Journal of CME

ISSN: 2833-8073 (Online) Journal homepage: www.tandfonline.com/journals/zjec21

Chatbot Underperformance in Biology and Image-Based Questions in Medical Education

Joyce Santana Rizzi, Lorraine Silva Requena, Angelica Maria Bicudo, Pedro Tadao Hamamoto Filho & Renato Ferretti

To cite this article: Joyce Santana Rizzi, Lorraine Silva Requena, Angelica Maria Bicudo, Pedro Tadao Hamamoto Filho & Renato Ferretti (2025) Chatbot Underperformance in Biology and Image-Based Questions in Medical Education, Journal of CME, 14:1, 2596550, DOI: 10.1080/28338073.2025.2596550

To link to this article: https://doi.org/10.1080/28338073.2025.2596550

Published online: 04 Dec 2025.

JOURNAL OF CME
2025, VOL. 14, 2596550
https://doi.org/10.1080/28338073.2025.2596550

Chatbot Underperformance in Biology and Image-Based Questions in Medical Education

Joyce Santana Rizzi [a], Lorraine Silva Requena [a], Angelica Maria Bicudo [b], Pedro Tadao Hamamoto Filho [c] and Renato Ferretti [a]

[a]Laboratory of Muscle Biology, Department of Structural and Functional Biology, Institute of Bioscience of Botucatu, Sao Paulo State University (UNESP), Botucatu, Brazil; [b]School of Medical Sciences, University of Campinas (UNICAMP), Campinas, Brazil; [c]Botucatu Medical School, Department of Neurosciences and Mental Health, São Paulo State University (UNESP), Botucatu, Brazil

ABSTRACT

AI chatbots have demonstrated variable performances across biological disciplines in medical education, particularly in multiple-choice and image-based assessments. However, their performance in addressing discipline-specific and image-based questions in biology remains unexamined. This study evaluated the accuracy and reliability of chatbots in answering biological questions from the Progress Test, a medical assessment applied across ten universities. We conducted an observational cross-sectional study by inputting 180 questions into the chatbots and categorising them according to morphology, function, and aggression. Each question was assessed for correctness across multiple chatbot attempts, and logistic regression and hierarchical clustering were applied to identify performance patterns. Although the chatbots answered functional and morphological questions accurately (from 85% (Gemini) to 91.7% (ChatGPT-4)), their accuracy decreased significantly for questions involving biological aggression and visual content. The agreement between chatbot responses remained weak, and Co-pilot displayed the lowest concordance. Chatbot accuracy decreased significantly in aggression-related disciplines and image-based questions. Logistic regression confirmed that the presence of images reduced the odds of correct answers by up to 17.6% (ChatGPT-4). Hierarchical clustering distinguished the two distinct response patterns, further validating these findings. These results highlight the potential of chatbots in medical education while emphasising their limitations in handling image-based and aggression-related content.

ARTICLE HISTORY
Received 6 August 2025
Accepted 20 November 2025

KEYWORDS
Undergraduate medical education; large language model; artificial intelligence; progress test; biological science disciplines

Introduction

The development of chatbots represents a collaborative effort between human expertise and artificial intelligence (AI), resulting in powerful tools for solving problems in morphological and functional disciplines [1–4]. These systems support higher-level thinking, interpretation, analysis, evaluation, and the formulation of evidence-based predictions [5]. However, their precision and reliability are contingent on specific field-related and image-based factors, making accuracy and efficacy crucial considerations in medical education [6,7]. Research on chatbots is extensive, encompassing studies on their design, applications, and potential across various fields in medical education [8–11].

The interdisciplinarity of biological science plays a crucial role in medical education by providing a comprehensive understanding of the human body at the macroscopic and microscopic levels [12]. A comprehensive understanding of anatomical structures, physiological functions, and host defence mechanisms against pathogens is essential for medical students to develop a deeper understanding of disease pathophysiology [6,13–16]. These interconnected domains elucidate the complex mechanisms underlying disease progression, the host's adaptive immune and physiological responses, and the development of innovative therapeutic strategies, thereby equipping medical students with a strong foundation for clinical reasoning and informed decision-making in diverse healthcare scenarios.

