



Research article

Comparative performance of artificial intelligence models on the Peruvian CODIRO 2023 dental residency examination

Edgar S. Povis-Flores^{1,*}, Mariana A. Ramírez-Benavides¹, Fabricio Meza-Ponce², Andrea G. Ramirez-Acuña¹, Mónica L. Ramírez Benavides¹ and Miguel A. Saravia-Rojas¹

¹ Faculty of Dentistry, Cayetano Heredia Peruvian University, Lima, Peru

² Faculty of Medicine, Cayetano Heredia Peruvian University, Lima, Peru

* **Correspondence:** Email: spovisflores@gmail.com; Tel: +51986366965.

Abstract: Artificial intelligence based on large language models has gained increasing relevance in educational and healthcare settings due to its ability to process information and generate coherent natural language responses. In the health sciences, several studies have evaluated the performance of these systems in academic and professional licensing examinations; however, evidence regarding their performance in dentistry, particularly in Latin American contexts, remains limited. This study aimed to compare the response accuracy of four artificial intelligence language models in answering the 2023 CODIRO Dental Residency admission examination in Peru, using the official answer key as the reference standard. An observational, cross-sectional, comparative study was conducted using the complete set of 100 multiple-choice questions from the examination. Claude, Copilot, Duck IA, and Qwen were each administered the same standardized prompt for every question, and responses were classified as correct or incorrect against the official key. Because all four models answered the same questions, accuracy was compared using Cochran's Q test, with pairwise exact McNemar tests and Holm correction applied when the omnibus test was significant. Copilot achieved the highest accuracy (100%), followed by Claude and Qwen (96% each) and Duck IA (86%); Cochran's Q test indicated a significant overall difference ($p < 0.001$). Duck IA performed significantly worse than the other three models, without significant differences among them. Agreement among models was highest in clinical disciplines and lowest in foundational disciplines such as anatomy and histology/embryology. Among items answered incorrectly, failing models frequently selected the same incorrect option, suggesting shared misconceptions rather than independent errors. These findings suggest that large language models, including platforms with limited representation in the dental education literature, can achieve high accuracy on dental residency examination questions; however,

this reflects knowledge recognition rather than clinical competence, and supports their potential only as complementary tools in dental education.

Keywords: artificial intelligence; large language models; education; dental; graduate; educational measurement; licensure; dental; Peru

1. Introduction

Artificial intelligence (AI) has experienced rapid development in recent decades, leading to significant advances in various fields of knowledge, particularly in the educational and healthcare sectors [1–3]. In this field, the development of systems based on natural language processing has enabled the creation of models capable of understanding and generating text with a high level of coherence. These models are known as large language models (LLMs) [4–6]. These technologies, which are gradually being integrated into academic and clinical settings, are used as support tools for educational content generation, information retrieval, and the enhancement of assisted learning processes [6–9].

Chatbots based on LLMs have been proposed as virtual assistants within educational environments, capable of providing continuous support to students and healthcare professionals, thereby facilitating access to information and promoting independent learning [6,10]. Their ability to process large amounts of information and generate context-based responses has generated growing interest in their application to medical and dental education. In this context, several studies have evaluated the performance of artificial intelligence models in professional licensing examinations in health-related fields such as medicine, pharmacy, dentistry, and nursing [1,7,11].

Some models, such as GPT-4, have achieved accuracy levels comparable to or even exceeding the minimum passing requirements for certain national professional licensing examinations [1], according to the findings of these studies. Likewise, other AI systems, such as Claude and GPT-4o, have demonstrated outstanding performance in specific academic assessments, achieving high accuracy rates across different areas of medical knowledge [11]. These findings suggest that language models may become valuable tools to support education in the health sciences and to develop AI-based educational resources.

Despite these advances, most of the available scientific evidence has been obtained from academic settings in Asia, Europe, and North America, primarily focusing on human medicine [9,11,12]. In contrast, studies evaluating the performance of artificial intelligence models in dental assessments remain limited, especially in Latin American countries. Differences in epidemiological profiles, educational curricula, and sociocultural contexts may influence the performance of these systems. This highlights the importance of evaluating their performance in professional examinations conducted within specific regional contexts.

