Received: 11 Aug. 2025 | Revised: 14 Aug. 2025 | Accepted: 16 Aug. 2025 | Published online: 25 Aug. 2025

RESEARCH ARTICLE

Contemporary Education and Teaching Research
2025, Vol. 6 (8)295-309
DOI: 10.61360/BoniCETR252018770801

Rapid Evolution of Large Language Models in Medical Education: Comparative Performance of ChatGPT-3.5, ChatGPT-5, and DeepSeek on Medical Microbiology MCQs

Malik Sallam[1,2,*], Amal Irshaid[2], Johan Snygg[3,4], Rula Albadri[5] & Mohammed Sallam[6,7,8,9]

[1]Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Jordan
[2]Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Jordan
[3]Department of Management, Mediclinic City Hospital, Mediclinic Middle East, Dubai, United Arab Emirates
[4]Department of Anesthesia and Intensive Care, University of Gothenburg, Sahlgrenska Academy, Sweden
[5]Department of Family Medicine, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates
[6]Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates
[7]Department of Management, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates
[8]Department of Management, School of Business, International American University, United States
[9]College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai, United Arab Emirates

Abstract: Rapid advances in large language models (LLMs) warrant specialty-specific benchmarking to assess their educational potential and limitations. We evaluated the newly released generative artificial intelligence (genAI) model ChatGPT-5, DeepSeek-R1, and the early ChatGPT-3.5 on 80 multiple-choice questions (MCQs) from a medical microbiology course examination, weighted for midterm and final components. Items were classified according to the revised Bloom's taxonomy. Performance was compared with that of more than 150 Doctor of Dental Surgery students. Content quality was assessed independently by two consultants in clinical microbiology using the validated CLEAR tool modified to assess AI content completeness, accuracy, and relevance. The mean total scores were 80.5 for ChatGPT-3.5, 96.0 for ChatGPT-5, and 95.5 for DeepSeek, versus a student mean of 86.21/100. ChatGPT-5 and DeepSeek-R1 significantly outperformed ChatGPT-3.5 in completeness and accuracy scores, with no differences between them. ChatGPT-5 maintained high accuracy across lower- and higher-order cognitive Bloom's domains, whereas DeepSeek-R1 showed a significant drop in higher-order items. For ChatGPT-3.5, incorrect responses had longer answer-choice word counts. CLEAR scores were significantly higher for correct versus incorrect responses in all models ($p < 0.001$). This study showed that the currently available LLMs can exceed average student performance in medical microbiology while providing high-quality explanations. Regular benchmarking is essential to ensure responsible integration of genAI into educational, pedagogical, and assessment tools.

Keywords: ChatGPT-5, artificial intelligence, large language models, medical education, medical microbiology, assessment

1. Introduction

The emergence of large language models (LLMs) represents one of the most rapid and consequential technological developments in recent educational history (Yusuf et al., 2024; Sharma et al., 2025). These generative artificial intelligence (genAI) tools, pioneered by OpenAI's ChatGPT models, have progressed from producing variable and often

Corresponding Author: Malik Sallam
[1]Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Jordan
[2]Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Jordan

unreliable outputs to delivering coherent, context- and domain-specific responses within a remarkably short time frame (Sallam, 2023; Azaria et al., 2023; Lin et al., 2024). In higher education, and particularly in medical and dental training, such AI-enabled advancements raise both opportunities for innovation and concerns about the implications for teaching, learning, and assessment (Sallam et al., 2023c; Hu et al., 2025; Kovalainen et al., 2025).

Medical microbiology constitutes a highly suitable domain for evaluating the capabilities of LLMs (Gurajala, 2024). This discipline requires both accurate recall of a broad factual base, including microbial classification, virulence mechanisms, and antimicrobial susceptibility patterns, and the application of knowledge to interpret diagnostic results, guide therapeutic decisions, and integrate microbiological data into clinical reasoning (Joshi, 2021; Mohseni & Ghorbani, 2024). Competence in medical microbiology thus involves multiple cognitive levels, from foundational recall to higher-order analytical and evaluative skills (Singh & Nagmoti, 2021). The multidimensional nature of the medical microbiology knowledge base provides a rigorous test for genAI tools, which must demonstrate more than superficial memorization to achieve clinically relevant performance (Sallam et al., 2023a; Sallam & Al-Salahat, 2023).

The current pace of LLM development is noteworthy. Unlike human learners, whose cognitive growth is gradual, LLMs evolve in distinct and substantial leaps with each model iteration, reflecting advances in architecture, training data, and fine-tuning methodologies (Parthasarathy et al., 2024; Matarazzo & Torlone, 2025). These rapid changes render performance evaluations quickly outdated, highlighting the need for systematic and recurrent benchmarking (Sallam et al., 2024d). In the higher education settings, without such monitoring, educators and policymakers risk basing decisions on outdated or anecdotal evidence, leading either to premature integration of genAI tools into high-stakes assessments or to underutilization of potentially transformative tools (Michel-Villarreal et al., 2023; Yusuf et al., 2024).

From the perspective of students, genAI tools present an attractive proposition (Abdaljaleel et al., 2024; Nelson et al., 2025; Kim et al., 2025). Generative AI tools can operate as on-demand learning companions, capable of generating explanations, simulating examination conditions, and providing individualized feedback without fatigue or variability in mood or availability (Zhu, 2025; Mirea et al., 2025). In principle, genAI models can supplement traditional instruction by offering repeated, tailored practice and by adapting content difficulty to the learner's current competency level (Katona & Gyonyoru, 2025; Vieriu & Petrea, 2025). However, the educational utility of LLMs in specialty-specific domains such as medical microbiology depends critically on their accuracy, reliability, and ability to engage meaningfully with higher-order cognitive tasks (Sallam et al., 2023a; Sallam & Al-Salahat, 2023).

Multiple-choice questions (MCQs) are a widely used method of competency assessment in medical and dental education (Parekh & Bahadoor, 2024; Haugen & de Lange, 2024). Their advantages include objectivity, scalability, and the capacity to assess a broad range of content within a single examination (Newton, 2020). When designed according to established psychometric principles, MCQs can probe beyond factual recall, testing understanding, application, and evaluation skills (Monrad et al., 2021). In the context of genAI evaluation, MCQs offer a standardized, reproducible platform for performance comparison across models and with human learners (Sallam & Al-Salahat, 2023; Sallam et al., 2024b; Sallam et al., 2024a; Bharatha et al., 2024; Herrmann-Werner et al., 2024). Such controlled conditions enable the measurement of correctness and the analysis of clarity, reasoning quality, and domain-specific weaknesses (Gilson et al., 2023; Newton & Xiromeriti, 2024). Thus, benchmarking genAI performance in specialty topics such as medical microbiology serves two essential functions. First, it provides empirical data to inform whether these tools can serve as legitimate educational adjuncts. Second, it establishes a trajectory of progress across model iterations, allowing stakeholders to anticipate future capabilities and prepare for potential shifts in the balance between human and machine contributions to the learning process.