Alessi and colleagues demonstrate that ChatGPT-3.5 achieved a 94.1% success rate for basic science multiple-choice questions (MCQs) from the 2021 Brazilian Progress Test (PT) exam. However, their study did not analyse the model's performance across biological disciplines [17]. Chatbot models trained on extensive datasets may inherit biases, produce skewed results, and provide unreliable output. However, few studies have employed rigorous

CONTACT Renato Ferretti r.ferretti@unesp.br Laboratory of Muscle Biology, Department of Structural and Functional Biology, Institute of Bioscience of Botucatu, Sao Paulo State University (UNESP), Botucatu, São Paulo, Brazil

statistical methodologies to comprehensively assess chatbot performance in the biological sciences in medical education. This can compromise educational quality, leading to superficial understanding instead of deep learning and critical engagement with scientific concepts. Understanding these capabilities and limitations is crucial for their use in medical education [3]. To overcome this gap, our study aims to comparatively analyse the performance, accuracy, and reliability of different chatbots in response to multiple-choice questions related to biological science disciplines in undergraduate medical education.

Materials and Methods

We conducted an observational and cross-sectional study to evaluate the performance of five chatbots (Copilot, Gemini, Claude 3.5 Sonnet, ChatGPT 3.5, and ChatGPT 4.0) on questions from an inter-institutional Brazilian consortium for Progress Test examinations from 2013 to 2024. PT examinations are annual (biannual in the last three years) cross-institutional assessments designed to evaluate medical students from the first to the sixth year across ten medical schools in Brazil. The university consortium was formed by São Paulo State University, Campinas State University, and the three campuses of São Paulo University (São Paulo, Bauru, and Ribeirão Preto); São Paulo Federal University; São José do Rio Preto Medical School; Marília Medical School; São Carlos Federal University; and Londrina State University.

Each exam comprised 120 MCQs covering a broad spectrum of medical disciplines grouped into six content areas: biological sciences, internal medicine, paediatrics, surgery, obstetrics and gynaecology, and public health. These questions are structured as clinically oriented case scenarios, often supplemented with diagnostic images and data tables, which challenge students to apply their theoretical knowledge to practical medical contexts. Each question had four answers to choose from, with only one correct random response. None of the questions included patient-identifying information or personal data.

In this study, we used only MCQs covering the biological sciences. The data were further classified into three categories: morphology, encompassing anatomy, embryology, and histology; function, including physiology, biochemistry, cellular biology, genetics, and pharmacology; and aggression, including immunology, microbiology, parasitology, and pathology.

The MCQs were manually uploaded to the chatbots, and a predefined prompt was used to guide the responses: "*What is the most appropriate answer to the clinical case?*" and "*Based on the interpretation of the question, provide a justification for the correct answer as well as for the incorrect options*". This structured approach ensured a standardised evaluation process, allowing for a comprehensive assessment of the chatbot's reasoning and explanatory capabilities. Responses were categorised as correct (score = 1) or incorrect (score = 0). Each justification was evaluated to identify and eliminate the potential instances of AI-generated hallucinations. Hallucination refers to responses that is misaligned with the prompt. In this study, we classified answers as hallucinated when the chatbot refused to answer or crashed. If hallucinations were detected, the session was immediately terminated, and a new session was initiated to mitigate errors and reduce the risk of memory bias. To ensure reliability and consistency, we assessed the chatbot's performance on three separate occasions, using two computers and three different users, and tested on distinct days between 10 January 2025, and 15 February 2025 – a period during which no revisions or subsequent versions of chatbots were made available. On each evaluation day, a new batch of MCQs was introduced to maintain the integrity of the assessment and prevent potential learning effects from the previous sessions. Ethical review was not applicable to this study because it did not involve research participants.

For the analysis, we considered the following independent variables: content area and the presence of visual content such as images (X-rays, magnetic resonance imaging, computed tomography, histological photomicrographs, and anatomical images) or tables. The answer (correct or incorrect) in each attempt was the dependent variable.

Statistical Analysis

Categorical data are presented as percentages. Differences in percentages were tested using chi-square or Fisher's exact tests. Variables with significant relationships with the dependent variables (correct answers or questions correctly answered two or three times) were included in a multivariate analysis using binary logistic regression. The agreement between the chatbots was tested by considering the best attempt of each chatbot (e.g. the attempt with the highest rate of correct answers) using Cohen's kappa coefficient for paired comparisons. The Fleiss generalised kappa coefficient was used to compare all the chatbots.

