

Embracing Large Language Models for Adult Life Support Learning

Serena Patel ¹, Rohit Patel ²

1. General Surgery, Imperial College NHS Trust, Ilford, GBR 2. Oral and Maxillofacial Surgery, Kings College Hospital, London, GBR

Corresponding author: Serena Patel, patelser96@gmail.com

Review began 12/13/2024

Review ended 12/16/2024

Published 12/18/2024

© Copyright 2024

Patel et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.75961

Abstract

Background

It is recognised that large language models (LLMs) may aid medical education by supporting the understanding of explanations behind answers to multiple choice questions. This study aimed to evaluate the efficacy of LLM chatbots ChatGPT and Bard in answering an Intermediate Life Support pre-course multiple choice question (MCQs) test developed by the Resuscitation Council UK focused on managing deteriorating patients and identifying causes and treating cardiac arrest. We assessed the accuracy of responses and quality of explanations to evaluate the utility of the chatbots.

Methods

The performance of the AI chatbots ChatGPT-3.5 and Bard were assessed on their ability to choose the correct answer and provide clear comprehensive explanations in answering MCQs developed by the Resuscitation Council UK for their Intermediate Life Support Course. Ten MCQs were tested with a total score of 40, with one point scored for each accurate response to each statement a-d. In a separate scoring, questions were scored out of 1 if all sub-statements a-d were correct, to give a total score out of 10 for the test. The explanations provided by the AI chatbots were evaluated by three qualified physicians as per a rating scale from 0-3 for each overall question and median rater scores calculated and compared. The Fleiss multi-rater kappa (κ) was used to determine the score agreement among the three raters.

Results

When scoring each overall question to give a total score out of 10, Bard outperformed ChatGPT although the difference was not significant ($p=0.37$). Furthermore, there was no statistically significant difference in the performance of ChatGPT compared to Bard when scoring each sub-question separately to give a total score out of 40 ($p=0.26$). The qualities of explanations were similar for both LLMs. Importantly, despite answering certain questions incorrectly, both AI chatbots provided some useful correct information in their explanations of the answers to these questions. The Fleiss multi-rater kappa was 0.899 ($p<0.001$) for ChatGPT and 0.801 ($p<0.001$) for Bard.

Conclusions

The performances of both Bard and ChatGPT were similar in answering the MCQs with similar scores achieved. Notably, despite having access to data across the web, neither of the LLMs answered all questions accurately. This suggests that there is still learning required of AI models in medical education.

Categories: Medical Education, Healthcare Technology

Keywords: advanced life support, bard, basic life support bls, chatgpt, google bard, large language models (llm), resus

Introduction

The advent of artificial intelligence (AI) since the 20th century has resulted in the development of machines with the capabilities of executing human intelligence processes to enable problem-solving and simplification of tasks. Since their conception, machine learning algorithms and neural networks have become more advanced with technologies transforming the way we work in industries ranging from manufacturing to medicines to finance [1]. Within medicine, the evolution of neural networks and deep learning has led to automated clinical image recognition and interpretation, offering the potential to transform radiomics and aid histopathological diagnosis [2,3] with high sensitivity and high specificity [4]. In the context of clinical education, AI has evolved to enable virtual patient systems [5], virtual-reality and robotic surgical simulation applications [6], and to support distance learning. Indeed, with medicine becoming increasingly digitalised, harnessing the power of AI to augment learning is crucial to ensure that students acquire the digital literacy needed to incorporate technology into their future practice and effectively utilize the data resources available to them [7].