This approach is particularly critical in light of the unprecedented rate of genAI performance improvement, which may compress decades of pedagogical evolution into a few years of technological change (Wong, 2024). At the same time, the accelerating capabilities of LLMs raise legitimate concerns about the role of educators (Barakat et al., 2025). If AI tools achieve accuracy levels equal to or exceeding those of students — and potentially rivaling those of domain experts — the traditional function of the educators and health professionals as the primary source of knowledge transmission and patient care may be fundamentally altered (Sallam, 2023; Rony et al., 2024; Hirani et al., 2024). Historical precedent demonstrates that new educational technologies, from the printing press to online learning platforms, have not eliminated the need for teachers but have reshaped their roles toward facilitation, mentorship, and the cultivation of higher-order thinking (Rajaram, 2023; Parveen & Ramzan, 2024).

However, the pace of genAI development suggests the possibility of more disruptive changes, in which certain instructional tasks are delegated entirely to machines (Xia et al., 2024; Storey et al., 2025). This evolving AI landscape has profound implications for curriculum design, assessment integrity, and the professional identity of educators (Ateeq et al., 2024; Tan et al., 2025; Bushuyev et al., 2025). If genAI tools can consistently outperform average students on well-constructed specialty examinations, educational institutions may need to reconsider the structure and purpose of their assessments to ensure they remain valid measures of human competence (Rudolph et al., 2023; Trikoili et al., 2025). Moreover, faculty may need to focus increasingly on guiding students in the critical appraisal of AI-generated information, fostering skills in verification, ethical application, and the integration of human judgment into decision-making processes (Fu & Weng, 2024; Sallam & Sallam, 2025).

Preliminary evaluations of ChatGPT-3.5 in medical microbiology demonstrated substantial capability but fell short of average human student performance (Sallam & Al-Salahat, 2023). Recent evidence suggests that newer genAI models, including ChatGPT-4o and emerging competitors such as DeepSeek, showed significant gains not only in accuracy but also in the ability to handle complex, higher-order cognitive tasks (Sallam et al., 2025b; Sallam et al., 2025a; Jiang et al., 2025; Jin et al., 2025). These gains represent more than incremental improvement; they would signal a qualitative shift in the potential role of genAI in health professions education with recognized benefits and risks (Sallam, 2023; Rodger et al., 2025). Against this background, the present study aimed to systematically benchmark the performance of three genAI tools — ChatGPT-3.5, ChatGPT-5, and DeepSeek — on a standardized set of medical microbiology MCQs drawn from an actual dental school examination. By comparing these results with those of the original student cohort, this study sought to quantify the progression of genAI capabilities within a specialty-specific, clinically relevant context. The goal is not only to measure current performance but also to illustrate the trajectory of improvement and to consider its implications for educational practice, assessment design, and the evolving responsibilities of educators in the era of genAI.

2. Methods

2.1 Study design

This study was conducted as a structured performance evaluation of genAI models, in accordance with the METRICS framework for the assessment of AI tools in healthcare (Sallam et al., 2024c). Model configurations, evaluation procedures, prompt formulations, and language settings were explicitly documented to ensure reproducibility. Responses generated by each genAI model were evaluated for completeness, factual accuracy, and contextual relevance using a modified version of the validated CLEAR assessment framework for AI-generated content (Sallam et al., 2023b). Two independent raters (M.S. and A.I.) — both clinical microbiologists with Jordan Medical Council (JMC) board certification and a combined 29 years of specialty experience — conducted the assessments, and inter-rater reliability was calculated to reduce subjectivity.

The dataset comprised 80 MCQs drawn from the Medical Microbiology examination for the Doctor of Dental Surgery (DDS) program at the University of Jordan, administered during the Spring Semester of the 2021–2022 academic year. The examination included 40 midterm questions (weight: 1 grade each) and 40 final questions (weight: 1.5 grades each). All MCQs were written in English, the official language of instruction for the DDS program. The exam was delivered online. Psychometric parameters, including difficulty and discrimination indices, were derived from the performance data of 153 students on the midterm and 154 students on the final examination. All MCQs were authored by the course instructor (first author, M.S.) and were free from copyright restrictions. Ethical approval was not required because all data were fully anonymized, derived from publicly available examination results, and based on original, copyright-free questions generated by the first author.

2.2 Classification of MCQs based on the revised Bloom's taxonomy and genAI prompting

In the early study by Sallam & Al-Salahat (2023), all MCQs were categorized according to the cognitive domains of the revised Bloom's taxonomy: (1) Remember, (2) Understand, (3) Analyze, and (4) Evaluate. "Remember" items required simple recall of factual information with minimal cognitive effort. "Understand" items assessed comprehension and the ability to link related concepts. "Analyze" items required breaking down information into components, identifying patterns, and making comparisons. "Evaluate" items involved forming judgments, assessing the quality of information, and making decisions, representing the highest level of cognitive demand.

The 80 MCQs were administered to three genAI models: ChatGPT-3.5 (tested on March 11, 2023), ChatGPT-5 (tested on August 9, 2025), and DeepSeek-R1 (tested on August 9, 2025). For ChatGPT-3.5, the following standardized prompt was used: "Select the most appropriate answer for the following MCQ with rationale for selecting this choice and excluding the other choices."

2.3 Generative AI content evaluation

Responses were first assessed for correctness against the answer key. Subjective evaluation was then performed using a modified CLEAR framework, which rated three dimensions: (1) completeness of the response, (2) accuracy, defined as the absence of false information and alignment with evidence-based knowledge, and (3) appropriateness and relevance, reflecting clarity, organization, and absence of extraneous content (Sallam et al., 2023b). Each dimension was scored on a 5-point Likert scale (1 = poor, 5 = excellent). To improve objectivity, pre-defined criteria specific to each MCQ were developed through consensus between the first and second authors (M.S. and A.I.). The outputs from each genAI model were evaluated independently by the two raters, both experienced clinical microbiologists (M.S. and A.I.). For each response, the CLEAR score was calculated as the mean of the three dimension scores, and overall average CLEAR scores were obtained by averaging the raters' scores.

