Babylonian Journal of Artificial Intelligence
Vol. **(2025)**, 2025, **pp**. 1–14
DOI: https://doi.org/10.58496/BJAI/2025/001 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BJAI

Research Article

Chinese Generative AI Models Challenge Western AI in Clinical Chemistry MCQs: A Benchmarking Follow-up Study on AI Use in Health Education

Malik Sallam[1,2,*], Kholoud Al-Mahzoum[3], Huda Eid[4], Khaled Al-Salahat[5], Mohammed Sallam[6,7,8,9], Guma Ali[10], Maad M. Mijwil[11,12]

[1] *Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Amman 11942, Jordan*
[2] *Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Amman 11942, Jordan*
[3] *Sheikh Jaber Al-Ahmad Al-Sabah Hospital, Ministry of Health, Kuwait City, Kuwait*
[4] *Danube Private University, Steiner Landstraße 124, Krems-Stein, 3500, Austria*
[5] *Faculty of Medicine, Aqaba Medical Sciences University (AMSU), Aqaba, Jordan*
[6] *Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates*
[7] *Department of Management, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates*
[8] *Department of Management, School of Business, International American University, Los Angeles, CA 90010, United States*
[9] *College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai, United Arab Emirates*
[10] *Department of Computer and Information Science, Faculty of Technoscience, Muni University, Arua, Uganda*
[11] *College of Administration and Economics, Al-Iraqia University, Baghdad, Iraq*
[12] *Computer Techniques Engineering Department, Baghdad College of Economic Sciences University, Baghdad, Iraq.*

ARTICLE INFO

Article History
Received 26 Dec 2024
Revised: 18 Jan 2025
Accepted 02 Feb 2025
Published 08 Feb 2025

Keywords
AI
Benchmarking
LLM
DeepSeek
Qwen

ABSTRACT

Background: The emergence of Chinese generative AI (genAI) models, such as DeepSeek and Qwen, has introduced strong competition to Western genAI models. These advancements hold significant potential in healthcare education. However, benchmarking the performance of genAI models in specialized medical disciplines is crucial to assess their strengths and limitations. This study builds on prior research evaluating ChatGPT (GPT-3.5 and GPT-4), Bing, and Bard against human postgraduate students in Medical Laboratory Sciences, now incorporating DeepSeek and Qwen to assess their effectiveness in Clinical Chemistry Multiple-Choice Questions (MCQs).
Methods: This study followed the METRICS framework for genAI-based healthcare evaluations, assessing six models using 60 Clinical Chemistry MCQs previously administered to 20 MSc students. The facility index and Bloom's taxonomy classification were used to benchmark performance. GenAI models included DeepSeek-V3, Qwen 2.5-Max, ChatGPT-4, ChatGPT-3.5, Microsoft Bing, and Google Bard, evaluated in a controlled, non-interactive environment using standardized prompts.
Results: The evaluated genAI models showed varying accuracy across Bloom's taxonomy levels. DeepSeek-V3 (0.92) and ChatGPT-4 (1.00) outperformed humans (0.74) in the Remember category, while Qwen 2.5-Max (0.94) and ChatGPT-4 (0.94) surpassed human performance (0.61) in the Understand category. ChatGPT-4 (+23.25%, $p < 0.001$), DeepSeek-V3 (+18.25%, $p = 0.001$), and Qwen 2.5-Max (+18.25%, $p = 0.001$) significantly outperformed human students. Decision tree analysis identified cognitive category as the strongest predictor of genAI accuracy ($p < 0.001$), with Chinese AI models performing comparably to ChatGPT-4 in lower-order tasks but exhibiting lower accuracy in higher-order domains.
Conclusions: The findings highlighted the growing capabilities of Chinese genAI models in healthcare education, proving that DeepSeek and Qwen can compete with, and in some areas outperform, Western genAI models. However, their relative weakness in higher-order reasoning raises concerns about their ability to fully replace human cognitive processes in clinical decision-making. As genAI becomes increasingly integrated into health education, concerns regarding academic integrity, genAI dependence, and the validity of MCQ-based assessments must be addressed. The study underscores the need for a re-evaluation of medical assessment strategies, ensuring that students develop critical thinking skills rather than relying on genAI for knowledge retrieval.

*Corresponding author. Email: malik.sallam@ju.edu.jo

1. INTRODUCTION

It is no exaggeration to say that healthcare education stands at the threshold of an epistemological revolution, one as profound as the transition from oral traditions to the written word or from handwritten manuscripts to the printing press [1]. The advent of generative artificial intelligence (genAI) does not merely introduce an incremental advancement in educational methodology; rather, it challenges the very structure of how medical knowledge is acquired, processed, and applied [2-4]. The classical pedagogical model of medical training, rooted in the painstaking accumulation of vast reservoirs of knowledge—memorized, internalized, and then regurgitated in the crucible of clinical practice—is undergoing a transformation of historical significance [5, 6].

For centuries, medical education was an exercise in intellectual endurance, demanding mastery of massive volumes of text, from *Gray's Anatomy* to *Harrison's Principles of Internal Medicine*, while relying on the traditional authority of professors and textbooks as the primary dispensers of knowledge [7]. The process was rigid, hierarchical, and unyielding [8]. One either learned the information through relentless repetition and clinical apprenticeship, or one risked falling behind [9]. But now, that time-honored structure is being dismantled—if not outright subverted—by artificial intelligence (AI) [3, 4, 10]. The emergence of OpenAI's ChatGPT, Microsoft's Bing/Copilot, and Google's Bard/Gemini has redefined the role of expertise in healthcare education and practice [11, 12]. No longer is the process of learning confined to passive absorption; rather, it has become an interactive dialogue between the learner and an AI system capable of mimicking human-like reasoning, responding to queries, and generating tailored educational content at an unprecedented scale [3, 13]. For example, a medical student struggling to grasp the nuances of metabolic pathways no longer needs to sift through pages of biochemistry textbooks; instead, a well-structured prompt to a genAI model produces a synthesized, contextually relevant, and dynamically adjusted response.

The AI transformation of healthcare education mirrors past disruptions in knowledge transmission—not unlike the impact of the Renaissance, when scholars transitioned from reliance on ancient authorities to direct engagement with primary sources [14]. The AI-driven democratization of information in medicine has similar implications: it flattens hierarchies, accelerates learning, and grants individuals access to an ever-expanding reservoir of knowledge without the mediation of traditional gatekeepers [15]. Yet, for all its promise, this AI revolution is not without its perils [12]. The inherent nature of AI—its reliance on statistical probabilities rather than genuine understanding—raises concerns about factual accuracy, bias, and over-reliance on machine-generated knowledge [3, 12, 16-18]. Thus, the integration of genAI into healthcare education is not a passing trend; it is a defining moment in the evolution of medical pedagogy [19]. Whether it heralds a golden age of accelerated learning or a crisis of epistemological erosion depends not on the technology itself but on how wisely and critically it is employed [20].