Finally, the bots' answers were used to generate a hierarchical cluster analysis with an agglomeration schedule and a proximity matrix to verify the response patterns across all AI bots. We used the cluster method of between-group linkage and the

interval of measure using the squared Euclidean distance. The resulting dendrogram was used to visualise the possible clusters created in the next step. The answers of the bots were included as variables for iterations (maximum of 10 iterations and convergence criterion of 0) and classified into clusters. Finally, we compared the characteristics of the clusters according to category, presence of images, and correct answers for each bot.

Analyses were performed using the Statistical Package for the Social Sciences (SPSS), version 24.0, for MacBook (IBM Corp., Armonk, NY, USA), and GraphPad Prism for MacBook, version 9.5.0 (San Diego, CA, USA). The level of significance was set at 5%.

Results

The dataset consisted of 180 multiple-choice questions (MCQs), of which 43 (23.9%) included images. Physiology, pharmacology, and genetics were disciplines with the highest frequencies of questions. The distribution of the questions according to the primary discipline and categorisation used is shown in Table 1.

AI chatbots demonstrated strong performance in answering biological science MCQs, achieving accuracy rates between 85% (Gemini) and 91.7% (ChatGPT-4) (Table 2). However, the agreement between the chatbots' best performances was weak, with Co-pilot showing the lowest concordance. GPT-4 and GPT-3.5 exhibit better agreement (k = 0.794, substantial, $p < 0.001$) (Table 3).

Co-pilot showed no difference in correct answers according to the MCQ category, whereas in all three attempts, the performance was lower for questions with images. On the best attempt (88.3% of correct answers), 93.4% of the questions without images were correctly marked, whereas the rate of correct answers decreased to 72.1% for questions with images.

Gemini showed differences in correct answers according to category in two of the three attempts; on the best attempt, the rates of accurate responses were 75.8% for morphology, 78.7% for aggression and defence, and 93.0% for functions ($p = 0.011$). Regarding the presence of images, in all attempts, there was a significant difference favouring MCQs without images (65.1–69.8% vs 89.8–92.0% for questions with and without images, respectively).

Claude showed differences by MCQ category in one of the three attempts, with a higher rate for functions (93.0%) and a lower rate for aggression and defence (76.6%), $p = 0.028$. The presence of images resulted in different rates of correct answers for all attempts (74.4 vs 92.0%, $p = 0.002$ for questions with and without images, respectively, on the best attempt).

ChatGPT-4 and ChatGPT-3.5 showed no difference by MCQ category between any attempt. However, as with the other bots, a significant difference was observed in all attempts in questions with images: on ChatGPT-4's best attempt, accuracy was 81.4% and 94.9% for questions with and without images, respectively ($p = 0.005$).

However, the frequency of images was lower for questions on functions (15.0%) and morphology (27.3%) than for those on aggression and defence (40.4%). Therefore, logistic models were built considering the covariation between categories and the presence of images. For all bots, MCQs from aggression and defence and with images significantly reduced the odds of correct answers from 12.9% (Claude) to 17.6% (ChatGPT-4) when comparing questions on functions. Gemini also performed poorly on questions about morphology and with an image. Table 4 lists the models that considered each chatbot's best attempts.

The hierarchical clustering provided a dendrogram (Figure 1) in which the two clusters could be distinguished. Cluster 1 had 158 questions, whereas Cluster 2 had 22. In the second cluster of questions, the bots had different patterns of answers, except for Co-pilot. These questions were more frequently answered incorrectly. They had more images than those from Cluster 1 and a predominance of

Table 1. Classification of disciplines according to their number of multiple-choice questions, categories and relative distribution.

Discipline	Number (%)	Category	Number (%)
Anatomy	13 (7.2)	Morphology	33 (18.3)
Histology	11 (6.1)		
Embryology	9 (5.0)		
Physiology	35 (19.4)	Functions	100 (55.6)
Pharmacology	25 (13.9)		
Genetics	23 (12.8)		
Biochemistry	13 (7.2)		
Cellular biology	4 (2.2)		
Microbiology	17 (9.4)	Aggression and Defence	47 (26.1)
Immunology	15 (8.3)		
Parasitology	12 6.7)		
Pathology	3 (1.7)		

Table 2. Comparison of chatbot performance based on lowest and highest scores.