2.4 Statistical analysis

All statistical analyses were performed using SPSS software, version 26.0 (Armonk, NY: IBM Corp). Two-sided p values of less than 0.05 were considered statistically significant. Descriptive statistics were used to summarize MCQ characteristics, model performance, and student scores. Categorical variables, including the proportion of correct answers within each revised Bloom's taxonomy category, were compared between genAI models using two-sided Fisher's exact tests (FETs). For continuous variables (e.g., stem word count, answer-choice word count, and CLEAR scores), normality was assessed using the Shapiro–Wilk test. Because distributions were non-normal, non-parametric tests were applied. Differences in word counts and CLEAR scores between correct and incorrect responses within each model were assessed using the Mann–Whitney U test (M-W). CLEAR scores (treated as scale variables) were compared across the three models using the Kruskal–Wallis H test (K-W), followed by pairwise post hoc comparisons with the M-W test when the overall result was significant. Comparisons were conducted for each CLEAR dimension (completeness, accuracy, relevance) and for the overall CLEAR score. Inter-rater reliability for correctness and CLEAR ratings was evaluated using Cohen's kappa (κ) statistic.

3. Results

3.1 MCQs features and overview of genAI performance

The 80 MCQs were classified according to the revised Bloom's taxonomy as follows: Remember (n = 26, 32.5%), Understand (n = 17, 21.3%), Analyze (n = 12, 15.0%), and Evaluate (n = 25, 31.3%). Most items (n = 76) addressed topics in medical virology; the remaining questions covered medical mycology (n = 2; one Remember, one Understand) and oral parasitology (n = 2; both Evaluate). Based on weighted scoring (1 point per midterm MCQ, 1.5 points per final MCQ; maximum total score, 100), ChatGPT-3.5 achieved a total score of 80.5, ChatGPT-5 scored 96.0, and DeepSeek scored 95.5. The mean student score for the same examination was 86.21/100.

3.2 Generative AI models' performance by revised Bloom's category

When stratified by cognitive domain, ChatGPT-3.5 correctly answered 37 of 43 (86.0%) items in the "Remember" or "Understand" categories and 27 of 37 (73.0%) items in the "Analyze" or "Evaluate" categories; the difference was not statistically significant (p = 0.170). ChatGPT-5 achieved 42 of 43 (97.7%) and 35 of 37 (94.6%) correct responses, respectively, with no significant difference between categories (p = 0.593). DeepSeek attained perfect accuracy for "Remember" or "Understand" items (43 of 43; 100%) but a lower proportion for "Analyze" or "Evaluate" items (33 of 37; 89.2%); this difference reached statistical significance (p = 0.042), making DeepSeek the only model to demonstrate a significant variation in performance between lower-order and higher-order cognitive domains (Table 1).

Table 1 Accuracy of Three Generative AI (genAI) Models in Answering Medical Microbiology Multiple-Choice Questions (MCQs)

GenAI model	Answer	Revised Bloom Taxonomy		*p* value
		Remember or Understand	Analyze or Evaluate	(two-sided FET)
		Count (%)	Count (%)	
ChatGPT-3.5	*Correct*	37 (86.0)	27 (73.0)	0.170
	Incorrect	6 (14.0)	10 (27.0)	
ChatGPT-5	*Correct*	42 (97.7)	35 (94.6)	0.593
	Incorrect	1 (2.3)	2 (5.4)	
DeepSeek	*Correct*	43 (100)	33 (89.2)	0.042
	Incorrect	0	4 (10.8)	

Notes:
Stratified by cognitive domain according to the revised Bloom's taxonomy. p values were calculated using the two-sided Fisher's exact test (FET) comparing lower-order ("Remember" or "Understand") and higher-order ("Analyze" or "Evaluate") items for each genAI model.

3.3 Generative AI models' performance based on MCQ complexity and CLEAR scores

Across all genAI models, stem word counts did not differ significantly between correct and incorrect responses (ChatGPT-3.5, p = 0.423; ChatGPT-5, p = 0.482; DeepSeek, p = 0.617). For ChatGPT-3.5, the mean answer-choice word count was higher for incorrectly answered items (32.5±16.9 words) compared with correct responses (22.1±17.7 words), a difference that reached statistical significance (p = 0.036). No significant differences in choice word counts were observed for ChatGPT-5 (p = 0.514) or DeepSeek (p = 0.678). For all three genAI models, CLEAR scores were significantly higher for correct responses compared with incorrect ones (ChatGPT-3.5, 4.9±0.3 vs. 2.5±0.8; ChatGPT-5, 5.0±0 vs. 2.8±0.2; DeepSeek, 5.0±0.3 vs. 3.3±1.1; all p < 0.001, Table 2).

Table 2 Mean Stem and Answer-Choice Word Counts and Mean CLEAR Scores for Correct and Incorrect Responses by Three Generative Ai Models

Variable		ChatGPT-3.5		ChatGPT-5		DeepSeek	
Correctness		Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Stem word count	Mean±SD	15.7±12.7	20.3±17.3	16.4±13.3	21.3±26.6	16.9±14	11.5±3.9
p value		0.423		0.482		0.617	
Choices word count	Mean±SD	22.1±17.7	32.5±16.9	23.9±18	30.3±18.7	24.1±18	25.8±18.6
p value		0.036		0.514		0.678	
Average CLEAR for ChatGPT-3.5	Mean±SD	4.9±0.3	2.5±0.8				
p value		<0.001					
Average CLEAR for ChatGPT-5	Mean±SD			5.0±0	2.8±0.2		
p value				<0.001			
Average CLEAR for DeepSeek	Mean±SD					5±0.3	3.3±1.1
p value						<0.001	

Notes:
p values are from Mann Whiteny U test comparing correct versus incorrect responses within each model. CLEAR scores range from 1 (poor) to 5 (excellent).