For much of the modern technological era, the dominance of Western AI models has been so complete that it was taken as a given, a natural extension of Silicon Valley's monopoly over the digital revolution [21-23]. OpenAI, Microsoft, and Google, fortified by vast computational resources and a firm grip on global AI research, had seemingly cemented their position as the uncontested arbiters of genAI [24, 25]. However, history seldom tolerates unilateral dominance for long, and as with the once-unchallenged supremacy of Western industrial power, so too is the genAI frontier witnessing a formidable challenge from the East. The emergence of Chinese genAI models, particularly DeepSeek and Qwen by Alibaba Cloud, marks a seismic shift in the genAI landscape [26]. These models are not mere imitations or belated attempts to replicate Western advancements. Instead, they represent a parallel and increasingly competitive genAI ecosystem, one that has grown not under the aegis of Silicon Valley but within China's vast and meticulously cultivated AI infrastructure, supported by state-led innovation initiatives and an increasingly sophisticated tech sector [27].

Crucially, the Chinese genAI models not only match their Western counterparts in performance—a feat that in itself signals a departure from past asymmetries—but they also introduce a game-changing element: democratization of access with transparency of AI algorithms [26-28]. Whereas OpenAI's GPT-4, Google's Gemini, and Microsoft's Copilot operate within heavily restricted ecosystems, DeepSeek and Qwen have adopted an open-access model, allowing researchers, students, and professionals to utilize cutting-edge AI technology without the financial and institutional barriers imposed by Western firms [26, 29]. The implications would be profound since the free availability of advanced genAI models disrupts the monopoly once held by tech conglomerates, placing transformative computational power into the hands of a broader, global audience. This shift is not merely technical; it is geopolitical. The long-standing perception of Western AI models as the pinnacle of AI development is being fundamentally challenged [26, 29]. The question is no longer whether China's AI revolution will reshape the digital landscape but rather to what extent Western dominance will be eroded as a result [26, 27, 29].

The critical evaluation of genAI models in education cannot be left to vague impressions or anecdotal assessments; it demands a rigorous, structured approach, one that dissects their cognitive capabilities with the same precision that has long defined academic scholarship [30, 31]. In this regard, the Bloom's revised taxonomy—an intellectual framework designed to classify human learning objectives—serves as both a yardstick and a challenge to AI's claim to cognitive prowess [32-35]. Developed as a means to systematize the progression of human knowledge acquisition, the Bloom's taxonomy organizes

cognition into a hierarchy: Remember, Understand, Apply, Analyze, Evaluate, and Create [36, 37]. This structure, first introduced in the mid-20th century, was not merely an academic exercise but a response to the evolving needs of education in the post-industrial world [38-40]. It was built on the recognition that true expertise is not a matter of rote memorization but a gradual ascent toward deeper intellectual synthesis [37].

To measure a genAI tool against Bloom's taxonomy framework is to subject it to the same standards once reserved exclusively for human intelligence. At the lowest rung—Remember— the genAI models would be expected to excel effortlessly, retrieving definitions, listing biochemical pathways, or recalling historical medical discoveries with near-perfect accuracy. This would be neither surprising nor particularly revelatory; after all, even the scribes of ancient Alexandria could transcribe knowledge without necessarily understanding it. But what distinguishes human cognition is the ability not merely to store and recall but to interpret, challenge, and innovate [41]. Moving upward, the Understand and Apply levels demand that AI demonstrate comprehension beyond repetition. It must take a concept such as the biochemical basis of metabolic acidosis and explain it in varied contexts or apply it to novel patient cases. While genAI models have shown remarkable proficiency here, particularly in structured multiple-choice questions, the cracks begin to emerge when complexity deepens and uncertainty is introduced as shown in various contexts (e.g., Medical microbiology, Clinical Chemistry) [35, 42].

It is at the Analyze, Evaluate, and Create levels that the great schism between human intelligence and artificial intelligence is expected to be apparent. Evaluation—requiring critical judgment, weighing competing hypotheses, and making reasoned clinical decisions based on conflicting data—has long been considered an innately human domain [43, 44]. Creation, meanwhile, represents the pinnacle of cognitive function: the synthesis of knowledge into something novel, insightful, and contextually rich [45]. To ask a genAI model to engage in true creativity is akin to asking a printing press to author an epic. It may rearrange words in aesthetically pleasing ways, but it lacks the lived experience, intuition, and intellectual defiance that have historically defined human innovation [46, 47]. This is not to say that AI cannot mimic aspects of creativity—it can, and sometimes with unnerving effectiveness; however, the distinction lies in origination versus recombination [46, 47]. Just as a skilled forger may reproduce the brushstrokes of a Renaissance master without capturing the essence of artistic intent, AI may generate text, code, or solutions that appear insightful but lack the foundational experience and intentionality that mark true human intelligence [48].

Thus, to subject genAI models to Bloom's taxonomy is not merely an academic exercise but a statement on the nature of intelligence itself. If genAI surpasses students in lower-order cognitive tasks yet struggles in higher-order reasoning, does this represent a triumph or a limitation? If AI can outperform humans on standardized exams yet falters in clinical judgment, does this indicate an augmentation of medical education or a dilution of expertise? These are not trivial questions. They are, in fact, the same questions that arose with every major technological leap in intellectual history—from the invention of writing, which Socrates feared would erode memory, to the printing press, which revolutionized access to knowledge while also unleashing an era of misinformation and dogmatic rigidity [49, 50]. To benchmark AI against human cognition is not merely an academic necessity; it is an intellectual reckoning with what it means to think, to reason, and ultimately, to create [31].

The application of genAI in Clinical Chemistry is therefore not a mere academic exercise; it is a critical stress test of genAI's capacity to engage with medical knowledge in both its most static and dynamic forms. Clinical Chemistry presents a unique intellectual battleground—a discipline that demands not only rote memorization of biochemical principles but also the nuanced application of those principles in clinical contexts. It is here, at the intersection of knowledge retention and problem-solving, that the true measure of an AI tool educational potential can be assessed. Therefore, this study builds upon previous research, which demonstrated that ChatGPT-4 outperformed its predecessors, while ChatGPT-3.5, Bing, and Bard displayed above-average competence [42]. The earlier findings were notable not only because these genAI models matched or exceeded postgraduate students in Clinical Chemistry multiple-choice questions (MCQs) but also because they successfully navigated diagnostic scenarios—interpreting simulated laboratory results to determine hypothetical patient conditions [42]. Such performance raised fundamental questions about AI's evolving role in healthcare education, particularly regarding academic integrity and the appropriateness of traditional assessment methods [51]. If AI can outperform human students on MCQs—the bedrock of standardized medical assessment—then should medical education rely on these tools, or must it evolve beyond them [52]. It is within this context of disruption and recalibration that the present study seeks to extend the investigation, incorporating new genAI models from China—DeepSeek and Qwen—into the comparative framework. The emergence of Chinese genAI models represents not just a technological shift but a geopolitical and educational one, challenging the long-held dominance of Western AI and offering free, unrestricted access to high-caliber large language models (LLMs) [26]. By subjecting these models to the same rigorous Clinical Chemistry MCQs, designed to span the breadth of Bloom's taxonomy, the study aims to benchmark their performance against Western genAI counterparts. The objective is not merely to determine which model scores highest, but rather to understand where, how, and why certain genAI models excel or falter in specialized healthcare education.