The admission examination for the Dental Residency Program, conducted in Peru and coordinated by the Consejo Directivo del Residentado Odontológico (CODIRO), is a highly demanding academic examination aimed at selecting dental surgeons for specialization programs. This examination is held annually and includes questions from several essential areas of dentistry; therefore, it represents an appropriate tool for evaluating the performance of artificial intelligence systems in solving complex academic assessments.

Within this framework, the present study aimed to compare the response accuracy of four artificial intelligence language models in answering the 2023 CODIRO examination, using the official answer key as the reference standard.

2. Materials and methods

2.1. Study design and question selection

This observational, cross-sectional, comparative study included the complete set of 100 multiple-choice questions from the 2023 CODIRO examination, with no subset selection by the authors. The distribution of questions across the 12 thematic areas (ranging from 3 questions in Histology/Embryology to 10 questions in 8 other areas) reflected the structure of the official examination itself, rather than a sampling decision. For each question, model responses were classified as correct or incorrect against the official CODIRO 2023 answer key, which served as the sole reference standard throughout the study; no AI model's output was used as a reference for evaluating another model's performance (Table S1).

2.2. AI models evaluated

Four freely accessible large language model platforms were evaluated, using their default configuration and free consumer tier (no paid subscription was used for any platform): Claude (Anthropic, Sonnet 4.6), Copilot (Microsoft, Smart Mode), Duck IA (DuckDuckGo AI Chat, GPT-5.4 nano backend), and Qwen (Alibaba, Qwen3.5-Plus).

ChatGPT was intentionally excluded from this comparison. Although it remains among the most widely used and extensively benchmarked AI platforms, its performance on national licensing-style examinations in dentistry, medicine, pharmacy, and nursing has already been documented across multiple prior studies, including a systematic review and meta-analysis [1] and a study on a comparable Peruvian national dental examination [13]. Given this existing evidence base, the present study instead prioritized platforms with comparatively limited representation in the dental education literature (Copilot, Duck IA, Qwen), broadening the range of evaluated tools rather than replicating findings already established for ChatGPT.

2.3. Testing conditions and prompt standardization

Testing was conducted between March 15 and March 30, 2026, with each model administered by a single, dedicated researcher throughout the study period. For each model, all 100 questions were submitted within a single, continuous conversation; conversation history was not cleared between questions, in order to approximate realistic everyday usage rather than an artificially isolated pattern. Each question was answered once (single attempt), with no follow-up, correction, or regeneration of responses requested; question wording, submission sequence, and response-recording procedures were kept identical across models. The recorded responses and original conversations were reviewed for citation of external sources; none were found for any of the four models.

Questions were entered using a standardized prompt, “Responda la siguiente pregunta de opción múltiple correspondiente al examen CODIRO 2023. Seleccione únicamente la alternativa correcta y

responda solo con la letra (A, B, C, D o E)”, followed by the question stem and its five answer choices.

2.4. Statistical analysis

Descriptive statistics (number and percentage of correct responses) were calculated overall and by thematic area, with 95% confidence intervals estimated using the Wilson score method. Because all four models answered the same 100 questions, responses were treated as paired (matched) data rather than independent samples. The overall difference in accuracy among the four models was assessed using Cochran’s Q test, and when significant differences were detected, pairwise comparisons between models were performed using the exact McNemar test, with the resulting p-values adjusted for the six comparisons using the Holm method; the odds ratio and its 95% confidence interval were reported for each pairwise comparison. All analyses were conducted in R (version 4.5.2), using the DescTools and exact2x2 packages. A two-sided p-value < 0.05 was considered statistically significant.

2.5. Ethics approval of research

The study uses publicly available data and does not involve human participants, biological samples, or personal data; therefore, approval from an ethics committee was not required.

3. Results

Overall accuracy was 100% (100/100, 95% CI: 96–100%) for Copilot, 96% (96/100, 95% CI: 90–98%) for both Claude and Qwen, and 86% (86/100, 95% CI: 78–92%) for Duck IA (Table 1). Cochran’s Q test indicated a significant overall difference in accuracy among the four models ($Q = 27.91$, $df = 3$, $p < 0.001$).