How to cite this article

Patel S, Patel R (December 18, 2024) Embracing Large Language Models for Adult Life Support Learning. Cureus 16(12): e75961. DOI 10.7759/cureus.75961

Both ChatGPT and Bard are large language model (LLM) generative AI chatbots trained on large datasets to recognize, predict, translate, or generate text. As such, they can be trained to problem-solve through document summarisation, text classification, question answering, and text generation. Their use in clinical education has been suggested in medical answering and reasoning, assignment feedback, simulated patient chatbots, and learning clinical reasoning [8]. ChatGPT, also known as Chat Generative Pre-trained Transformer, developed by OpenAI and originally released to the public in November 2022 (based on data pre-September 2021), has different versions trained on different datasets. It has already been shown to have a role in supporting healthcare professionals with medical education and medical writing [9]. Bard is a newer chatbot, developed and released by Google in March 2023 [10], and has real-time access to data through Google's search engine. Similar to ChatGPT, Bard has the ability to generate text with conversational responses with a potential to impact medical education.

The Resuscitation Council UK [11] is a national charitable organisation responsible for setting standards for cardiopulmonary resuscitation (CPR) and related disciplines in the UK. In addition, the organisation is responsible for providing mandatory life support training courses including intermediate and advanced life support for healthcare professionals including doctors, dentists, nurses, and midwives for the recognition and management of deteriorating patients and appropriate response to cardiac arrest events. Completion of life support qualifications is mandatory for clinical medical students and junior doctors and courses involve pre-reading as well as an in-person or e-learning course that delegates must do to achieve certification. Several of these clinical courses provide pre-course test questions to aid student learning and post-course tests that delegates must pass for certification.

Previous studies have assessed the utility of LLMs in answering clinical medical questions in various disciplines (pharmacology, microbiology, pathology, clinical knowledge) in medical licensing examinations [12–14]. This study aimed to evaluate the efficacy of LLM ChatGPT and Bard in answering resuscitation-based medical questions, relating to causes and management of cardiac arrest. We assessed the accuracy of responses and the quality of explanations to evaluate the utility of LLMs as educational tools to support healthcare professionals in their preparation for life support courses and their understanding of the assessment and management of unwell patients, including response in cardiac arrest.

Materials And Methods

Artificial Intelligence

ChatGPT-3.5 (Open AI, San Francisco, USA) and Bard (Google, Mountain View, USA), both freely accessible to the public, were the AI chatbots utilised in this study. ChatGPT used the Generative Pre-Trained Transformer 3.5 whilst Bard utilised the Language Model for Dialogue Applications (LaMDA). At the time of this study, ChatGPT-3.5 was trained and confined to datasets up to 2021, limiting access to the latest research. In contrast, Bard had the capability of searching the Internet in real-time to find the most recent research on a topic. Both AI chatbots allow text input and provide text output to answer questions and provide explanations.

Test questions

The performance of the AI chatbots was tested on true-false test questions taken from an Intermediate Life Support pre-course question test developed by the Resuscitation Council UK. Thus, all inputs represent true training questions for cardiopulmonary resuscitation teaching in the United Kingdom and are essential knowledge for the performance of cardiopulmonary resuscitation by healthcare professionals in the United Kingdom. The main questions had a similar structure each consisting of a stem with four true-false sub-questions (a-d), with the chatbots asked to choose all correct options and then confirm. These constituted 10 main questions, which with the sub-questions totalled 40 true-false questions.

Testing

All inputs into LLMs occurred on 4th October 2023 by a single tester with the other author observing. Both ChatGPT and Bard were given the instruction to 'act as a medical doctor' to answer the questions. The questions were submitted by reproducing the original Resuscitation Council UK's question verbatim. For each question, the chatbot was advised to 'Choose ALL correct options and then confirm' as in the original question. The answers selected by the chatbots were compared to the true answers to the questions on the Resuscitation Council UK's pre-course test paper. Scoring occurred in two ways. The 40 sub-questions were tested to give a total score out of 40, with 1 point scored for each accurate response to each statement a-d. In a separate scoring, each of the 10 main questions was scored 1 if all the responses to sub-questions a-d were correct, to give a total score out of 10.