2. METHODS

2.1 Study design and ethics

This study was meticulously designed in accordance with the METRICS (Model, Evaluation, Timing, Range/ Randomization, Individual factors, Count, and Specificity of prompts and language) checklist, to ensure a methodologically rigorous framework for assessing the performance of six different genAI models in a healthcare education setting [53]. The study was based on the utilization of a set of 60 MCQs from a Clinical Chemistry examination administered to masters (MSc) students enrolled in the Medical Laboratory Sciences program at the School of Medicine, the University of Jordan. These MCQs were previously employed in a prior benchmarking study to test four different genAI models [42]. The primary reference dataset for AI benchmarking was derived from an in-person examination taken by 20 postgraduate students during the Autumn Semester of the 2019/2020 academic year. This cohort served as a real-world comparator, enabling a direct performance assessment between genAI models and human students.

The MCQs were developed by the first author (Malik Sallam), a Clinical Pathology consultant certified by the Jordan Medical Council, with extensive teaching experience in Clinical Chemistry since the 2018/2019 academic year. The use of originally authored questions mitigated concerns regarding intellectual property and copyright violations, ensuring that the study adhered to the highest academic standards as stated in [42]. To study received Institutional Review Board (IRB) approval from the Deanship of Scientific Research, the University of Jordan (Reference Number: 2/2024/19), with a waiver granted due to the non-sensitive, anonymized nature of the dataset [42].

2.2 METRICS features used to guide the study design

To ensure robust performance benchmarking, 60 MCQs from a Clinical Chemistry examination for MSc students at the University of Jordan, School of Medicine, were employed. The facility index (FI), defined as the proportion of students who correctly answered a given MCQ out of the total 20-student cohort [42]. The cognitive complexity of each MCQ was categorized using Bloom's revised taxonomy into four levels: Remember, Understand, Apply, and Analyze [37]. This classification was achieved through consensus between two authors (Malik Sallam and Khaled Al-Salahat) as stated in [42], both certified Clinical Pathologists with extensive teaching and assessment experience.

Six genAI models were tested as follows: ChatGPT-3.5 (OpenAI, San Francisco, CA): Evaluated using its default settings (last update: January 2022) [54]. ChatGPT-4 (OpenAI, San Francisco, CA): Latest version assessed at April 2023 update [54]. Bing Chat (GPT-4 Turbo) (Microsoft, Redmond, WA): Tested with its "Balanced" conversation mode, last updated April 2023 [55]. Bard (Gemini 1.0) (Google, Mountain View, CA): Version assessed at October 4, 2023 update [56]. DeepSeek V3 (DeepSeek, China): Evaluated in January 2025 [57, 58]. Qwen 2.5-Max (Alibaba Cloud, China): Evaluated in January 2025 [59]. Each genAI model was engaged in a controlled, non-interactive testing environment, ensuring uniformity across evaluations: GPT-3.5, GPT-4, Bard, DeepSeek, and Qwen 2.5-Max were tested using a single-page input system to maintain consistency. Bing Chat was tested using the "New Topic" option for each question to mitigate context retention bias, given the model's response limit (50 exchanges per session). No model was permitted to "regenerate" responses, nor was feedback provided during interactions, preventing adaptive learning bias that could skew results.

To ensure clarity, precision, and reproducibility, a standardized prompt was designed for genAI testing using the following exact phrasing: "*For the following 60 Clinical Chemistry MCQs that will be provided one by one, please select the most appropriate letter only for each MCQ without explanation. Please note that only one choice is correct while the other four choices are incorrect. Please note that these questions were designed for master's students in medical laboratory sciences*." Each MCQ was subsequently inputted sequentially, with no alterations or additional context provided. English was exclusively used, as it is the official language of instruction in the MSc program at the University of Jordan as stated in [42]. The genAI-generated responses were assessed for correctness against the pre-established answer key. Human involvement in evaluation was limited to verifying accuracy, ensuring objective assessment without subjective interpretation. All MCQs were originally authored by the first author (Malik Sallam), the sole instructor of the Clinical Chemistry course, ensuring full control over content validity and eliminating concerns related to copyright infringement. The material taught in the course was derived from the following textbooks: Tietz Textbook of Clinical Chemistry and Molecular Diagnostics [60]; Clinical Chemistry: Principles, Techniques, and Correlations [61]; Henry's Clinical Diagnosis and Management by Laboratory Methods [62]. The MCQs encompassed a comprehensive range of Clinical Chemistry topics, reflecting the curricular scope of the MSc program: Adrenal Function; Amino Acids and Proteins; Body Fluid Analysis; Clinical Enzymology; Electrolytes; Gastrointestinal Function; Gonadal Function; Liver Function; Nutritional Assessment; Pancreatic Function; Pituitary Function; Thyroid Gland; and Trace Elements as stated in [42].

2.3 Statistical analysis

All statistical analyses were performed using IBM SPSS Statistics (version 26.0; IBM Corp, Armonk, NY). Descriptive statistics were used to summarize genAI performance across the revised Bloom's taxonomy levels. Paired-samples t-tests

were conducted to compare genAI model accuracy with human performance in Clinical Chemistry MCQs, reporting 95% confidence intervals (CIs) and *p* values to assess statistical significance. Bonferroni post hoc comparisons were applied to evaluate pairwise differences among AI models while controlling for multiple comparisons. To assess genAI model performance across cognitive domains, Mann-Whitney *U* tests were used to compare lower-order (Remember, Understand) and higher-order (Apply, Analyze) cognitive tasks, with statistical significance set at $p < 0.05$. Pearson correlation analyses were conducted to explore associations between genAI accuracy and MCQ complexity factors, including stem word count and choice word count, with correlation significance determined at $p < 0.05$. A decision tree analysis using the Chi-squared Automatic Interaction Detection (CHAID) method was conducted to identify key predictors of genAI accuracy, stratifying performance based on cognitive category and genAI model type. The model's risk estimate and classification accuracy were calculated to assess predictive reliability. Statistical significance was set at $\alpha = 0.05$, with adjustments applied for multiple comparisons where appropriate.

3. RESULTS

3.1 Benchmarking of genAI performance compared to humans

Generative AI models exhibited varying performance across Bloom's taxonomy levels (**TABLE I**). DeepSeek-V3 (0.92) and ChatGPT-4 Legacy Model (1.00) achieved the highest accuracy in the Remember category, exceeding human performance (0.74). In the Understand category, Qwen 2.5-Max (0.94) and ChatGPT-4 (0.94) outperformed humans (0.61). Microsoft Bing (0.80) demonstrated the highest accuracy in the Apply category, while ChatGPT-3.5 (0.40) had the lowest. In the Analyze category, Qwen 2.5-Max (0.77) and ChatGPT-4 (0.77) outperformed humans (0.60), whereas Google Bard (0.38) showed the lowest accuracy (**TABLE I**).