At the level of individual thematic areas, no significant differences among models were detected (all $p > 0.05$; Table 1), with Orthodontics coming closest to significance ($p = 0.061$). Agreement among models was highest in areas emphasizing clinical and procedural knowledge: in Oral Surgery, Periodontics, Radiology, and Operative Dentistry/Cariology, all four models answered every question correctly, precluding a formal statistical comparison in these four areas. In contrast, the greatest discrepancies were observed in foundational and basic science areas, particularly Histology/Embryology, where accuracy ranged from 33% (Claude, Duck IA) to 100% (Copilot, Qwen), and Anatomy, where Duck IA (71%) underperformed relative to the other three models.

Pairwise comparisons (Table 2) revealed that Duck IA performed significantly worse than Claude, Copilot, and Qwen (all Holm-adjusted $p \leq 0.010$). No significant differences were found among Claude, Copilot, and Qwen in any pairwise comparison (all Holm-adjusted $p \geq 0.375$); Claude and Qwen split evenly across their four discordant items (two favoring each model), yielding an odds ratio of 1.00.

Duck IA, consistent with its significantly lower accuracy, was the only model to answer incorrectly in 8 of the 14 discordant items. On the remaining 6 items, where a second model (Claude or Qwen) also failed, the same incorrect option was selected by both in 5 cases, suggesting a shared pattern of misconception rather than independent, random errors. The exception was a question on Kennedy classification (Oral Rehabilitation), where the failing models selected different incorrect options (Supplementary Table S3).

Table 1. Accuracy of the four artificial intelligence language models by thematic area.

Thematic area	Total questions	Claude	Copilot	Duck IA	Qwen	P value*
Overall (global)	100	96, 96% (95% CI: 90–98%)	100, 100% (95% CI: 96–100%)	86, 86% (95% CI: 78–92%)	96, 96% (95% CI: 90–98%)	< 0.001
Oral Rehabilitation	5	4, 80% (95% CI: 38–96%)	5, 100% (95% CI: 57–100%)	3, 60% (95% CI: 23–88%)	4, 80% (95% CI: 38–96%)	0.262
Anatomy	7	6, 86% (95% CI: 49–97%)	7, 100% (95% CI: 65–100%)	5, 71% (95% CI: 36–92%)	6, 86% (95% CI: 49–97%)	0.262
Histology/Embryology	3	1, 33% (95% CI: 6–79%)	3, 100% (95% CI: 44–100%)	1, 33% (95% CI: 6–79%)	3, 100% (95% CI: 44–100%)	0.112
Oral Surgery	10	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	NA*
Oral Pathology	10	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	9, 90% (95% CI: 60–98%)	10, 100% (95% CI: 72–100%)	0.392
Pediatric Dentistry	10	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	9, 90% (95% CI: 60–98%)	10, 100% (95% CI: 72–100%)	0.392
Orthodontics	10	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	7, 70% (95% CI: 40–89%)	8, 80% (95% CI: 49–94%)	0.061
Periodontics	10	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	NA*
Radiology	10	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	NA*
Public Health/Research	10	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	9, 90% (95% CI: 60–98%)	10, 100% (95% CI: 72–100%)	0.392
Endodontics	5	5, 100% (95% CI: 57–100%)	5, 100% (95% CI: 57–100%)	3, 60% (95% CI: 23–88%)	5, 100% (95% CI: 57–100%)	0.112
Operative Dentistry/Cariology	10	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	10, 100% (95% CI: 72–100%)	NA*

Note: Data are presented as n, % (95% CI). 95% CI: 95% confidence interval, calculated using the Wilson score method. *P value from Cochran's Q test. NA: not applicable; the test statistic is undefined because all four models gave identical responses (no variability) in that area.

Table 2. Pairwise comparison of correct-response rates between language models.

Comparison	Discordant pairs	Favors first model	Favors second model	OR (95% CI)	P value (raw)	P value (Holm)
Copilot vs. Claude	4	4	0	0.00 (0.00–1.51)	0.125	0.375
Claude vs. Duck IA	10	10	0	0.00 (0.00–0.45)	0.002	0.010
Claude vs. Qwen	4	2	2	1.00 (0.07–13.80)	1.000	1.000
Copilot vs. Duck IA	14	14	0	0.00 (0.00–0.30)	< 0.001	< 0.001
Copilot vs. Qwen	4	4	0	0.00 (0.00–1.51)	0.125	0.375
Qwen vs. Duck IA	10	10	0	0.00 (0.00–0.45)	0.002	0.010

Note: OR: odds ratio; 95% CI: 95% confidence interval. Within each comparison, models are listed in order of discordant pairs won, so OR is oriented ≥ 1 . P values adjusted for the six comparisons (Holm method).