Chatbot	Lowest score (%)	Highest score (%)
Co-pilot	157 (87.2)	159 (88.3)
Gemini	153 (85.0)	155 (86.1)
Claude	156 (86.7)	158 (87.8)
GPT-4	164 (91.1)	165 (91.7)
GPT-3.5	162 (90.0)	163 (90.6)

Table 3. Pairwise comparison of chatbot performance with agreement metrics (k) and statistical significance (*p* values).

	Co-pilot	Gemini	Claude	GPT-4	GPT-3.5
Co-pilot		77.8% K = 0.04 *p* = 0.955	77.2% K = −0.083 *p* = 0.267	80.0% K = −0.108 *p* = 0.142	78.9% K = −0.117 *p* = 0.115
Gemini			87.2% K = 0.438 *p* < 0.001	86.1% K = 0.386 *p* < 0.001	88.9% K = 0.463 *p* < 0.001
Claude				91.7% K = 0.550 *p* < 0.001	90.6% K = 0.512 *p* < 0.001
GPT-4					96.7% K = 0.794 *p* < 0.001
GPT-3.5					

Table 4. Odds ratios (OR), confidence intervals (95% CI), and *p* values for chatbot performance across different thematic areas and imagen interactions.

Bot	Area*Image	OR	95% CI	*p*-value
Co-pilot	Functions*Image	Ref		
	Morphology*Image	0.300	0.056–1.607	0.160
	Aggression*Image	0.147	0.049–0.443	0.001
Gemini	Functions*Image	Ref		
	Morphology*Image	0.127	0.030–0.527	0.005
	Aggression*Image	0.174	0.059–0.513	0.002
Claude	Functions*Image	Ref		
	Morphology*Image	0.748	0.087–6.457	0.792
	Aggression*Image	0.129	0.044–0.376	< 0.001
GPT-4	Functions*Image	Ref		
	Morphology*Image	0.503	0.057–4.478	0.538
	Aggression*Image	0.176	0.052–0.599	0.005
GPT-3.5	Functions*Image	Ref		
	Morphology*Image	0.563	0.064–4.692	0.605
	Aggression*Image	0.153	0.048–0.487	0.001

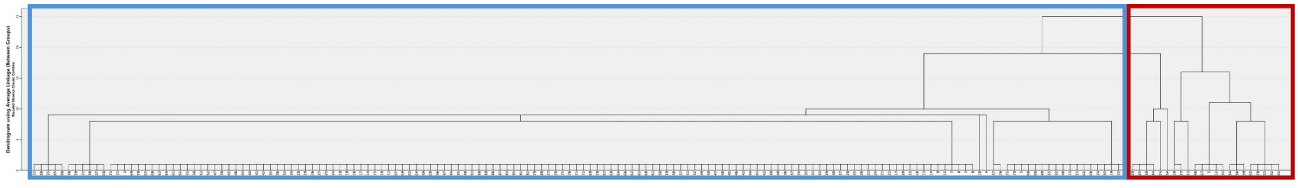

Figure 1. Hierarchical clustering dendrogram of data points based on similarity metrics. The dendrogram represents the hierarchical clustering of data points, illustrating the relationships and grouping patterns based on similarity metrics. Two clusters could be identified: a larger one (Cluster 1, blue rectangle) containing 158 MCQs, and a smaller one (Cluster 2, red rectangle).

themes related to aggression and defence, whereas those from Cluster 1 were predominantly about function (Table 5). Accordingly, in the agreement analysis, Co-pilot had the lowest kappa values compared to the other bots. The present analysis reinforces that finding.

Notably, MCQs related to "aggression" disciplines – immunology, microbiology, parasitology, and pathology – as well as image-based questions, pose the greatest challenge, leading to higher error rates across all bots.

Discussion

AI chatbots have demonstrated strong performance in answering multiple-choice questions in biological science, with GPT-4 achieving the highest accuracy. However, their agreement varied, with GPT-4 and GPT-3.5 showing substantial concordance, whereas Co-pilot had the lowest. The accuracy and reliability of Co-pilot, Gemini, Claude, and ChatGPT (versions 4 and 3.5) have not been previously assessed in the context of basic science education using

Table 5. Comparison of parameters between clusters based on image usage, category distribution, and chatbot accuracy.