TABLE I. OVERALL PERFORMANCE OF GENERATIVE AI MODELS STRATIFIED BASED ON THE REVISED BLOOM'S TAXONOMY CATEGORIES

Revised Bloom Taxonomy	**Remember**	**Understand**	**Apply**	**Analyze**
GenAI[a] model/human	**Mean**	**Mean**	**Mean**	**Mean**
DeepSeek-V3	0.92	1.00	0.60	0.62
Qwen 2.5-Max	0.87	0.94	0.60	0.77
ChatGPT-3.5	0.79	0.83	0.40	0.62
ChatGPT-4 Legacy Model	1.00	0.94	0.60	0.77
Microsoft Bing	0.83	0.83	0.80	0.54
Google Bard	0.75	0.78	0.60	0.38
Humans	0.74	0.61	0.71	0.60

[a.] genAI: Generative AI

Paired-samples t-tests assessed differences in accuracy between genAI models and human performance in Clinical Chemistry MCQs. DeepSeek-V3 (+18.25%, 95% CI: 8.11% to 28.39%, $t(59) = 3.601$, $p = 0.001$), Qwen 2.5-Max (+18.25%, 95% CI: 7.57% to 28.94%, $t(59) = 3.418$, $p = 0.001$), and ChatGPT-4 Legacy Model (+23.25%, 95% CI: 13.65% to 32.85%, $t(59) = 4.846$, $p < 0.001$) demonstrated significantly higher accuracy than human students. ChatGPT-3.5 (+6.58%, 95% CI: -6.59% to 19.75%, $t(59) = 1.00$, $p = 0.321$), Microsoft Bing (+9.92%, 95% CI: –1.44% to 21.27%, $t(59) = 1.747$, $p = 0.086$), and Google Bard (–0.08%, 95% CI: –12.39% to 12.22%, $t(59) = –0.014$, $p = 0.989$) showed no statistically significant difference from human performance.

Bonferroni post hoc comparisons indicated that ChatGPT-4 Legacy Model achieved significantly higher accuracy than Google Bard ($p = 0.023$), while no other genAI model comparisons showed statistically significant differences ($p > 0.05$). Pairwise differences between DeepSeek-V3, Qwen 2.5-Max, ChatGPT-3.5, Microsoft Bing, and Google Bard were non-significant across all comparisons.

3.2 GenAI model performance based on Bloom's taxonomy

Mann-Whitney U tests showed significant differences in correct answers between cognitive categories for DeepSeek-V3 ($U = 249.000$, $Z = –3.364$, $p = 0.001$), ChatGPT-3.5 ($U = 282.000$, $Z = –2.022$, $p = 0.043$), ChatGPT-4 Legacy Model ($U = 282.000$, $Z = –2.980$, $p = 0.003$), and Google Bard ($U = 258.000$, $Z = –2.370$, $p = 0.018$). Qwen 2.5-Max ($p = 0.072$) and Microsoft Bing ($p = 0.064$) showed no significant differences (**Fig. 1**).

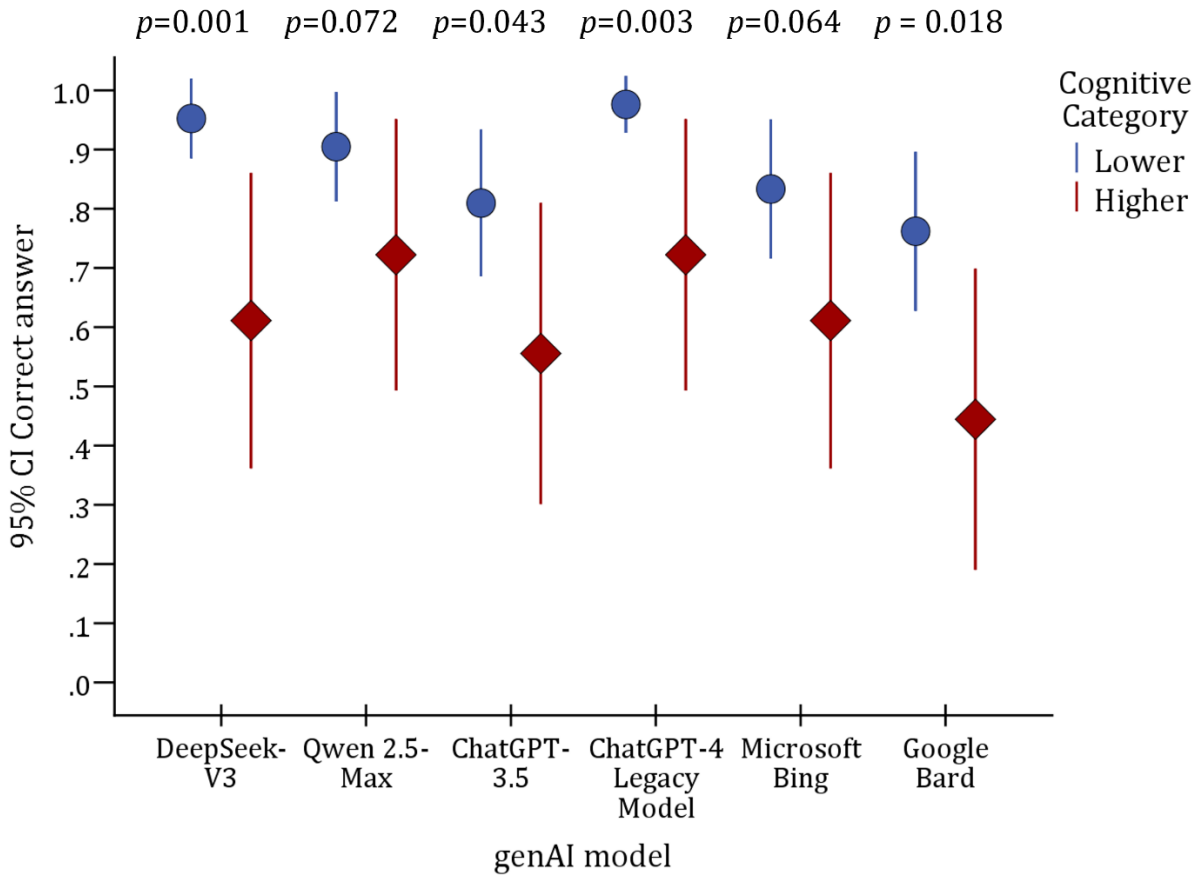

Fig. 1. Performance of generative AI (genAI) models across cognitive categories based on Bloom's taxonomy. Mean accuracy and 95% confidence intervals (CI) are shown for lower-order cognitive tasks (Remember, Understand; blue circles) and higher-order cognitive tasks (Apply, Analyze; red diamonds).