4. Discussion

This study evaluated the performance of four artificial intelligence models in answering the 2023 CODIRO Dental Residency Admission Examination. Copilot achieved the highest accuracy (100%), followed by Claude and Qwen (96% each) and Duck IA (86%). These findings suggest that contemporary large language models possess substantial potential for answering multiple-choice questions in dentistry, consistent with previous studies reporting high performance of artificial intelligence systems in licensing examinations in the health sciences [1,7,14].

Earlier evaluations found that models such as GPT-4 achieved accuracy rates ranging from 64.4% to 100% in national examinations related to pharmacy, medicine, nursing, and dentistry [1]. Artificial intelligence models have also demonstrated the ability to meet minimum passing standards in several medical disciplines through evaluations conducted using the Japanese National Medical Licensing Examination [12]. In this context, the results of the present study fall within the performance range previously described in the literature, despite reflecting substantially more advanced model generations.

The models demonstrated greater accuracy in clinical disciplines such as oral surgery, periodontics, radiology, and operative dentistry/cariology, and larger discrepancies in foundational disciplines such as anatomy and histology/embryology. Similar variability has been reported in previous investigations, suggesting that the performance of language models may differ across specialties due to the conceptual complexity of the content and the manner in which such information is represented within training datasets [11,15]. However, these subject-level comparisons should be interpreted with caution: with only three items in Histology/Embryology, for example, a single additional correct or incorrect response would shift the observed accuracy by approximately 33 percentage points, illustrating the limited precision of subject-level estimates in areas with few questions.

Statistical analysis confirmed an overall difference in accuracy among the four models. Duck IA performed significantly worse than Claude, Copilot, and Qwen, whereas no significant differences were found among these latter three models despite Copilot's descriptively highest accuracy. Notably, Claude and Qwen were statistically indistinguishable, splitting evenly across the items on which they disagreed. These findings indicate that, among current widely accessible models, meaningful performance gaps may be concentrated in a subset of platforms rather than reflecting a uniform hierarchy across all evaluated tools. Differences in model architecture, parameter scale, and training methodology may partly explain Duck IA's comparatively lower performance [5,16], a possibility discussed further in the Limitations section in relation to the specific model underlying this platform.

An additional finding was that, among the six discordant items in which more than one model answered incorrectly, the failing models converged on the same incorrect option in five cases. Because large language models do not provide verifiable access to their internal reasoning, the exact cause of any individual error cannot be confirmed; even so, for some items, a plausible content-based account can be proposed, while for others none is apparent. In the question on the number of bones of the head, for example, the three models that erred (Claude, Duck IA, and Qwen) all selected 22, the sum of the 8 cranial and 14 facial bones, without the six auditory ossicles that bring the total to 28 [17]. In the two questions on mandibular growth direction, Duck IA and Qwen both selected the direction opposite to the correct one, echoing a recognized point of confusion between the local direction of condylar growth and the net forward-downward displacement of the mandible [18]. By contrast, in the question on the onset of third-molar calcification, Claude and Duck IA converged on an incorrect age for which no plausible source could be identified. This overall pattern is nonetheless consistent with recent work showing that large language models tend to select the same incorrect distractors that most commonly mislead human learners, rather than failing in an independent or random manner [19].

From an educational perspective, the findings suggest that artificial intelligence models may serve as complementary tools for learning and preparation for academic assessments in dentistry. Several studies have reported that AI-based chatbots can support self-directed learning, provide immediate feedback, and facilitate innovative pedagogical approaches in higher education [6,20–22]. However, high accuracy on multiple-choice questions reflects only the knowledge base of clinical competence and does not necessarily extend to diagnostic reasoning, clinical skills, or real-world performance [23]. Given the risk of inaccurate outputs and the importance of preserving students' clinical reasoning and critical thinking skills, AI models should be used to complement, rather than substitute for, structured dental education and clinical supervision [4,24,25].