The explanations provided by the AI chatbots were evaluated by three qualified physicians as per a rating scale from 0–3 for each overall question. All physicians have medical degrees from the United Kingdom, hold advanced life support certification from the Resuscitation Council UK, and have experience teaching undergraduate medicine. A score of zero denoted the explanation was wholly incorrect, providing no insight or knowledge; a score of one was used when the explanation had some correct or useful information but lacked significant detail or contained errors; a score of two was used when the explanation was correct and

relevant but lacked detail or did not provide a sufficient nuanced explanation for all options; a score of three was given for clear, comprehensive, and correct explanations which sufficiently explained all of the options for each question. The median score for each question was calculated and rounded to the nearest integer for analysis.

Ethical approval

Ethical approval was not required for this study as there were no animal or human subjects.

Data analysis

Data was inputted into Microsoft Excel 365 (Microsoft Corporation, Redmond, USA). Statistical analysis was conducted using R (version 4.3.1, R Foundation for Statistical Computing, Vienna, Austria). The Chi-squared non-parametric test was used to determine the association of correct answers between ChatGPT and Bard. The Fleiss kappa test was used to assess score agreement between the three raters. Results were statistically considered statistically significant if $p < 0.05$.

Results

Accuracy of answers

In assessing the accuracy of correct answers to the questions, the performances of both Bard and ChatGPT were noted to be similar. When scoring the main questions to give a total score out of 10, Bard outperformed ChatGPT although the difference was not significant (6 vs 4, $p=0.37$). Furthermore, there was no statistically significant difference in the performance of ChatGPT compared to Bard when scoring each sub-question separately to give a total score out of 40 (30 vs 34, $p=0.26$). The total scores achieved by both ChatGPT and Bard on the tests are shown in Figure 1.

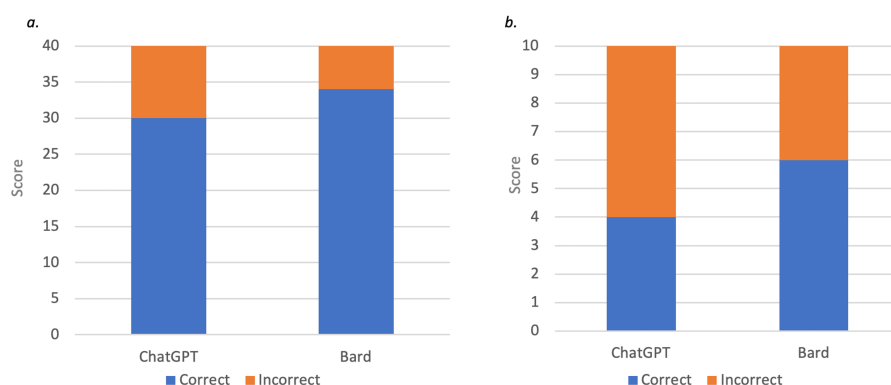


FIGURE 1: Comparison of performance of ChatGPT and Bard when scoring a) out of 40 b) out of 10.

Interestingly, of the 10 main questions, all questions answered incorrectly by Bard were also answered incorrectly by ChatGPT, with additional questions also answered incorrectly by ChatGPT. Four questions were answered correctly by both LLMs, four questions were answered incorrectly by both LLMs, and two questions were answered correctly by Bard and incorrectly by ChatGPT. Figure 2 shows the correct and incorrect responses by main question number for each of the LLMs tested.