Bonferroni post hoc comparisons revealed that ChatGPT-4 Legacy Model achieved significantly higher accuracy than Google Bard in lower-order cognitive tasks ($p = 0.045$), while no other significant differences were observed among genAI models in this category. Pairwise comparisons between DeepSeek-V3, Qwen 2.5-Max, ChatGPT-3.5, and Microsoft Bing demonstrated no statistically significant differences ($p > 0.05$). In higher-order cognitive tasks, no genAI model significantly outperformed another, with all pairwise comparisons yielding non-significant results ($p > 0.05$).

3.3 Benchmarking genAI performance based on MCQ complexity

Pearson correlation analysis revealed no significant relationship between MCQ stem word count and AI model accuracy across all tested genAI models ($p > 0.05$). DeepSeek-V3 (r = 0.011, $p = 0.932$), Qwen 2.5-Max (r = 0.119, $p = 0.366$), ChatGPT-3.5 (r = –0.064, $p = 0.627$), ChatGPT-4 Legacy Model (r = –0.055, $p = 0.678$), Microsoft Bing (r = 0.074, $p = 0.577$), and Google Bard (r = 0.111, $p = 0.397$) all demonstrated weak, non-significant correlations with MCQ stem word count, indicating that question length did not influence genAI model accuracy in this dataset.

Pearson correlation analysis demonstrated a significant negative correlation between MCQ choice word count and accuracy for ChatGPT-3.5 (r = –0.263, $p = 0.042$) and Google Bard (r = –0.269, $p = 0.038$), indicating that longer answer choices were associated with lower accuracy for these models. No significant correlations were observed for DeepSeek-V3 (r = 0.015, $p = 0.909$), Qwen 2.5-Max (r = –0.049, $p = 0.711$), ChatGPT-4 Legacy Model (r = –0.099, $p = 0.454$), or Microsoft Bing (r = –0.152, $p = 0.245$), suggesting that MCQ choice length did not significantly impact performance for these models.

3.4 Decision tree analysis on genAI performance across cognitive levels and model types

Decision tree analysis using the CHAID method identified cognitive category (Lower vs. Higher) as the most significant predictor of genAI accuracy ($p < 0.001$, F = 34.634). GenAI models exhibited a higher mean accuracy in lower-order tasks (Remember, Understand) at 87.3%, compared to 61.1% in higher-order tasks (Apply, Analyze). Among lower-order tasks, DeepSeek-V3, Qwen 2.5-Max, and ChatGPT-4 Legacy Model achieved a significantly higher accuracy (94.4%) than ChatGPT-3.5, Microsoft Bing, and Google Bard, which scored 80.0% ($p = 0.019$, F = 12.061) (**Fig. 2**). The overall risk

estimate of the model was 0.145 ± 0.011, indicating a low classification error rate. These findings confirm that Chinese genAI models (DeepSeek-V3, Qwen 2.5-Max) perform comparably to ChatGPT-4 Legacy Model in lower-order cognitive tasks, while Western AI models exhibit lower accuracy overall.

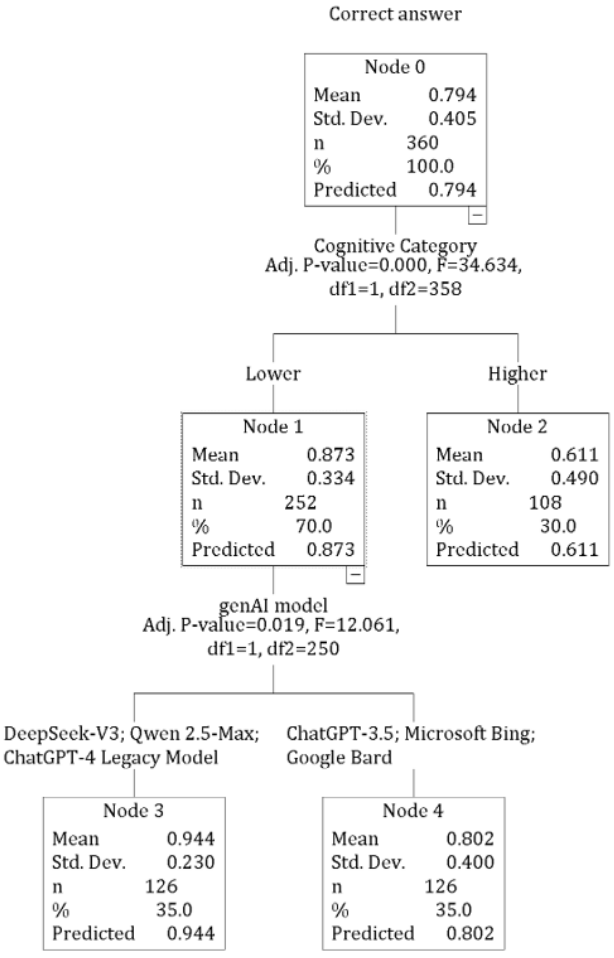

Fig. 2. Decision tree analysis illustrating the impact of cognitive category and generative AI (genAI) model type on accuracy in Clinical Chemistry multiple-choice questions (MCQs).

4. DISCUSSION

The remarkable performance of Chinese genAI models marks a new milestone in the trajectory of genAI abilities. In the highly specialized and cognitively demanding field of Clinical Chemistry assessed in this study, DeepSeek-V3 and Qwen 2.5-Max not only rivaled but in several aspects equaled or exceeded their Western counterparts, a feat that would have been inconceivable just a few months ago [63]. The main study results force a reckoning with the unprecedented cognitive capabilities of Chinese genAI models and their potential to contribute in reshaping the very foundations of healthcare education. If a machine can now outperform human postgraduate students in an advanced discipline requiring precision, pattern recognition, and applied problem-solving, then the implications extend beyond pedagogy—they call into question the assumed superiority of human cognition in structured medical assessments.

The remarkable performance of Chinese genAI models in this study marks a significant shift in the landscape of AI-driven healthcare education. Previous studies have extensively evaluated Western genAI models, particularly OpenAI's ChatGPT series, across diverse medical disciplines [25, 35, 42, 64-73]. This evaluation included the United States Medical Licensing Examination (USMLE) exam [25, 74]. However, the inclusion of DeepSeek-V3 and Qwen 2.5-Max in this study offers an

unprecedented benchmark of non-Western genAI models in a specialized and cognitively demanding field like Clinical Chemistry.

In an early scoping review on ChatGPT performance in MCQs by Newton and Xiromeriti, ChatGPT-4 vastly outperformed its predecessor ChatGPT-3, achieving a passing rate of 93% compared to 20%, with superior performance over human students in 35% of the assessed exams [75]. While earlier reports have highlighted the evolution of OpenAI's models, our study is among the first to rigorously benchmark Chinese genAI models in healthcare education, confirming that their claimed benchmarks hold up under independent evaluation [26, 57]. Collectively, our findings compared to similar literature assessing genAI performance in healthcare education underscore a new reality: the dominance of Western genAI is no longer assured. DeepSeek-V3 and Qwen 2.5-Max have proven their capacity to match, if not surpass, their Western genAI rivals in structured knowledge recall and cognitive processing. This not only affirms China's growing AI prowess but also raises critical questions about the future trajectory of genAI development, access, and integration into healthcare education globally [76, 77].