4.1. Limitations

This study has several limitations. Each question was submitted once per model, without repeated trials to assess response consistency; because large language models can produce different answers to the same prompt across separate attempts, the accuracy estimates reported here may not fully capture each model's range of possible responses. Performance was also evaluated using multiple-choice questions only, chosen to allow direct comparison with the examination's actual format; this scope excludes the open-ended clinical reasoning, diagnostic synthesis, and image interpretation required in real practice, and these findings should not be extrapolated to more complex clinical competencies. A separate concern involves whether any model could have accessed the examination outside of the recorded conversation. Although no evidence of external search was found in any model's recorded responses, and each model denied, when asked directly, having used the examination or its answer key as a source, only the latter claim relies on self-report and cannot be independently verified; the possibility that one or more models had encountered this publicly available examination during training also cannot be excluded by either check. Accuracy was also examined across the sequence of questions as a check for context-carryover effects from the continuous-conversation format, and no decline was observed for any model. Finally, only four artificial intelligence platforms were evaluated, and their performance may vary with prompt formulation, system updates, and the specific model version assessed; because these platforms are updated frequently, the reported accuracy rates correspond to the versions evaluated at the time of testing.

5. Conclusions

This study demonstrated substantial accuracy among four freely accessible artificial intelligence models in answering the 2023 CODIRO Dental Residency Admission Examination, although Duck IA performed significantly worse than the other three. Performance was more consistent in clinical disciplines than in foundational disciplines, and failing models frequently converged on the same incorrect option, suggesting shared misconceptions rather than independent errors. Notably, three of the four platforms evaluated, despite receiving comparatively little attention in the dental education literature relative to more widely studied models, achieved accuracy levels similar to or exceeding those reported for better-known systems, suggesting that these platforms merit further, more in-depth evaluation in future research. These findings suggest that artificial intelligence models may serve as valuable complementary tools for academic preparation in dentistry; however, high accuracy on multiple-choice questions is not equivalent to clinical competence, and their use should remain subject to critical supervision by educators and healthcare professionals.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Authors' contributions

Conceptualization: Mariana A. Ramírez-Benavides, DDS; Miguel A. Saravia-Rojas, PhD; Methodology: Edgar S. Povis-Flores, DDS; Fabricio Meza-Ponce, MD; Data curation: Andrea G. Ramirez-Acuña, DDS; Formal analysis: Mariana A. Ramírez-Benavides, DDS; Mónica L. Ramírez Benavides, DDS; Fabricio Meza-Ponce, MD; Writing - original draft: Edgar S. Povis-Flores, DDS; Mariana A. Ramírez-Benavides, DDS; Fabricio Meza-Ponce, MD; Writing - review & editing: All authors.

Conflict of interest

The authors declared that they have no conflicts of interest.

References

1. H. K. Jin, H. E. Lee, E. Kim, Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: A systematic review and meta-analysis, *BMC Med. Educ.*, **24** (2024), 1013. <https://doi.org/10.1186/s12909-024-05944-8>
2. A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, H. Müller, AI for life: Trends in artificial intelligence for biotechnology, *New Biotechnol.*, **74** (2023), 16–24. <https://doi.org/10.1016/j.nbt.2023.02.001>
3. A. Montejo-Ráez, S. M. Jiménez-Zafra, Current approaches and applications in natural language processing, *Appl. Sci.*, **12** (2022), 4859. <https://doi.org/10.3390/app12104859>
4. F. F. H. Nah, R. Zheng, J. Cai, K. Siau, L. Chen, Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration, *J. Inf. Technol. Case Appl. Res.*, **25** (2023), 277–304. <https://doi.org/10.1080/15228053.2023.2233814>