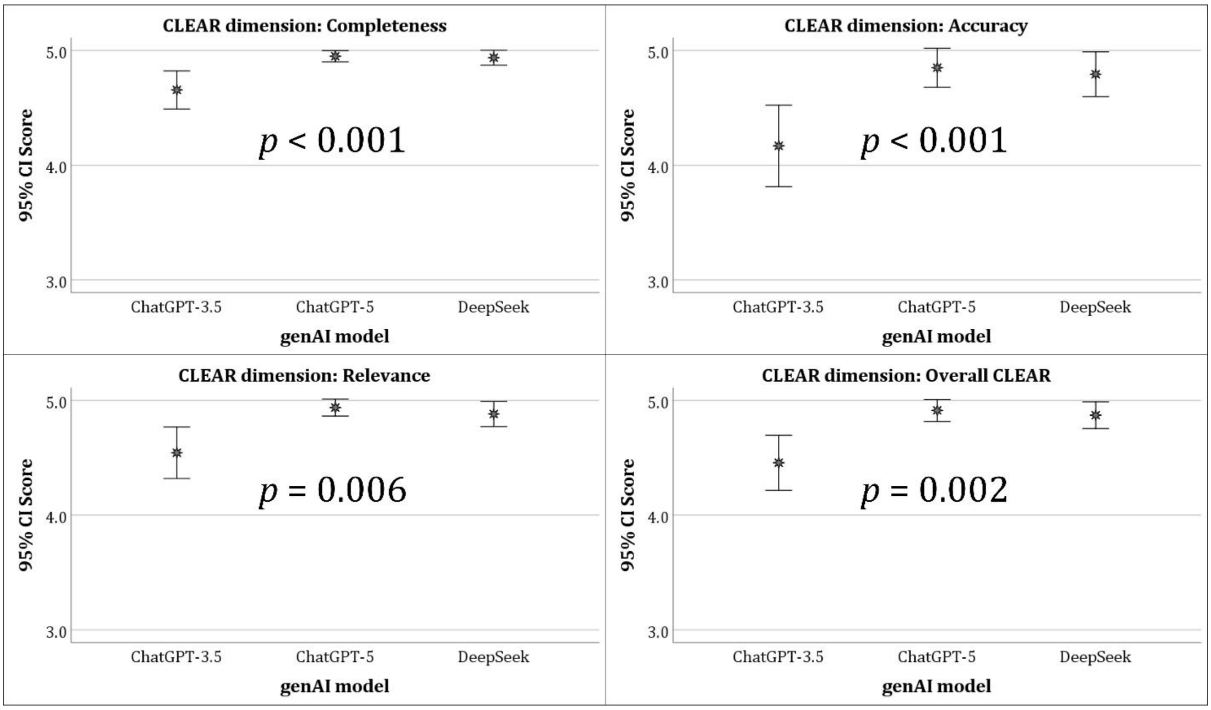
3.4 Head-to-head benchmarking of genAI based on CLEAR components

Inter-rater agreement for completeness ratings was almost perfect for ChatGPT-3.5 (κ = 0.827, p < 0.001) and perfect for ChatGPT-5 (κ = 1.000, p < 0.001) and DeepSeek (κ = 1.000, p < 0.001). For accuracy, ratings showed near-perfect agreement for ChatGPT-3.5 (κ = 0.965, p < 0.001) and ChatGPT-5 (κ = 1.000, p < 0.001), and strong agreement for DeepSeek (κ = 0.883, p < 0.001). Relevance ratings demonstrated near-perfect agreement for ChatGPT-3.5 (κ = 0.961, p < 0.001) and perfect agreement for ChatGPT-5 (κ = 1.000, p < 0.001), with moderate agreement for DeepSeek (κ = 0.527, p < 0.001).

Using the CLEAR framework, ChatGPT-3.5 achieved mean scores of 4.66±0.74 for completeness, 4.17±1.60 for accuracy, and 4.54±1.01 for relevance, yielding an overall score of 4.46±1.08. ChatGPT-5 recorded 4.95±0.22, 4.85±0.76, and 4.94±0.33 for the respective dimensions, with an overall score of 4.91±0.43. DeepSeek recorded 4.94±0.29, 4.79±0.88, and 4.88±0.49, with an overall score of 4.87±0.52.

Kruska-Wallis testing demonstrated significant differences across the three genAI models for completeness (H = 15.800, p < 0.001), accuracy (H = 16.999, p < 0.001), relevance (H = 10.198, p = 0.006), and overall CLEAR score (H = 12.373, p = 0.002). In M-W analyses, both ChatGPT-5 and DeepSeek scored significantly higher than ChatGPT-3.5 for completeness (p = 0.002 for both) and accuracy (p ≤ 0.003 for both). For relevance, ChatGPT-5 scored significantly higher than ChatGPT-3.5 (p = 0.002), whereas DeepSeek's difference from ChatGPT-3.5 did not reach significance (p = 0.075). Overall, CLEAR scores were significantly higher for ChatGPT-5 (p = 0.001) and DeepSeek (p = 0.031) compared with ChatGPT-3.5. No statistically significant differences were observed between ChatGPT-5 and DeepSeek in any dimension (all p > 0.05, Figure 1).

Figure 1 Error-Bar Plots Showing Mean CLEAR Scores with 95% Confidence Intervals (CIs) for Completeness, Accuracy, and Relevance across the Three Generative Artificial Intelligence (genAI) Models



Notes:
p values are from Kruskal–Wallis tests comparing scores among models for each dimension.

4. Discussion

This study demonstrated a notable acceleration in the capabilities of LLMs applied to medical microbiology assessment. Using a fixed bank of specialty MCQs and human student performance as a reference, ChatGPT-5 and DeepSeek achieved examination scores of 96.0 and 95.5 (of 100), respectively, exceeding the student mean of 86.21/100 and substantially surpassing the prior-generation ChatGPT-3.5 (80.5). Quality ratings using the modified CLEAR tool tracked these gains as follows. Newer genAI models produced more complete, accurate, and relevant explanations with near-ceiling scores and robust inter-model differences on K-W testing. Inter-rater reliability was excellent for most CLEAR domains, strengthening confidence in these associations. Together, the findings of this study indicated that general-purpose LLMs have crossed a performance threshold at which they can, in many circumstances, meet or exceed average learner performance on specialty content.

Several features of the results merit closer interpretation. First, the advantage of newer genAI models, such as ChatGPT-5 did not depend on lower-order cognitive tasks. ChatGPT-5 showed uniformly high accuracy across "Remember/Understand" and "Analyze/Evaluate" revised Bloom's categories, suggesting genuine gains in reasoning over and above factual retrieval and emphasizing that this advanced model is characterized by its sophisticated multimodal features (Oyekunle et al., 2024). DeepSeek, while achieving perfect accuracy on lower-order items, showed a statistically significant decrement on higher-order items. This pattern—high proficiency in knowledge recall with emerging but not complete facility in multistep appraisal—points to genAI model-specific differences in instruction tuning, error calibration, or handling of distractor structure. It also illustrates why benchmarking by cognitive level is essential (Holzinger et al., 2023; Ying et al., 2025). Aggregate genAI accuracy can mask meaningful weaknesses at the very skills educators prize for safe practice.