The findings of this study were not just remarkable—they can be described as unsettling. DeepSeek-V3, Qwen 2.5-Max, and ChatGPT-4 did not merely perform well; they decisively outclassed human postgraduate students in Clinical Chemistry, a field that demands precision and analytical depth. That a genAI model—devoid of experience, intuition, or human judgment—can surpass individuals who have dedicated extensive time and efforts to mastering this discipline is not merely an achievement of engineering; it is a stark warning. This is no longer a question of AI assisting health education—it is outperforming the very individuals it was designed to support. With models like ChatGPT-4, Qwen 2.5-Max, and DeepSeek-V3 achieving success rates exceeding 85%, compared to the human average of 67%, we must confront deeply uncomfortable questions: what happens when genAI no longer complements human expertise but surpasses it? If memorization, pattern recognition, and structured reasoning can be executed more efficiently by machines, then the very pedagogical foundations of health education—lectures, standardized exams, rote learning—may be rendered obsolete [78, 79].

The implications of remarkable performance of Western and Chinese genAI models are unsettling. What role remains for the human physician in a world where genAI can diagnose faster, recall more accurately, and analyze data with unparalleled efficiency? Does this signal a necessary shift toward an education system that prioritizes judgment, ethics, and human intuition—qualities genAI cannot yet replicate? Or does it mark the beginning of an era where the traditional notion of human expertise is increasingly diminished, replaced by an ever-advancing algorithm that never tires, never forgets, and never needs training beyond its initial parameters? If these findings are not met with urgent reflection, then the field of healthcare risks entering an age where the physician is no longer the ultimate authority, but a consultant to an omnipresent, infallible machine [80-82]. What we have witnessed in this study is not merely progress—it is a profound restructuring of the intellectual hierarchy that has governed medicine for centuries [83].

A fundamental and deeply unsettling reality was revealed by the results of this study: healthcare education assessments, as currently structured, are not evaluating intelligence—they are predominantly testing recall. The unyielding dominance of genAI models in lower-order cognitive tasks, as dictated by the revised Bloom's taxonomy, confirms what should have been clear long ago: MCQs are an anachronism, an outdated relic of a system that has long mistaken memorization for mastery [52, 84]. In this study, genAI models such as DeepSeek-V3, Qwen 2.5-Max, and ChatGPT-4 demonstrated near-perfect proficiency in Remember and Understand tasks, achieving 87.3% accuracy in lower-order cognitive domains. Yet when challenged with higher-order tasks (Apply, Analyze), their accuracy dropped significantly. This discrepancy should not reassure educators—it should alarm them. The reality is that a well-trained machine can now outscore human students on the very exams that determine medical competence, yet it remains less capable of critical reasoning, real-time decision-making, or patient-centered judgment. If genAI models can effortlessly excel in recall-based MCQs, what, then, is the value of these assessments? This study does not merely confirm genAI's strengths—it highlights the urgent inadequacy of current medical education assessments [84-87]. If institutions continue to prioritize rote learning over problem-solving, they will produce graduates who pass exams but lack the skills to navigate the uncertainties of real-world clinical practice [43, 88].

Based on the study results and consistent with a wide range of studies, it appears that genAI has exposed the intellectual hollowness of traditional assessment methods; it is now incumbent upon health educators to correct course [25, 35, 42, 64-73, 87]. The path forward is clear: medical education must abandon its overreliance on knowledge retrieval and instead emphasize clinical reasoning, synthesis, and adaptive decision-making—domains where genAI might still falter [89]. If assessments in healthcare education remain unchanged, we risk training a generation of health professionals whose primary skill is competing with a machine that will always be faster, more precise, and more reliable in pure factual recall [90, 91]. The objective of healthcare education must not be to replicate genAI's strengths but to cultivate uniquely human capabilities—intuition, ethical discernment, and the ability to think critically amid uncertainty [92]. To ignore these findings is to surrender the future of medical expertise to algorithms. The imperative is clear: adapt the curriculum, evolve the assessments, or risk irrelevance in the age of AI [12, 93, 94].

Despite their dominance, not all genAI models in this study withstood the subtleties of complex MCQs. ChatGPT-3.5 and Google Bard faltered when faced with longer, more nuanced answer choices, showing a negative correlation between choice

length and accuracy. This suggests that even sophisticated genAI remains vulnerable to linguistic complexity, distractors, and phrasing subtleties—an Achilles' heel that health educators could exploit. By designing MCQs with heightened complexity, assessment frameworks could shift from testing recall to truly challenging cognitive depth, ensuring that AI-driven pattern recognition does not replace genuine analytical reasoning. This presents a necessary safeguard against uncritical AI reliance in health education [95, 96].

An emerging hierarchy of intelligence among genAI models was evident in this study, with DeepSeek-V3, Qwen 2.5-Max, and ChatGPT-4 consistently outperforming their counterparts, ChatGPT-3.5, Microsoft Bing, and Google Bard. This stratification signifies more than a mere performance gap—it represents a fundamental shift in digital competency, where select genAI models achieve near-expert proficiency, while others remain constrained in both factual recall and analytical reasoning. Just as the choice of authoritative medical texts once shaped health education, the selection of genAI models would now directly influence a student's comprehension, accuracy, and cognitive development. The unregulated use of lower-performing genAI models introduces a risk—students engaging with suboptimal AI tools may develop an incomplete or distorted understanding of complex medical concepts [97, 98]. To preserve academic integrity and ensure consistent, high-quality AI integration, institutions must consider standardized benchmarks, guiding the incorporation of only the most reliable AI models into medical training [31]. As genAI becomes increasingly interwoven with learning, the quality of knowledge acquisition is now directly tied to the caliber of the AI model employed, necessitating a strategic and evidence-based approach to its adoption in healthcare education [99, 100].

Finally, this study has several limitations that warrant consideration. First, the assessment was limited to MCQs, favoring structured recall over clinical reasoning. GenAI models excel in such formats, but their real-world decision-making capabilities remain uncertain. Future research should incorporate case-based discussions and problem-solving assessments for a broader evaluation. Second, the study focused solely on Clinical Chemistry MCQs, limiting generalizability to other healthcare disciplines. Third, the benchmarking did not account for genAI training datasets or knowledge cutoffs, which may influence performance. The extent to which proprietary datasets contribute to genAI success remains unclear and warrants further study. Fourth, the reasoning behind genAI-generated responses was not analyzed in this study as opposed to the first study [42], leaving potential hallucinations, biases, or AI-generated content errors unexamined. Finally,, all MCQs were in English, limiting insights into genAI performance in non-English languages since evaluating genAI's linguistic robustness is essential for global applicability.