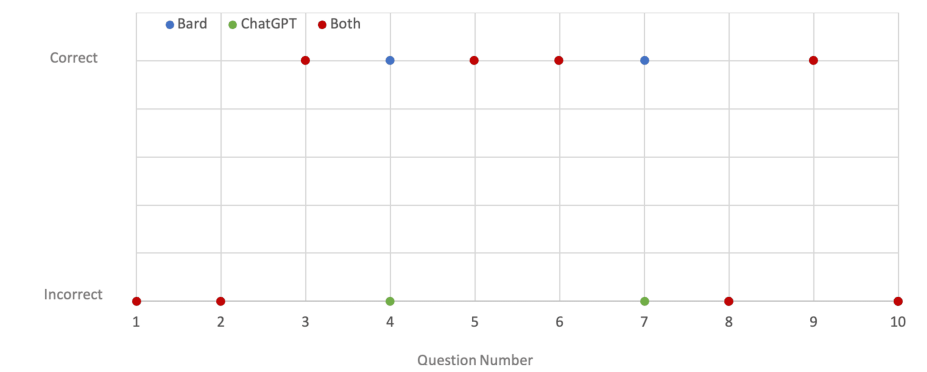


FIGURE 2: Correct and incorrect responses by question number

When considering the 40 sub-questions, a similar pattern was noted. All questions answered incorrectly by Bard were also answered incorrectly by ChatGPT, and additional questions that were correct by Bard were incorrect by ChatGPT. Thirty questions were answered correctly by both LLMs, six were answered incorrectly by both LLMs, and four were answered correctly by Bard but incorrectly by ChatGPT. This is shown in Figure 3.

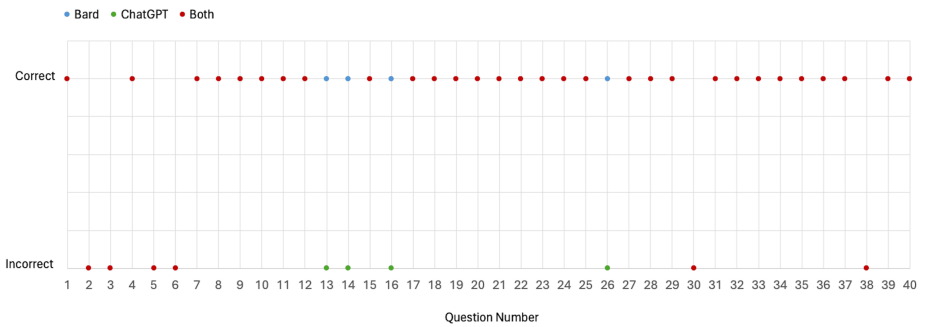


FIGURE 3: Correct and incorrect responses for each of the 40 sub-questions

Quality of explanations

Explanations provided by ChatGPT and Bard for each main question were scored out of 3. The median rater scores for each question are shown in Figure 4. Both ChatGPT and Bard had the same median rater score for five questions, four of these had a median rater score of 3, and one had a median rater score of 2. In three cases, the median rater score was higher for ChatGPT than Bard and in two cases, the score was higher for Bard than ChatGPT.

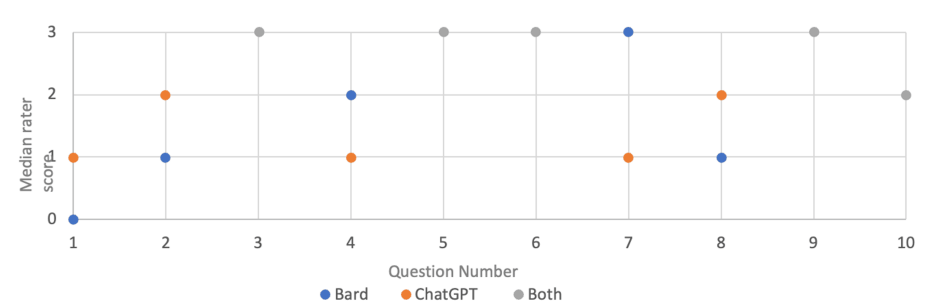


FIGURE 4: Median rater scores for each chatbot for each main question

Important to note is that despite answering certain questions incorrectly, both AI chatbots provided some useful correct information in their explanations of the answers to these questions. The Fleiss multi-rater kappa was 0.899 ($p < 0.001$) for ChatGPT and 0.801 ($p < 0.001$) for Bard. This shows strong inter-rater

reliability for both AI chatbots in providing explanations for the answers to the questions.