Second, item structure appeared to matter for earlier genAI models. For ChatGPT-3.5, incorrect answers were associated with longer answer-choice text, whereas stem length showed no association for any model (Sallam & Al-Salahat, 2023). This suggests that distractor complexity, not case narrative length per se, increased error susceptibility—possibly by amplifying lexical overlap, introducing superficially plausible alternatives, or taxing earlier genAI models' ability to weigh multiple qualifiers in parallel. The absence of this effect in ChatGPT-5 and DeepSeek indicates that newer genAI models better parse and prioritize information within dense option sets, an ability that aligns with their higher CLEAR accuracy and completeness as highlighted in this study.

Third, genAI explanation quality co-varied with correctness in all models, with large and consistent differences in CLEAR scores between correct and incorrect responses. Pedagogically, this matters as it implies that high-performing genAI models not only choose correct answers more often but also produce explanations that are clearer, more complete, and more accurate—attributes that affect how learners internalize reasoning patterns. Notably, for DeepSeek, interrater agreement on the relevance dimension—defined as whether content is clear, concise, unambiguous, well-organized, and free from irrelevant information—was only moderate. This may suggest that some responses, while fluent, varied in how well they met these criteria, or that raters occasionally differed in judging whether minor deviations from clarity or conciseness were acceptable. This underlines the need for transparent, item-specific rubrics and calibrated exemplars when using genAI explanations in teaching or assessment.

Taken together, the results of this study have direct implications for medical education. With near-expert performance on a specialty examination and high-clarity rationales, modern LLMs can function as scalable, always-available tutors (Skryd & Lawrence, 2024; Wu et al., 2024; Scarlatos et al., 2025). The advanced genAI models can generate stepwise explanations, rehearse variant phrasings of the same concept, and adapt practice to a learner's current misconceptions (Pesovski et al., 2024; Mawarsih et al., 2025). For students with limited access to faculty time or commercial preparation resources, this capability could narrow achievement gaps (Khan et al., 2024; Schmidt et al., 2025). In microbiology specifically, LLMs can coach students through interpretation of culture results, antimicrobial resistance mechanisms, or syndromic differentials, accelerating movement from rote memorization to conceptual understanding.

In this study, because newer genAI model explanation quality was high when answers are correct, structured use of these tools (e.g., "explain, then verify with source material") can magnify educational value. Nevertheless, the same properties that make LLMs attractive tutors create challenges for high-stakes assessment (Richardson & Clesham, 2021; Córdova-Esparza, 2025). If genAI models can exceed average student performance on secure, well-constructed MCQs, the discriminatory power of those items to distinguish levels of human competence diminishes. Over-reliance on MCQs with predictable distractor patterns may invite "teaching to the model," in which item formats drift toward what genAI handles well. Moreover, fluent but occasionally inaccurate explanations can instill false confidence; learners may adopt plausible but incorrect heuristics that are hard to unlearn (Chelli et al., 2024; Martens et al., 2025).

The moderate inter-rater agreement on the relevance dimension for DeepSeek in this study could be a signal that even expert human supervisors may struggle to detect subtle issues in genAI output when the prose is polished. This finding calls for deliberate "AI-aware" assessment design (Karahan & Emekli, 2025). MCQs should increasingly privilege discrimination at higher cognitive levels, incorporate counterintuitive distractors that require integration across domains (e.g., microbiology with pharmacokinetics or infection-control logistics), and vary option length and structure to avoid spurious lexical cues. Sequential test formats that require reasoning across linked items, justification prompts that demand brief constructed responses, and viva-style defenses of diagnostic choices can surface reasoning quality that a single best-answer format may conceal.

Where feasible, performance tasks (e.g., interpreting antibiograms with patient-specific constraints) and Objective Structured Clinical Examination (OSCE)-style stations can complement MCQs to assess applied judgment less susceptible to current LLM strengths. The results support using LLMs as formative companions rather than as unsupervised arbiters of competence. Practical guardrails include (1) requiring students to annotate genAI outputs with cited course resources, (2) embedding "reflect-verify-revise" cycles in assignments, (3) logging prompts and rationales to make the learning process auditable, and (4) coaching learners to generate counter-explanations ("why the other options are wrong"). Faculty development should focus on interpreting genAI explanations, identifying subtle inaccuracies, and curating prompt templates that elicit transparent reasoning rather than purely declarative answers.

A central message of this study is the short shelf-life of conclusions about genAI capability. Two years transformed ChatGPT that underperformed relative to students into successors that outperform them. Institutions should therefore adopt scheduled benchmarking of commonly used genAI models on domain-specific item banks, with dashboards that track performance by cognitive level and content area. Thresholds can trigger item review (e.g., if a general-purpose model persistently exceeds 90% on an item, consider revising or retiring it) and inform exam security protocols. Because genAI model performance can drift with updates, documenting model identity and test dates—as done in this study—is essential for reproducibility. Although the examination language in this study was English, many programs teach and assess in multilingual contexts. Performance parity across languages cannot be assumed. Future work should replicate these analyses in Arabic and other languages to ensure that genAI-augmented learning does not exacerbate inequities for students who study or test in non-English environments (Weng & Fu, 2025). Additionally, specialty sub-domains underrepresented in training data (e.g., regional pathogen epidemiology) may show different performance profiles and deserve targeted benchmarking.

Limitations of the study should be acknowledged at this stage as follows. First, generalizability was constrained by a single institution, a single course, and an item pool dominated by virology. Second, the evaluation setting did not simulate full exam conditions for genAI tools (e.g., time limits, multimodal inputs). Third, while inter-rater reliability was high, the moderate agreement in one dimension for DeepSeek highlights the need for more granular scoring rubrics. Future studies should (1) expand to larger, multi-institutional item banks spanning bacteriology, mycology, and parasitology; (2) include constructed-response and sequential reasoning tasks; (3) evaluate robustness to adversarial distractors; and (4) examine the longitudinal impact of genAI-assisted study on independent human performance.