5. CONCLUSIONS

The rise of Chinese genAI with remarkable performance as shown in this study is nothing short of extraordinary. DeepSeek-V3 and Qwen 2.5-Max have not merely challenged but, in critical domains, matched the formidable ChatGPT-4 model—a feat that would have been inconceivable just months ago when Western genAI reigned unchallenged. The results of this study mark the dismantling of Silicon Valley's once-unquestioned monopoly, as China's genAI infrastructure surges forward, demanding a recalibration of global expectations. No longer a follower, China has emerged as a formidable contender, proving that genAI innovation is no longer the sole province of San Francisco or Seattle but may soon be led by Beijing and Hangzhou. Beyond statistical superiority, the performance of DeepSeek-V3 and Qwen 2.5-Max signals a paradigm shift. These genAI models demonstrated exceptional accuracy in lower-order cognitive tasks, rivaling their Western counterparts while surpassing human postgraduate students in knowledge recall and structured analysis. If genAI can now outperform human learners in factual domains, then the very structure of healthcare education must evolve accordingly. The assumption that genAI will remain an adjunct to human expertise is rapidly becoming obsolete—these models are no longer tools; they are competitors. Yet, the results also expose the persistent limitations of all genAI models. While these models excel in recall and structured reasoning, they falter in higher-order cognitive tasks requiring judgment, synthesis, and clinical intuition—domains still firmly within human expertise. This gap presents both an imperative and an opportunity: genAI developers must refine these models to strengthen critical reasoning, while health educators must adapt curricula to emphasize uniquely human faculties—intuition, ethical judgment, and problem-solving in unpredictable clinical scenarios. Ultimately, this study is a testament to China's genAI mastery. The days of unquestioned genAI leadership in the West are over, and the future of AI-driven healthcare education will be dictated not by geography but by innovation. Those who fail to recognize this seismic shift, who assume that genAI excellence will remain a Western prerogative, do so at their peril.

Abbreviations

AI: Artificial intelligence
CHAID: Chi-squared Automatic Interaction Detection
CI: Confidence interval
FI: Facility index
genAI: Generative artificial intelligence

LLM: Large language models
MCQ: Multiple-choice question
METRICS: Model, Evaluation, Timing, Range/Randomization, Individual factors, Count, and Specificity of prompts and language
NLP: Natural language processing
USMLE: The United States Medical Licensing Examination

Conflicts Of Interest

The authors declare no conflicts of interest.

Funding

The authors clearly indicate that the research was conducted without any funding from external sources.

Acknowledgment

The first draft of the manuscript was written by the author, Malik Sallam. No external writing assistance was provided. OpenAI's ChatGPT-4o model was used to refine the language and structure of the manuscript. All final edits and intellectual contributions were made solely by the authors, and no external funding or payment was involved in the writing process.

References

[1] G. Schubring, "Textbooks Before the Invention of the Printing Press: Orality and Teaching," in *Analysing Historical Mathematics Textbooks*, G. Schubring Ed. Cham: Springer International Publishing, 2022, pp. 15-53.

[2] M. Karabacak, B. B. Ozkara, K. Margetis, M. Wintermark, and S. Bisdas, "The Advent of Generative Language Models in Medical Education," (in eng), *JMIR Med Educ*, vol. 9, p. e48163, Jun 6 2023, doi: 10.2196/48163.

[3] M. Sallam, N. A. Salim, M. Barakat, and A. B. Al-Tammemi, "ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations," (in eng), *Narra J*, vol. 3, no. 1, p. e103, Apr 2023, doi: 10.52225/narra.v3i1.103.

[4] M. Mijwil, M. Abotaleb, A. L. I. Guma, and K. Dhoska, "Assigning Medical Professionals: ChatGPT's Contributions to Medical Education and Health Prediction," *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol. 2024, pp. 76-83, 07/20 2024, doi: 10.58496/MJAIH/2024/011.

[5] M. S. Nilsson, S. Pennbrant, E. Pilhammar, and C.-G. Wenestam, "Pedagogical strategies used in clinical medical education: an observational study," *BMC Medical Education*, vol. 10, no. 1, p. 9, 2010/01/28 2010, doi: 10.1186/1472-6920-10-9.

[6] M. M. Mir *et al.*, "Application of Artificial Intelligence in Medical Education: Current Scenario and Future Perspectives," (in eng), *J Adv Med Educ Prof*, vol. 11, no. 3, pp. 133-140, Jul 2023, doi: 10.30476/jamp.2023.98655.1803.

[7] N. Mansuri and A. B. Zelenski, "Flourishing as an Aim of Medical Education: Are We Hitting the Target?," *Medical Science Educator*, 2025/01/29 2025, doi: 10.1007/s40670-024-02255-x.

[8] M. Vanstone and L. Grierson, "Thinking about social power and hierarchy in medical education," (in eng), *Med Educ*, vol. 56, no. 1, pp. 91-97, Jan 2022, doi: 10.1111/medu.14659.

[9] K. M. Ludmerer, "The development of American medical education from the turn of the century to the era of managed care," (in eng), *Clin Orthop Relat Res*, no. 422, pp. 256-62, May 2004, doi: 10.1097/01.blo.0000131257.59585.b0.

[10] S. Narayanan, R. Ramakrishnan, E. Durairaj, and A. Das, "Artificial Intelligence Revolutionizing the Field of Medical Education," (in eng), *Cureus*, vol. 15, no. 11, p. e49604, Nov 2023, doi: 10.7759/cureus.49604.

[11] A. Alhur, "Redefining Healthcare With Artificial Intelligence (AI): The Contributions of ChatGPT, Gemini, and Co-pilot," (in eng), *Cureus*, vol. 16, no. 4, p. e57795, Apr 2024, doi: 10.7759/cureus.57795.

[12] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," (in eng), *Healthcare (Basel)*, vol. 11, no. 6, p. 887, Mar 19 2023, doi: 10.3390/healthcare11060887.

[13] C. Hervás-Gómez, M. D. Díaz Noguera, and F. Vera, *The Education Revolution through Artificial Intelligence Enhancing Skills, Safeguarding Rights, and Facilitating Human-Machine Collaboration*. 2024.

[14] S. Hacking, "ChatGPT and Medicine: Together We Embrace the AI Renaissance," (in eng), *JMIR Bioinform Biotechnol*, vol. 5, p. e52700, May 7 2024, doi: 10.2196/52700.

[15] G. Rubeis, K. Dubbala, and I. Metzler, ""Democratizing" artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term," (in eng), *Front Genet*, vol. 13, p. 902542, 2022, doi: 10.3389/fgene.2022.902542.