Parameter	Cluster 1 ($n = 158$)	Cluster 2 ($n = 22$)	P-value
Image – yes n (%)	31 (19.6)	12 (54.5)	< 0.001
Category			0.041
Aggression	37 (23.4)	10 (45.5)	
Function	93 (58.9)	7 (31.8)	
Morphology	28 (17.7)	5 (22.7)	
Correct answers			
Co-pilot	138 (87.3)	21 (95.5)	0.478
Gemini	150 (94.9)	5 (22.7)	< 0.001
Claude	153 (96.8)	5 (22.7)	< 0.001
GPT-4	157 (99.4)	8 (36.4)	< 0.001
GPT-3.5	157 (99.4)	6 (27.3)	< 0.001

clinically based questions. This investigation elucidates the performance of AI chatbots in biological science MCQs using a databank of 11 years of Progress Test Examinations conducted across the most esteemed medical institutions in Brazil.

Clinic-based MCQs in the Progress Test (PT) examinations are carefully designed for rigorous, annual, cross-institutional assessment [18,19]. Developed by a consortium of universities, this examination evaluates medical students from their first to sixth years across medical schools in Brazil [19]. PT serves as a validated tool for formulating both interdisciplinary and discipline-specific MCQs, ensuring a comprehensive assessment of students' knowledge across biological and clinical sciences [12,20]. By integrating core concepts from basic medical sciences, the exam effectively measures students' development of progressive competencies and readiness for clinical practice [21].

AI tools have demonstrated success rates of up to 60% for medical examination questions in various fields. Among various chatbots, GPT has outperformed Co-pilot and Gemini in clinical chemistry [4,22] and anatomy assessments [3]. However, they fail to generate accurate and consistent responses to non-expert MCQs in physiology [23]. Furthermore, AI's knowledge in specialised fields, such as aggression-related disciplines, remains limited [7], likely because of its inability to interpret figures, graphs, and tables, restricted access to country-specific or non-indexed epidemiological data, and insufficient training in solving domain-specific MCQs.

Image-based questions significantly reduced accuracy across all models, with Co-pilot performing worse than other models, such as Gemini, Claude, GPT-3.5, and GPT-4. Co-pilot had under a 3% accuracy gap between text and image questions on musculoskeletal and bone structures, while GPT-3.5 and Gemini scored 6.5% and 12.4% lower for images, respectively [24]. Accordingly, Sau (2025) found that ChatGPT-4 and Gemini showed lower performance on image-based questions on neurosurgery board questions [6]. Moreover, Newton et al. (2025) found reduced performance of ChatGPT-4 on questions containing images when the answer options were added to an image as text labels [25]. Despite the limitations exhibited by GPT-4 in responding to gross anatomy questions, it demonstrated statistically superior knowledge bases and sophisticated language understanding over Copilot, GPT-3.5, and Gemini [3]. Image-based questions are crucial in subjects like anatomy, physiology, and histology. They present a particular challenge, as they require spatial recognition and detailed visual interpretation [26–29]. This emphasises the need for advancements in multimodal AI capabilities to enhance performance on image-based assessments.

Similar findings were observed with human test-takers. Questions addressing images that require more cognitive processes tend to be more difficult, following a hierarchical structure [25,30]. However, the use of images is insufficient to pose an increased difficulty [14,31]. In anatomy education, schematic images may facilitate student performance, whereas cross-sectional images may require additional cognitive abilities [15]. Future studies should address how the AI bot performance varies according to the type of images used. Logistic models and hierarchical clustering confirmed that image-based questions from disciplines, such as immunology, microbiology, parasitology, and pathology, classified as aggression and defence, decreased the odds of correct responses by up to 17.6% (in GPT-4). Chatbots struggled with reasoning-based MCQs that required deep physiological understanding and underperformed Korean medical students in a parasitology exam [7,23]. However, these studies presented the limitation of using chatbot versions that did not allow image inclusion in the analysis. Therefore, whether the lack of image-based inputs influenced the lower performance in disciplines such as parasitology was unclear.

AI-based image analysis relies on the quality and diversity of data. Factors such as image brightness, contrast, and noise significantly impact the accuracy of AIs. Variability in staining techniques (including H&E, immunohistochemistry, and fluorescence staining), image resolution, magnification, and imaging equipment further affect the robustness of chatbots [21]. Additionally, ethical concerns demand a deeper analysis of the regulatory challenges posed by real patient images used in chatbot models. Finally, despite the need for advancements in AI capabilities and the extensive data processed by generative AI, its effective use and application in medical education still rely on careful guidance from educators to ensure precise and meaningful learning [32].