Conclusions

In medical microbiology, modern LLMs now perform at or above the level of average learners and provide high-quality explanations. Used well, genAI models can extend access, personalize study, and accelerate feedback. Used uncritically, the same models risk eroding assessment validity and fostering superficially plausible but fragile understanding. The educational task ahead is not to exclude these revolutionary and inevitable tools, but to utilize them with design, governance, and pedagogy that keep human judgment at the center—while acknowledging that the definition of "human-only" cognitive work is itself evolving. Continuous, specialty-level benchmarking—of the kind illustrated here—should become routine infrastructure for any program that seeks to integrate genAI into teaching and assessment responsibly.

Conflict of interest

The authors declare that they have no conflicts of interest in this work.

Funding

This research received no external funding.

Acknowledgement

ChatGPT-5 was used for language refinement (improving grammar, sentence structure, and readability of the manuscript) and technical writing assistance (providing suggestions for structuring complex technical descriptions more effectively). We confirm that all AI-assisted processes were critically reviewed by the authors to ensure the integrity and reliability of the results. The final decisions and interpretations presented in this article were solely made by the authors.

References

Abdaljaleel, M., Barakat, M., Alsanafi, M., Salim, N. A., Abazid, H., Malaeb, D*., et al.* (2024). A multinational study on the factors influencing university students' attitudes and usage of ChatGPT. *Scientific Reports, 14*(1), 1983. doi:10.1038/s41598-024-52549-8

Ateeq, A., Alzoraiki, M., & Milhem, M. (2024). Artificial intelligence in education: implications for academic integrity and the shift toward holistic assessment. *Frontiers in Education, 9*, 1470979. doi:10.3389/feduc.2024.1470979

Azaria, A., Azoulay, R., & Reches, S. (2023). ChatGPT is a Remarkable Tool—For Experts. *Data Intelligence, 6*, 1-49. doi:10.1162/dint_a_00235

Barakat, M., Salim, N. A., & Sallam, M. (2025). University Educators Perspectives on ChatGPT: A Technology Acceptance Model-Based Study. *Open Praxis, 17*(1), 129–144. doi:10.55982/openpraxis.17.1.718

Bharatha, A., Ojeh, N., Fazle Rabbi, A. M., Campbell, M. H., Krishnamurthy, K., Layne-Yarde, R. N. A*., et al.* (2024). Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy. *Adv Med Educ Pract, 15*, 393-400. doi:10.2147/amep.S457408

Bushuyev, S., Puziichuk, A., Bushueva, N., Bushuyeva, V., & Bushuyev, D. (2025). The evolving landscape of education under the influence of AI. *Bulletin of NTU KhPI Series Strategic Management Portfolio Program and Project Management*, 3-8. doi:10.20998/2413-3000.2024.9.1

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M*., et al.* (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *J Med Internet Res, 26*, e53164. doi:10.2196/53164

Córdova-Esparza, D.-M. (2025). AI-Powered Educational Agents: Opportunities, Innovations, and Ethical Challenges. *Information, 16*(6), 469. doi:10.3390/info16060469

Fu, Y., & Weng, Z. (2024). Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices. *Computers and Education: Artificial Intelligence, 7*, 100306. doi:10.1016/j.caeai.2024.100306

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A*., et al.* (2023). 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, 9*, e45312. doi:10.2196/45312

Gurajala, S. (2024). Artificial intelligence (AI) and medical microbiology: A narrative review. *Indian Journal of Microbiology Research, 11*, 156-162. doi:10.18231/j.ijmr.2024.029

Haugen, H. J., & de Lange, T. (2024). Multiple choice as formative assessment in dental education. *Eur J Dent Educ, 28*(3), 757-769. doi:10.1111/eje.13002

Herrmann-Werner, A., Festl-Wietek, T., Holderried, F., Herschbach, L., Griewatz, J., Masters, K*., et al.* (2024). Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study. *J Med Internet Res, 26*, e52113. doi:10.2196/52113

Hirani, R., Noruzi, K., Khuram, H., Hussaini, A. S., Aifuwa, E. I., Ely, K. E*., et al.* (2024). Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities. *Life (Basel), 14*(5), 557. doi:10.3390/life14050557

Holzinger, A., Saranti, A., Angerschmid, A., Finzel, B., Schmid, U., & Mueller, H. (2023). Toward human-level concept learning: Pattern benchmarking for AI algorithms. *Patterns (N Y), 4*(8), 100788. doi:10.1016/j.patter.2023.100788

Hu, C., Li, F., Wang, S., Gao, Z., Pan, S., & Qing, M. (2025). The role of artificial intelligence in enhancing personalized learning pathways and clinical training in dental education. *Cogent Education, 12*(1), 2490425. doi:10.1080/2331186X.2025.2490425

Jiang, Q., Gao, Z., & Karniadakis, G. (2025). DeepSeek vs. ChatGPT: A Comparative Study for Scientific Computing and Scientific Machine Learning Tasks. *arXiv*. doi:10.48550/arXiv.2502.17764

Jin, I., Tangsrivimol, J. A., Darzi, E., Hassan Virk, H. U., Wang, Z., Egger, J*., et al.* (2025). DeepSeek vs. ChatGPT: prospects and challenges. *Front Artif Intell, 8*, 1576992. doi:10.3389/frai.2025.1576992

Joshi, L. T. (2021). Using alternative teaching and learning approaches to deliver clinical microbiology during the COVID-19 pandemic. *FEMS Microbiol Lett, 368*(16). doi:10.1093/femsle/fnab103

Karahan, B. N., & Emekli, E. (2025). Comparison of applicability, difficulty, and discrimination indices of multiple-choice questions on medical imaging generated by different AI-based chatbots. *Radiography (Lond), 31*(5), 103087. doi:10.1016/j.radi.2025.103087

Katona, J., & Gyonyoru, K. I. K. (2025). AI-based Adaptive Programming Education for Socially Disadvantaged Students: Bridging the Digital Divide. *TechTrends*. doi:10.1007/s11528-025-01088-8

Khan, M. S., Umer, H., & Faruqe, F. (2024). Artificial intelligence for low income countries. *Humanities and Social Sciences Communications, 11*(1), 1422. doi:10.1057/s41599-024-03947-w

Kim, J., Yu, S., Detrick, R., & Li, N. (2025). Exploring students' perspectives on Generative AI-assisted academic writing.