Figure 5 shows the frequency of median rater scores for each Chatbot across the questions. Five questions answered by Bard had a median rater score of 3, two had a median score of 2, two had a median score of 1, and one had a median rater score of 0. In contrast, four of the explanations provided by ChatGPT had a median rater score of 3, three had a median score of 2, three had a score of 1, and none had a score of 0.

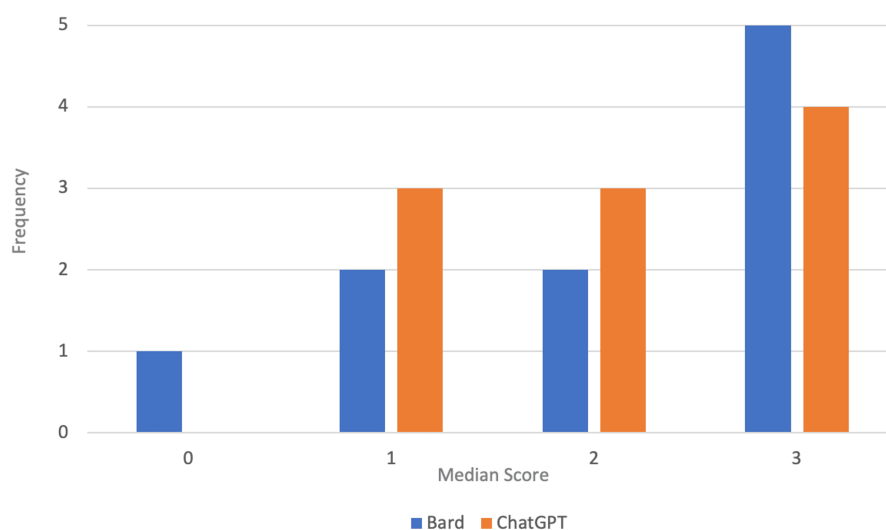


FIGURE 5: Frequency of median rater scores for each chatbot for the main questions

Discussion

The potential for ChatGPT across medical education has been noted [5-7]. It has been suggested to facilitate evidence-based decision-making and generate differential diagnoses triggering self-directed learning on diagnoses unfamiliar to the student and as a knowledge resource [8]. In the context of exam preparation, ChatGPT has been suggested to generate level-appropriate Objective Structured Clinical Examination (OSCE) scenarios (although notably marking domains and learning outcomes differ between institutions, and therefore it is hard to achieve reliability, validity, and calibration), even providing answers if prompted [15], and assist as a ‘virtual teaching assistant’ providing feedback and clarifying concepts [16].

This study aimed to assess the accuracy of responses and quality of explanations of Bard and ChatGPT in answering medical questions presented in the Resuscitation Council UK’s ILS paper focused on managing deteriorating patients and identifying causes and treating cardiac arrest. The performances of both Bard and ChatGPT were similar in answering the MCQs with similar scores achieved. Notably, despite having access to data across the web, neither of the LLMs answered all questions accurately. This suggests that there is still learning required of AI models in medical education. Evaluation of the explanations provided by the LLMs found that both Bard and ChatGPT were similar in the quality of explanation provided, with the majority of the explanations providing correct or useful information to aid learning of medical information. Further to this, it was noted that even in cases where the AI chatbots answered questions incorrectly, they were able to provide useful insights in their explanations.

Previous studies have sought to investigate the role of LLMs in other contexts of medical testing. A systematic review showed the impressive performance of generative AI models on standardised tests [17], including in the United States Medical Licensing Examination [18] and in specialised tests ranging from cardiology [19] to biochemistry [20] and general surgery [21]. However, there were cases where models failed exams such as the Taiwanese Family Medicine Board Exam [22]. In other studies, ChatGPT was shown to correctly answer two-thirds of all questions at a German state licensing exam level Progress Test [23] and performed at or near a passing level of 60% on a test of 350 items of the United States Medical Licensing Examination (USMLE) [18]. In other studies of medical physiology, ChatGPT secured around 75% [24,25]. In scoring each stem separately in our study, ChatGPT scored similarly at 75%. Although the difference was not noted to be significant, Bard scored 85% in our study.