Another practical implication of our findings is that, for high-stakes examinations, the use of image-based questions may be a security tool against cheating on tests with the aid of AI. This is especially relevant in the context of the increasing use of computer-based tests for residency applications or licencing examinations [7,33].

Limitations

This study has several limitations. Variables such as Internet speed, online traffic, specific versions of the chatbots used, and potential processing delays may have affected the AI's responses. Co-pilot and Claude have limited image analysis capabilities. Non-English questions may have resulted in less accurate responses owing to possible misinterpretations compared to English-language content. Additionally, because PT is based on clinical cases from the Brazilian context, chatbot performance may differ on international certification exams. Finally, our findings cannot be generalised to other medical disciplines because chatbots' knowledge bases are continuously updated through user interactions and feedback.

Conclusion

The performance of chatbots on tests in biological disciplines remained consistent; however, their efficacy varied between biological fields and image-based questions in undergraduate medical education. Although GPT-4 demonstrated potential as a tool for medical education, its limitations in discipline-specific image interpretation emphasise the need for continuous advancement.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research received no external funding.

Author contribution

Renato Ferretti contributed to the conception and design of the study. Material preparation, data collection, and analysis were performed by Renato Ferretti and Pedro Tadao Hamamoto Filho. The first draft of the manuscript wa written by Renato Ferretti, Joyce Santana Rizzi, Lorraine Silva Requena, Angelica Maria Bicudo, and Pedro Tadao Hamamoto Filho. All authors read, commented, and approved the final manuscript.

Data Availability Statement

On reasonable request, the research project's data are made available.

ORCID

Joyce Santana Rizzi http://orcid.org/0000-0002-5212-6006
Lorraine Silva Requena http://orcid.org/0000-0002-4211-5514
Angelica Maria Bicudo http://orcid.org/0000-0003-3043-5147
Pedro Tadao Hamamoto Filho http://orcid.org/0000-0001-6436-9307
Renato Ferretti http://orcid.org/0000-0003-3944-1906

References

[1] Haupt CE, Marks M. AI-generated medical advice—gpt and beyond. Jama. 2023;329(16):1349–1350 Preprint at doi: 10.1001/jama.2023.5321
[2] Morreel S, Mathysen D, Verhoeven V. Aye, AI! ChatGPT passes multiple-choice family medicine exam. Med Teach. 2023;45(6):665–666 doi: 10.1080/0142159X.2023.2187684
[3] Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in gross anatomy course: comparative analysis. Clin Anat. 2024;38(2):200–210. doi: 10.1002/ca.24244
[4] Sallam M, Al-Salahat K, Eid H, et al. Human versus artificial intelligence: ChatGPT-4 outperforming Bing, Bard, ChatGPT-3.5 and humans in clinical chemistry multiple-choice questions. Adv Med Educ Pract. 2024;15:857–871. doi: 10.2147/AMEP.S479801
[5] OpenAI. GPT-4 technical report. 2023.
[6] Sau S, George DD, Singh R. Accuracy and quality of ChatGPT-4o and Google Gemini performance on image-based neurosurgery board questions. Neurosurg Rev. 2025;48(1):320. doi: 10.1007/s10143-025-03472-7