Education and Information Technologies, 30(1), 1265-1300. doi:10.1007/s10639-024-12878-7

Kovalainen, T., Pramila-Savukoski, S., Kuivila, H.-M., Juntunen, J., Jarva, E., Rasi, M., *et al.* (2025). Utilising artificial intelligence in developing education of health sciences higher education: An umbrella review of reviews. *Nurse Education Today, 147*, 106600. doi:10.1016/j.nedt.2025.106600

Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review, 57*(9), 243. doi:10.1007/s10462-024-10896-y

Martens, D., Shmueli, G., Evgeniou, T., Bauer, K., Janiesch, C., Feuerriegel, S., *et al.* (2025). Beware of “Explanations” of AI. *arXiv*. doi:10.48550/arXiv.2504.06791

Matarazzo, A., & Torlone, R. (2025). A Survey on Large Language Models with some Insights on their Capabilities and Limitations. *arXiv*. doi:10.48550/arXiv.2501.04040

Mawarsih, P. B., Nadzifah, H., Puspa Widuri, A. W., & Kurniawati, E. (2025). Generative AI in higher education: the ChatGPT effect. *Asia Pacific Journal of Education*, 1-3. doi:10.1080/02188791.2024.2420309

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT. *Education Sciences, 13*(9), 856. doi:10.3390/educsci13090856

Mirea, C.-M., Bologa, R., Toma, A., Clim, A., Plăcintă, D.-D., & Bobocea, A. (2025). Transforming Learning with Generative AI: From Student Perceptions to the Design of an Educational Solution. *Applied Sciences, 15*(10), 5785. doi:10.3390/app15105785

Mohseni, P., & Ghorbani, A. (2024). Exploring the synergy of artificial intelligence in microbiology: Advancements, challenges, and future prospects. *Computational and Structural Biotechnology Reports, 1*, 100005. doi:10.1016/j.csbr.2024.100005

Monrad, S., Zaidi, L., Grob, K., Kurtz, J., Tai, A., Hortsch, M., *et al.* (2021). What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy. *Medical teacher, 43*, 1-12. doi:10.1080/0142159X.2021.1879376

Nelson, A. S., Santamaría, P. V., Javens, J. S., & Ricaurte, M. (2025). Students’ Perceptions of Generative Artificial Intelligence (GenAI) Use in Academic Writing in English as a Foreign Language. *Education Sciences, 15*(5), 611. doi:10.3390/educsci15050611

Newton, P., & Xiromeriti, M. (2024). ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assessment & Evaluation in Higher Education, 49*(6), 781-798. doi:10.1080/02602938.2023.2299059

Newton, P. M. (2020). Guidelines for Creating Online MCQ-Based Exams to Evaluate Higher Order Learning and Reduce Academic Misconduct. In S. E. Eaton (Ed.), *Handbook of Academic Integrity* (pp. 1-17). Singapore: Springer Nature Singapore.

Oyekunle, D., Nwaiku, M., Matthew, U., Onyedibe, N., Onyedibe, O., Nwanakwaugwu, A., *et al.* (2024). Transition to Sustainable Human-Centric Education in Emerging Artificial Intelligence Industry 5.0: Conversational AI With User-Centric ChatGPT-5. In (pp. 37-76).

Parekh, P., & Bahadoor, V. (2024). The Utility of Multiple-Choice Assessment in Current Medical Education: A Critical Review. *Cureus, 16*(5), e59778. doi:10.7759/cureus.59778

Parthasarathy, V., Zafar, A., Khan, A., & Shahid, A. (2024). The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An

Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. *arXiv*. doi:10.48550/arXiv.2408.13296

Parveen, D., & Ramzan, S. (2024). The Role of Digital Technologies in Education: Benefits and Challenges. *International Research Journal on Advanced Engineering and Management (IRJAEM), 2*, 2029-2037. doi:10.47392/IRJAEM.2024.0299

Pesovski, I., Santos, R., Henriques, R., & Trajkovik, V. (2024). Generative AI for Customizable Learning Experiences. *Sustainability, 16*, 3034. doi:10.3390/su16073034

Rajaram, K. (2023). Future of Learning: Teaching and Learning Strategies. In K. Rajaram (Ed.), *Learning Intelligence: Innovative and Digital Transformative Learning Strategies: Cultural and Social Engineering Perspectives* (pp. 3-53). Singapore: Springer Nature Singapore.

Richardson, M., & Clesham, R. (2021). Rise of the machines? The evolving role of Artificial Intelligence (AI) technologies in high stakes assessment. *London Review of Education, 19*. doi:10.14324/LRE.19.1.09

Rodger, D., Mann, S. P., Earp, B., Savulescu, J., Bobier, C., & Blackshaw, B. P. (2025). Generative AI in healthcare education: How AI literacy gaps could compromise learning and patient safety. *Nurse Education in Practice, 87*, 104461. doi:10.1016/j.nepr.2025.104461

Rony, M. K. K., Parvin, M. R., Wahiduzzaman, M., Debnath, M., Bala, S. D., & Kayesh, I. (2024). "I Wonder if my Years of Training and Expertise Will be Devalued by Machines": Concerns About the Replacement of Medical Professionals by Artificial Intelligence. *SAGE Open Nurs, 10*, 23779608241245220. doi:10.1177/23779608241245220

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of applied learning and teaching, 6*(1), 342-363. doi:10.37074/jalt.2023.6.1.9

Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel), 11*(6), 887. doi:10.3390/healthcare11060887

Sallam, M., Al-Mahzoum, K., Almutawaa, R. A., Alhashash, J. A., Dashti, R. A., AlSafy, D. R.*, et al.* (2024a). The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-choice questions: a comparative analysis of English and Arabic responses. *BMC Research Notes, 17*(1), 247. doi:10.1186/s13104-024-06920-7

Sallam, M., Al-Mahzoum, K., Eid, H., Al-Salahat, K., Sallam, M., Ali, G.*, et al.* (2025a). Chinese Generative AI Models Challenge Western AI in Clinical Chemistry MCQs: A Benchmarking Follow-up Study on AI Use in Health Education. *Babylonian Journal of Artificial Intelligence, 2025*, 1-14. doi:10.58496/BJAI/2025/001

Sallam, M., Al-Mahzoum, K., Sallam, M., & Mijwil, M. M. (2025b). DeepSeek: Is it the End of Generative AI Monopoly or the Mark of the Impending Doomsday? *Mesopotamian Journal of Big Data, 2025*, 26-34. doi:10.58496/MJBD/2025/002

