and students. The Scale increases vertically from less capable student less capable/more straight-forward question to more skilled/more difficult question.

The use of virtual agents in academia raises validity issues. To ensure the responsible use of these tools, it is necessary to reach a consensus on regulating their use to prevent possible abuse and guarantee academic integrity. Virtual agents should not be used for cheating but to enhance personalized learning and provide new opportunities for

scholars and students. Teachers can integrate these tools into their teaching by creating new exercises that encourage critical thinking and problem-based learning, using AI to support and enrich the educational process [3]. Potential advantages of virtual agents for creating exams are their availability at any time, scalability to respond to significant work demands in less time and effortlessly, and customization to individual needs with high variability.

Our results showed that it is necessary to conduct tests and studies to determine the quality of the information these tools provide. Academics and students should always verify the content validity of the questions with these tools. Scholars and students who use virtual agents must be transparent about their use and take responsibility for the veracity of the information [10, 11]. As we can see, more work is still needed to improve the performance of virtual agents for academic use, [4] especially in specific areas such as biomedical engineering. Finally, we consider it essential to educate teachers and students about the limitations of virtual agents and how to use them effectively [3]. Giving scholars the authority and independence to use them is also essential [12]. We also consider it necessary to guarantee that everyone has the same access to these technologies [13].

5 Conclusion

In conclusion, using artificial intelligence and virtual agents based on GPT language models in academia raises several practical and ethical considerations. Implementing clear guidelines, a code of ethics, and consensus in regulating its use are required. Academics must be transparent in their use, assuming responsibility for the information generated. Literacy in this area, constant evaluation, and continuous improvement are essential to maximize the benefits and minimize the risks associated with these technologies.

References

1. D'Amico, R.S., White, T.G., Shah, H.A., et al.: I asked a ChatGPT to write an editorial about how we can incorporate chatbots into neurosurgical research and patient care. *Neurosurgery* **92**, 663–664 (2023)
2. Sallam, M.: ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* **11**, 887 (2023)
3. Yang, H.: How I use ChatGPT responsibly in my teaching. *Nature* (2023). <https://doi.org/10.1038/d41586-023-01026-9>
4. Das, D., Kumar, N., Longjam, L.A., et al.: Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. *Cureus* **15**, 3–11 (2023)
5. Ariyaratne, S., Iyengar, K.P., Nischal, N., et al.: A comparison of ChatGPT-generated articles with human-written articles. *Skelet. Radiol.* **52**, 1755–1758 (2023)
6. Eysenbach, G.: The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med. Educ.* **9**, e46885 (2023)

7. Sedaghat, S.: Early applications of ChatGPT in medical practice, education and research. *Clin. Med. (Northfield Il)* **23**, 278–279 (2023)
8. Agarwal, M., Sharma, P., Goswami, A.: Analysing the applicability of ChatGPT, bard, and bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus* (2023). <https://doi.org/10.7759/cureus.40977>
9. The Jamovi Project. Jamovi (2023). <https://www.jamovi.org>
10. Gurha, P., Ishaq, N., Marian, A.J.: ChatGPT and other artificial intelligence chatbots and biomedical writing. *J. Cardiovasc. Aging* 3–5 (2023)
11. Hosseini, M., Horbach, S.P.J.M.: Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res. Integr.* 1–14 (2023)
12. Anders, B.A.: Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking? *Patterns* **4**, 100694 (2023)
13. Salvagno, M., Taccone, F.S., Gerli, A.G.: Can artificial intelligence help for scientific writing? *Crit. Care* **27**, 75 (2023)