5. P. Lee, S. Bubeck, J. Petro, Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine, *N. Engl. J. Med.*, **388** (2023), 1233–1239. <https://doi.org/10.1056/NEJMSr2214184>
6. L. Labadze, M. Grigolia, L. Machaidze, Role of AI chatbots in education: Systematic literature review, *Int. J. Educ. Technol. High. Educ.*, **20** (2023), 56. <https://doi.org/10.1186/s41239-023-00426-1>
7. A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, et al., How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment, *JMIR Med. Educ.*, **9** (2023), e45312. <https://doi.org/10.2196/45312>
8. A. J. López-Galisteo, O. Borrás-Gené, The creation and evaluation of an AI assistant (GPT) for educational experience design, *Information*, **16** (2025), 117. <https://doi.org/10.3390/info16020117>
9. T. Camarata, L. McCoy, R. Rosenberg, K. R. T. Grellinger, K. Brettschnieder, J. Berman, LLM-generated multiple choice practice quizzes for preclinical medical students, *Adv. Physiol. Educ.*, **49** (2025), 758–763. <https://doi.org/10.1152/advan.00106.2024>
10. G. J. Hwang, H. Xie, B. W. Wah, D. Gašević, Vision, challenges, roles and research issues of artificial intelligence in education, *Comput. Educ. Artif. Intell.*, **1** (2020), 100001. <https://doi.org/10.1016/j.caeai.2020.100001>
11. S. Y. Lin, Y. Y. Hsu, S. W. Ju, P. C. Yeh, W. H. Hsu, C. H. Kao, Assessing AI efficacy in medical knowledge tests: A study using Taiwan's internal medicine exam questions from 2020 to 2023, *Digit. Health*, **10** (2024), 20552076241291404. <https://doi.org/10.1177/20552076241291404>
12. T. Kawahara, Y. Sumi, GPT-4/4V's performance on the Japanese National Medical Licensing Examination, *Med. Teach.*, **47** (2025), 450–457. <https://doi.org/10.1080/0142159X.2024.2342545>
13. M. A. Saravia-Rojas, C. Mendiola-Aquino, F. Orejuela-Ramirez, W. Tunkuipa-Chacón, R. Geng-Vivanco, ¿Pueden los grandes modelos de lenguaje de inteligencia artificial aprobar el Examen Nacional de Odontología en el Perú, *Rev. Estomatol. Herediana*, **35** (2025), 305–311. <https://doi.org/10.20453/reh.v35i4.6253>
14. C. C. Lin, J. S. Sun, C. H. Chang, Y. H. Chang, J. Z. C. Chang, Performance of artificial intelligence chatbots in National dental licensing examination, *J. Dent. Sci.*, **20** (2025), 2307–2314. <https://doi.org/10.1016/j.jds.2025.05.012>
15. T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, et al., Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models, *PLoS Digital Health*, **2** (2023), e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
16. R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, et al., On the opportunities and risks of foundation models, preprint, arXiv:2108.07258.
17. N. Norton, *Anatomía de Cabeza y Cuello Para Odontólogos*, 4th edition, Elsevier, Barcelona, 2007.
18. T. Rakosi, *Ortopedia Maxilar*, SALVAT, Barcelona, 1992.
19. N. Liu, S. Sonkar, R. G. Baraniuk, Do LLMs make mistakes like students? Exploring natural alignment between language models and human error patterns, in *Artificial Intelligence in Education (AIED 2025)*, **15880** (2025), 364–377. https://doi.org/10.1007/978-3-031-98459-4_26
20. E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al., ChatGPT for good? On opportunities and challenges of large language models for education, *Learn. Individ. Differ.*, **103** (2023), 102274. <https://doi.org/10.1016/j.lindif.2023.102274>

21. M. Sallam, ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns, *Healthcare*, **11** (2023), 887. <https://doi.org/10.3390/healthcare11060887>
22. H. A. Aluko, A. Aluko, G. A. Offiah, F. Ogunjimi, A. O. Aluko, F. M. Alalade, et al., Exploring the effectiveness of AI-generated learning materials in facilitating active learning strategies and knowledge retention in higher education, *Int. J. Organ. Anal.*, **34** (2026), 1243–1267. <https://doi.org/10.1108/IJOA-07-2024-4632>
23. G. E. Miller, The assessment of clinical skills/competence/performance, *Acad. Med.*, **65** (1990), S63–S67. <https://doi.org/10.1097/00001888-199009000-00045>
24. Z. Niu, X. Kuang, J. Chen, X. Cai, P. Zhang, The application and challenges of ChatGPT in laboratory medicine, *Adv. Lab. Med.*, **6** (2025), 397–410. <https://doi.org/10.1515/almed-2025-0080>
25. A. Rao, J. Kim, M. Kamineni, M. Pang, W. Lie, K. J. Dreyer, et al., Evaluating ChatGPT as an adjunct for radiologic decision-making, *J. Am. Coll. Radiol.*, **20** (2023), 990–997. <https://doi.org/10.1016/j.jacr.2023.05.003>



AIMS Press

©2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)