Sallam, M., & Al-Salahat, K. (2023). Below average ChatGPT performance in medical microbiology exam compared to university students. *Frontiers in Education, 8*, 1333415. doi:10.3389/feduc.2023.1333415

Sallam, M., Al-Salahat, K., & Al-Ajlouni, E. (2023a). ChatGPT Performance in Diagnostic Clinical Microbiology Laboratory-Oriented Case Scenarios. *Cureus, 15*(12), e50629. doi:10.7759/cureus.50629

Sallam, M., Al-Salahat, K., Eid, H., Egger, J., & Puladi, B. (2024b). Human versus Artificial Intelligence: ChatGPT-4 Outperforming

Bing, Bard, ChatGPT-3.5 and Humans in Clinical Chemistry Multiple-Choice Questions. *Adv Med Educ Pract, 15*, 857-871. doi:10.2147/amep.S479801

Sallam, M., Barakat, M., & Sallam, M. (2023b). Pilot Testing of a Tool to Standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-Based Models. *Cureus, 15*(11), e49373. doi:10.7759/cureus.49373

Sallam, M., Barakat, M., & Sallam, M. (2024c). A Preliminary Checklist (METRICS) to Standardize the Design and Reporting of Studies on Generative Artificial Intelligence-Based Models in Health Care Education and Practice: Development Study Involving a Literature Review. *Interact J Med Res, 13*, e54704. doi:10.2196/54704

Sallam, M., Khalil, R., & Sallam, M. (2024d). Benchmarking Generative AI: A Call for Establishing a Comprehensive Framework and a Generative AIQ Test. *Mesopotamian Journal of Artificial Intelligence in Healthcare, 2024*, 69-75. doi:10.58496/MJAIH/2024/010

Sallam, M., Salim, N. A., Barakat, M., & Al-Tammemi, A. B. (2023c). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J, 3*(1), e103. doi:10.52225/narra.v3i1.103

Sallam, M., & Sallam, M. (2025). Ethical aspects of implementing generative artificial intelligence in medical education: a narrative review. *History and Philosophy of Medicine, 7*, 18–25. doi:10.53388/HPM2025020

Scarlatos, A., Liu, N., Lee, J., Baraniuk, R., & Lan, A. (2025). Training LLM-based Tutors to Improve Student Learning Outcomes in Dialogues. *arXiv*. doi:10.48550/arXiv.2503.06424

Schmidt, D. A., Alboloushi, B., Thomas, A., & Magalhaes, R. (2025). Integrating artificial intelligence in higher education: perceptions, challenges, and strategies for academic innovation. *Computers and Education Open, 9*, 100274. doi:10.1016/j.caeo.2025.100274

Sharma, S., Mittal, P., Kumar, M., & Bhardwaj, V. (2025). The role of large language models in personalized learning: a systematic review of educational impact. *Discover Sustainability, 6*(1), 243. doi:10.1007/s43621-025-01094-z

Singh, S. P., & Nagmoti, J. M. (2021). Strengthening clinical microbiology skill acquisition; a nationwide survey of faculty perceptions & practices on teaching & assessment of practical skills to undergraduate students. *Indian Journal of Medical Microbiology, 39*(2), 154-158. doi:10.1016/j.ijmmb.2020.11.003

Skryd, A., & Lawrence, K. (2024). ChatGPT as a Tool for Medical Education and Clinical Decision-Making on the Wards: Case Study. *JMIR Form Res, 8*, e51346. doi:10.2196/51346

Storey, V. C., Yue, W. T., Zhao, J. L., & Lukyanenko, R. (2025). Generative Artificial Intelligence: Evolving Technology, Growing Societal Impact, and Opportunities for Information Systems Research. *Information Systems Frontiers*. doi:10.1007/s10796-025-10581-7

Tan, X., Cheng, G., & Ling, M. H. (2025). Artificial intelligence in teaching and teacher professional development: A systematic review. *Computers and Education: Artificial Intelligence, 8*, 100355. doi:10.1016/j.caeai.2024.100355

Trikoili, A., Georgiou, D., Pappa, C. I., & Pittich, D. (2025). Critical Thinking Assessment in Higher Education: A Mixed-Methods Comparative Analysis of AI and Human Evaluator. *International Journal of Human–Computer Interaction*, 1-14. doi:10.1080/10447318.2025.2499164

Vieriu, A. M., & Petrea, G. (2025). The Impact of

Artificial Intelligence (AI) on Students' Academic Development. *Education Sciences, 15*(3), 343. doi:10.3390/educsci15030343

Weng, Z., & Fu, Y. (2025). Generative AI in Language Education: Bridging Divide and Fostering Inclusivity. *International Journal of Technology in Education, 8*, 395-420. doi:10.46328/ijte.1056

Wong, W. K. O. (2024). The sudden disruptive rise of generative artificial intelligence? An evaluation of their impact on higher education and the global workplace. *Journal of Open Innovation: Technology, Market, and Complexity, 10*(2), 100278. doi:10.1016/j.joitmc.2024.100278

Wu, Y., Zheng, Y., Feng, B., Yang, Y., Kang, K., & Zhao, A. (2024). Embracing ChatGPT for Medical Education: Exploring Its Impact on Doctors and Medical Students. *JMIR Med Educ, 10*, e52483. doi:10.2196/52483

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*(1), 40. doi:10.1186/s41239-024-00468-z

Ying, L., Collins, K., Wong, L., Sucholutsky, I., Liu, R., Weller, A., *et al.* (2025). On Benchmarking Human-Like Intelligence in Machines. *arXiv*. doi:10.48550/arXiv.2502.20502

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: a threat to academic integrity or reformation? Evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education, 21*(1), 21. doi:10.1186/s41239-024-00453-6

Zhu, Y. (2025). Revolutionizing simulation-based clinical training with AI: Integrating FASSLING for enhanced emotional intelligence and therapeutic competency in clinical psychology education. *Journal of Clinical Technology and Theory, 2*, 38-54. doi:10.54254/3049-5458/2025.21247

How to Cite: Sallam, M., Irshaid, A., Snygg, J., Albadri, R., & Sallam, M. (2025).Rapid Evolution of Large Language Models in Medical Education: Comparative Performance of ChatGPT-3.5, ChatGPT-5, and DeepSeek on Medical Microbiology MCQs. *Contemporary Education and Teaching Research*, *06*(8), 295-309. https://doi.org/10.61360/BoniCETR252018770801