The comparative performance of different chatbots in answering medical questions has been inconsistent in previous studies. ChatGPT outperformed Bard in a study of lung cancer questions [26] and neurosurgery board questions [27], which differed from our study. It is possible that performance may differ according to different subspecialties of medicine depending on the training algorithms the AI chatbots have been

exposed to. Furthermore, a separate study by Kung et al. has shown that the performance of AI chatbots such as ChatGPT can vary depending on the exact prompt inputted for each question (e.g. forced justification) [18]. Thus, the choice of words for prompts could explain the differences in AI performance. Furthermore, the complexity of the input can further influence the performance accuracy of each of the LLMs in answering medical questions [9]. Improvement of AI models through training may allow them to be more accurate to further support learning.

The overall quality of the explanations provided by the LLMs for each question was found to be similar in our study. Interestingly, despite answering some questions incorrectly, the LLMs were able to provide useful and relevant information to aid in understanding the correct answers to these questions. This suggests that AI can support students in understanding material for questions even if answers are inaccurate; the challenge lies in students relying on AI to provide the correct answer without applying their own critical thinking as this may hinder understanding, particularly in cases where it may be used by students dishonestly to obtain answers. In fact, these AI LLMs may be more appropriately used to support the learning process and problem-solving without providing the answers, instead giving students information that guides them to deduce an answer for themselves. This supports constructivist theories that learning occurs as students form connections between new information to previous knowledge allowing them to build an understanding [28]. However, a further problem that students may face is that LLMs can often convincingly justify incorrect responses, in a phenomenon known as “hallucination” [17] limiting their utility for medical education. On the other hand, facing incorrect responses allows students to problem-solve and question the LLM responses, and through further discussion with educators students are able to develop their own knowledge.

Our study supports the notion that LLMs can be increasingly utilised across the medical education curriculum, providing students with vast amounts of knowledge. However, in order to use LLMs appropriately and ensure students are still able to acquire the knowledge and understanding they need, students will benefit from engaging with educators who can guide students through complex concepts needed to answer medical questions and extend students’ learning beyond what they can achieve through independent use of LLMs. This is underpinned by Vygotsky’s social-cultural theory, which emphasises the importance of the cultural conditions under which knowledge is developed and places importance on the teacher-student relationship [29]; the purpose of tutor discussion is to capitalise on the fact that there is a difference between what a learner can learn independently and what they can learn with an experienced person and this difference is identified as Vygotsky’s Zone of Proximal Development [29,30]. Considering this in the context of our study, educators on Resuscitation Council courses would be able to support students’ learning in addition to the use of LLMs.

Limitations

This study had notable limitations. This study was based on a small sample size of 40 true-false questions and 10 main questions from an ILS pre-course paper from the Resuscitation Council UK. These questions were selected since the Resuscitation Council UK’s life support courses are one of the only mandatory courses required for all UK medical doctors and thus assess the standard knowledge expected of all practicing physicians in the United Kingdom. However, the limited data set is arguably insufficient to comprehensively cover all aspects of resuscitation-based training. Also, whilst the questions are based on this topic they may only be representative of UK guidelines for management. Further studies with larger data sets are needed to comprehensively address this topic, including advanced life support and advanced cardiac life support guidelines. Our study focused only on ChatGPT-3.5 and Bard although other LLMs also exist including Claude-2 and Microsoft Bing. Furthermore, only the free version of ChatGPT - Chat GPT3.5 - was accessed for this study, and it is possible that ChatGPT-4 may perform differently in this setting, especially given its live internet browsing capabilities since mid-2023 [31]. A previous study of the Japanese licensing exam has shown that ChatGPT-4 outperformed ChatGPT-3.5 [32]. Additionally, although inter-rater reliability was noted to be strong, there may be subjectivity and human error in the evaluation of the explanations presenting a further limitation of our study.