[7] Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof. 2023;20:1. doi: 10.3352/jeehp.2023.20.1
[8] Kung TH, Cheatham M, Medenilla A. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198
[9] Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. JMIR Med Educ. 2024;10:e50965. doi: 10.2196/50965
[10] Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. J Med Educ Curric Dev. 2024;11:23821205241238641.
[11] Bharatha A, Ojeh N, Fazle Rabbi AM. Comparing the performance of ChatGPT-4 and medical students on MCQs at varied levels of Bloom's taxonomy. Adv Med Educ Pract. 2024;15:393–400. doi: 10.2147/AMEP.S457408
[12] Johnson TR, Khalil MK, Peppler RD, et al. Use of the NBME comprehensive basic science examination as a progress test in the preclerkship curriculum of a new medical school. How We Teach: Generalizable Educ Res Adv Physiol Educ. 2014;38(4):315–320. doi: 10.1152/advan.00047.2014
[13] Tavakol M, O'Brien D. Psychometrics for physicians: everything a clinician needs to know about assessments in medical education. Int J Med Educ. 2022;13:100–106 doi: 10.5116/ijme.625f.bfb1
[14] Holland J, O'Sullivan R, Arnett R. Is a picture worth a thousand words: an analysis of the difficulty and discrimination parameters of illustrated vs. text-alone vignettes in histology multiple choice questions assessment and evaluation of admissions, knowledge, skills and attitudes. BMC Med Educ. 2015;15(1). doi: 10.1186/s12909-015-0452-9
[15] Sagoo MG, Vorstenbosch MATM, Bazira PJ. Online assessment of applied anatomy knowledge: the effect of images on medical students' performance. Anat Sci Educ. 2021;14(3):342–351. doi: 10.1002/ase.1965
[16] Ali M, Benfante V, Basirinia G. Applications of artificial intelligence, deep learning, and machine learning to support the analysis of microscopic images of cells and tissues. J Imag. 2025;11(2): 59. doi: 10.3390/jimaging11020059
[17] Rodrigues Alessi M, Gomes HA, Lopes de Castro M, et al. Performance of ChatGPT in solving questions from the Progress Test (Brazilian National Medical Exam): a potential artificial intelligence tool in medical practice. Cureus. 2024. doi: 10.7759/cureus.64924
[18] Schuwirth LWT, van der Vleuten CPM. The use of progress testing. Perspect Med Educ. 2012;1(1):24–30 doi: 10.1007/s40037-012-0007-2
[19] Bicudo AM, Hamamoto Filho PT, Abbade JF, et al. Teste de Progresso em Consórcios para Todas as Escolas Médicas do Brasil. Rev Bras Educ Med. 2019;43(4):151–156. doi: 10.1590/1981-52712015v43n4rb20190018
[20] Al Alwan I. The progress test as a diagnostic tool for a new PBL curriculum. 2011. Available from: http://www.educationforhealth.net/
[21] Ali K, Cockerill J, Bennett JH, et al. Transfer of basic science knowledge in a problem-based learning curriculum. Eur J Dent Educ. 2020;24(3):542–547. doi: 10.1111/eje.12535
[22] Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Switz). 2023;11(6): 887. doi: 10.3390/healthcare11060887
[23] 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. doi: 10.7759/cureus.40977
[24] Guerra GA, Hofmann HL, Le JL. ChatGPT, Bard, and Bing chat are large language processing models that answered OITE questions with a similar accuracy to first-year orthopaedic surgery residents. Arthrosc: The J Arthroscopic Relat Surg. 2024;41(3):557–562. doi: 10.1016/j.arthro.2024.08.023
[25] Newton PM, Summers CJ, Zaheer U. Can ChatGPT-4o really pass medical science exams? A pragmatic analysis using novel questions. Med Sci Educ. 2025;35(2):721–729. doi: 10.1007/s40670-025-02293-z
[26] Mogali SR. Initial impressions of ChatGPT for anatomy education. Anat Sci Educ. 2024;17(2):444–447 doi: 10.1002/ase.2261
[27] Arun G, Perumal V, Urias FPJB. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: a comparative pilot study. Anatomical Sci Ed. 2024;17(7):1396–1405. doi: 10.1002/ase.2502
[28] Collins BR, Black EW, Rarey KE. Introducing AnatomyGPT: a customized artificial intelligence application for anatomical sciences education. Clin Anat. 2024;37(6):661–669. doi: 10.1002/ca.24178
[29] Totlis T, Natsis K, Filos D. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat. 2023;45(10):1321–1329. doi: 10.1007/s00276-023-03229-1
[30] Phillips AW, Smith SG, Straus CM. Driving deeper learning by assessment. An adaptation of the revised Bloom's taxonomy for medical imaging in gross anatomy. Acad Radiol. 2013;20(6):784–789. doi: 10.1016/j.acra.2013.02.001
[31] Schewior L, Lindner MA. Revisiting picture functions in multimedia testing: a systematic narrative review and taxonomy extension. Educ Psychol Rev. 2024;36(2). doi: 10.1007/s10648-024-09883-0
[32] Montiel-Romero S, Rajme-López S, Román-Montes CM. Recommended antibiotic treatment agreement between infectious diseases specialists and ChatGPT®. BMC Infect Dis. 2025;25(1). doi: 10.1186/s12879-024-10426-9
[33] Borges MC, Santos LL, Manso PH. Increased accessibility of computer-based testing for residency application to a hospital in Brazil with item characteristics comparable to paper-based testing: a psychometric study. J Educ Eval Health Prof. 2024;21:32. doi: 10.3352/jeehp.2024.21.32