Conclusions

Our study provides insight into the suitability of AI LLMs to support students in learning concepts related to cardiopulmonary resuscitation. The results indicate a reasonable performance of ChatGPT and Bard in answering relevant questions but more importantly the ability of these LLMs to provide correct and insightful explanations that can aid learning. However, currently, these AI models require further training to improve their accuracy. The results of our study should be interpreted with caution given the limitations of the research design. As LLM technology evolves, future studies must be designed to address the limitations and provide further results that can be generalised in medical education.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Serena Patel, Rohit Patel

Acquisition, analysis, or interpretation of data: Serena Patel, Rohit Patel

Drafting of the manuscript: Serena Patel, Rohit Patel

Critical review of the manuscript for important intellectual content: Serena Patel, Rohit Patel

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

We are thankful to Dr Nitya Matcha and Dr Siddhartha Matcha for their help in evaluating the responses. Data are available on reasonable request.

References

1. Rane NL, Mallick SK, Kaya Ö, Rane J: Chapter 7. Applications of deep learning in healthcare, finance, agriculture, retail, energy, manufacturing, and transportation: a review. *Applied Machine Learning and Deep Learning: Architectures and Techniques*. Rane NL, Mallick SK, Kaya Ö, Rane J (ed): Deep Science Publishing, Mumbai; 2024. [10.70593/978-81-981271-4-3](#)
2. Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A: Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat Rev Clin Oncol*. 2022, 19:132-46. [10.1038/s41571-021-00560-7](#)
3. Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M: Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun*. 2016, 7:12474. [10.1038/ncomms12474](#)
4. Gulshan V, Peng L, Coram M, et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016, 316:2402-10. [10.1001/jama.2016.17216](#)
5. Mir MM, Mir GM, Raina NT, et al.: Application of artificial intelligence in medical education: current scenario and future perspectives. *J Adv Med Educ Prof*. 2023, 11:133-40. [10.30476/JAMP.2023.98655.1803](#)
6. Mandler AG: Touch surgery: a twenty-first century platform for surgical training. *J Digit Imaging*. 2018, 31:585-90. [10.1007/s10278-018-0102-y](#)
7. Jamal A, Solaiman M, Alhasan K, Temsah MH, Sayed G: Integrating ChatGPT in medical education: adapting curricula to cultivate competent physicians for the AI era. *Cureus*. 2023, 15:e43036. [10.7759/cureus.43036](#)
8. Ahn S: The impending impacts of large language models on medical education. *Korean J Med Educ*. 2023, 35:103-7. [10.3946/kjme.2023.253](#)
9. Dave T, Athaluri SA, Singh S: ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023, 6:1169595. [10.3389/frai.2023.1169595](#)
10. Google begins opening access to its ChatGPT competitor Bard | Reuters. (2023). Accessed: December 17, 2023: <https://www.reuters.com/technology/google-begins-opening-access-its-chatgpt-competitor-bard-2023-03-21/#:~:text=March....>
11. Home | Resuscitation Council UK. (2024). Accessed: November 11, 2023: <https://www.resus.org.uk/>.
12. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D: How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023, 9:e45312. [10.2196/45312](#)
13. Wang X, Gong Z, Wang G, et al.: ChatGPT performs on the Chinese National Medical Licensing Examination. *J Med Syst*. 2023, 47:86. [10.1007/s10916-023-01961-0](#)
14. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, Knoedler L: Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. *Ann Biomed Eng*. 2024, 52:1542-5. [10.1007/s10439-023-03338-3](#)
15. Tsang R: Practical applications of ChatGPT in undergraduate medical education. *J Med Educ Curric Dev*. 2023, 10:23821205231178449. [10.1177/23821205231178449](#)
16. Lee H: The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. 2024, 17:926-31. [10.1002/ase.2270](#)
17. Preiksaitis C, Rose C: Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR Med Educ*. 2023, 9:e48785. [10.2196/48785](#)
18. Kung TH, Cheatham M, Medenilla A, et al.: Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023, 2:e0000198. [10.1371/journal.pdig.0000198](#)
19. Skolidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S: ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story?. *Eur Heart J Digit Health*. 2023, 4:279-81. [10.1093/ehjdh/ztad029](#)
20. Ghosh A, Bir A: Evaluating ChatGPT's ability to solve higher-order questions on the competency-based

- medical education curriculum in medical biochemistry. *Cureus*. 2023, 15:e37023. [10.7759/cureus.37023](https://doi.org/10.7759/cureus.37023)
21. Oh N, Choi GS, Lee WY: 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*. 2023, 104:269-73. [10.4174/asttr.2023.104.5.269](https://doi.org/10.4174/asttr.2023.104.5.269)
22. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ: ChatGPT failed Taiwan's Family Medicine Board Exam . *J Chin Med Assoc*. 2023, 86:762-6. [10.1097/JCMA.0000000000000946](https://doi.org/10.1097/JCMA.0000000000000946)
23. Friederichs H, Friederichs WJ, März M: ChatGPT in medical school: how successful is AI in progress testing? . *Med Educ Online*. 2023, 28:2220920. [10.1080/10872981.2023.2220920](https://doi.org/10.1080/10872981.2023.2220920)
24. Dhanvijay AK, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H: Performance of large language models (ChatGPT, Bing Search, and Google Bard) in solving case vignettes in physiology. *Cureus*. 2023, 15:e42972. [10.7759/cureus.42972](https://doi.org/10.7759/cureus.42972)
25. Banerjee A, Ahmad A, Bhalla P, Goyal K: Assessing the efficacy of ChatGPT in solving questions based on the core concepts in physiology. *Cureus*. 2023, 15:e43314. [10.7759/cureus.43314](https://doi.org/10.7759/cureus.43314)
26. Rahsepar AA, Tavakoli N, Kim GH, Hassani C, Abtin F, Bedayat A: How AI responds to common lung cancer questions: ChatGPT vs Google Bard. *Radiology*. 2023, 307:e230922. [10.1148/radiol.230922](https://doi.org/10.1148/radiol.230922)
27. Ali R, Tang OY, Connolly ID, et al.: Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*. 2023, 93:1090-8. [10.1227/neu.0000000000002551](https://doi.org/10.1227/neu.0000000000002551)
28. Taylor DC, Hamdy H: Adult learning theories: implications for learning and teaching in medical education: AMEE Guide No. 83. *Med Teach*. 2013, 35:e1561-72. [10.3109/0142159X.2013.828153](https://doi.org/10.3109/0142159X.2013.828153)
29. Vygotsky LS, Cole M: *Mind in Society: Development of Higher Psychological Processes* . Cole M, John-Steiner V, Scribner S, Souberman E (ed): Harvard University Press, Cambridge; 1978.
30. Chaiklin S: The zone of proximal development in Vygotsky's theory of learning and school instruction . *Vygotsky's Educational Theory in Cultural Context*. Kozulin A, Gindis B, Ageyev VS, Miller SM (ed): Cambridge University Press, Cambridge; 2003. 39:64.
31. Temsah MH, Jamal A, Aljamaan F, Al-Tawfiq JA, Al-Eyadhy A: ChatGPT-4 and the Global Burden of Disease Study: advancing personalized healthcare through artificial intelligence in clinical and translational medicine. *Cureus*. 2023, 15:e39384. [10.7759/cureus.39384](https://doi.org/10.7759/cureus.39384)
32. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M: Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: evaluation study. *JMIR Form Res*. 2023, 7:e48023. [10.2196/48023](https://doi.org/10.2196/48023)