[16] C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review," *Smart Learning Environments*, vol. 11, no. 1, p. 28, 2024/06/18 2024, doi: 10.1186/s40561-024-00316-7.
[17] Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," *Humanities and Social Sciences Communications*, vol. 10, no. 1, p. 567, 2023/09/13 2023, doi: 10.1057/s41599-023-02079-x.
[18] S. M. Williamson and V. Prybutok, "The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation," *Information*, vol. 15, no. 6, p. 299, 2024, doi: 10.3390/info15060299.
[19] M. M. Triola and A. Rodman, "Integrating Generative Artificial Intelligence Into Medical Education: Curriculum, Policy, and Governance Strategies," (in eng), *Acad Med*, Dec 20 2024, doi: 10.1097/acm.0000000000005963.
[20] V. Rampton, M. Mittelman, and J. Goldhahn, "Implications of artificial intelligence for medical education," (in eng), *Lancet Digit Health*, vol. 2, no. 3, pp. e111-e112, Mar 2020, doi: 10.1016/s2589-7500(20)30023-6.
[21] C. Mulligan and P. Godsiff, "Datalism and Data Monopolies in the Era of A.I.: A Research Agenda," *arXiv*, 2023, doi: 10.48550/arXiv.2307.08049.
[22] C. Durand, *How Silicon Valley Unleashed Techno-Feudalism: The Making of the Digital Economy*. Verso Books, 2024.
[23] K.-F. Lee, *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin, 2018.
[24] O. Strozhemin. "Google, Microsoft, OpenAI Square Up on Generative AI." Informa TechTarget. https://aibusiness.com/companies/google-microsoft-openai-square-up-on-generative-ai#close-modal (accessed 2025).
[25] M. Sallam, "Bibliometric top ten healthcare-related ChatGPT publications in the first ChatGPT anniversary," (in eng), *Narra J*, vol. 4, no. 2, p. e917, Aug 2024, doi: 10.52225/narra.v4i2.917.
[26] M. Sallam, K. Al-Mahzoum, M. Sallam, and M. M. Mijwil, "DeepSeek: Is it the End of Generative AI Monopoly or the Mark of the Impending Doomsday?," *Mesopotamian Journal of Big Data*, vol. 2025, pp. 26-34, 2025, doi: 10.58496/MJBD/2025/002.
[27] E. Gibney, "China's cheap, open AI model DeepSeek thrills scientists," (in eng), *Nature*, vol. 638, no. 8049, pp. 13-14, Feb 2025, doi: 10.1038/d41586-025-00229-6.
[28] J. Smith, "Daily briefing: The pros and cons of DeepSeek," (in eng), *Nature*, Jan 30 2025, doi: 10.1038/d41586-025-00330-w.
[29] D. Normile, "Chinese firm's large language model makes a splash," (in eng), *Science*, vol. 387, no. 6731, p. 238, Jan 17 2025, doi: 10.1126/science.adv9836.
[30] J. Luo, "A critical review of GenAI policies in higher education assessment: a call to reconsider the "originality" of students' work," *Assessment & Evaluation in Higher Education*, vol. 49, no. 5, pp. 651-664, 2024/07/03 2024, doi: 10.1080/02602938.2024.2309963.
[31] M. Sallam, R. Khalil, and M. Sallam, "Benchmarking Generative AI: A Call for Establishing a Comprehensive Framework and a Generative AIQ Test," *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol. 2024, pp. 69-75, 2024, doi: 10.58496/MJAIH/2024/010.
[32] M. F. Klein, "Use of Taxonomy of Educational Objectives (Cognitive Domain) in Constructing Tests for Primary School Pupils," *The Journal of Experimental Education*, vol. 40, no. 3, pp. 38-50, 1972. [Online]. Available: http://www.jstor.org/stable/20157277.
[33] A. Herrmann-Werner *et al.*, "Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study," (in eng), *J Med Internet Res*, vol. 26, p. e52113, Jan 23 2024, doi: 10.2196/52113.
[34] F. Eisinger *et al.*, "What's Going On With Me and How Can I Better Manage My Health? The Potential of GPT-4 to Transform Discharge Letters Into Patient-Centered Letters to Enhance Patient Safety: Prospective, Exploratory Study," (in eng), *J Med Internet Res*, vol. 27, p. e67143, Jan 21 2025, doi: 10.2196/67143.
[35] M. Sallam and K. Al-Salahat, "Below average ChatGPT performance in medical microbiology exam compared to university students," *Frontiers in Education*, Original Research vol. 8, p. 1333415, 2023, doi: 10.3389/feduc.2023.1333415.
[36] N. E. Adams, "Bloom's taxonomy of cognitive learning objectives," (in eng), *J Med Libr Assoc*, vol. 103, no. 3, pp. 152-3, Jul 2015, doi: 10.3163/1536-5050.103.3.010.
[37] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview," *Theory Into Practice*, vol. 41, no. 4, pp. 212-218, 2002. [Online]. Available: http://www.jstor.org/stable/1477405.
[38] E. J. Furst, "Bloom's Taxonomy of Educational Objectives for the Cognitive Domain: Philosophical and Educational Issues," *Review of Educational Research*, vol. 51, no. 4, pp. 441-453, 1981, doi: 10.2307/1170361.

[39] M. J. Gierl, "Comparing Cognitive Representations of Test Developers and Students on a Mathematics Test with Bloom's Taxonomy," *The Journal of Educational Research*, vol. 91, no. 1, pp. 26-32, 1997. [Online]. Available: http://www.jstor.org/stable/27542125.

[40] M. J. Benson, M. J. Sporakowski, and A. J. Stremmel, "Writing Reviews of Family Literature: Guiding Students Using Bloom's Taxonomy of Cognitive Objectives," *Family Relations*, vol. 41, no. 1, pp. 65-69, 1992, doi: 10.2307/585395.

[41] G. E. Gignac and E. T. Szodorai, "Defining intelligence: Bridging the gap between human and artificial perspectives," *Intelligence*, vol. 104, p. 101832, 2024/05/01/ 2024, doi: 10.1016/j.intell.2024.101832.

[42] M. Sallam, K. Al-Salahat, H. Eid, J. Egger, and B. Puladi, "Human versus Artificial Intelligence: ChatGPT-4 Outperforming Bing, Bard, ChatGPT-3.5 and Humans in Clinical Chemistry Multiple-Choice Questions," (in eng), *Adv Med Educ Pract*, vol. 15, pp. 857-871, 2024, doi: 10.2147/amep.S479801.

[43] P. Benner, R. G. Hughes, and M. Sutphen, "Advances in Patient Safety Clinical Reasoning, Decisionmaking, and Action: Thinking Critically and Clinically," in *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, R. G. Hughes Ed. Rockville (MD): Agency for Healthcare Research and Quality (US), 2008.

[44] H. J. Einhorn and R. M. Hogarth, "Behavioral Decision Theory: Processes of Judgment and Choice," *Journal of Accounting Research*, vol. 19, no. 1, pp. 1-31, 1981, doi: 10.2307/2490959.

[45] I. Nonaka, "A Dynamic Theory of Organizational Knowledge Creation," *Organization Science*, vol. 5, no. 1, pp. 14-37, 1994. [Online]. Available: http://www.jstor.org/stable/2635068.

[46] J. Haase and P. H. P. Hanel, "Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity," *Journal of Creativity*, vol. 33, no. 3, p. 100066, 2023/12/01/ 2023, doi: 10.1016/j.yjoc.2023.100066.

[47] M. B. Garcia, "The Paradox of Artificial Creativity: Challenges and Opportunities of Generative AI Artistry," *Creativity Research Journal*, pp. 1-14, 2024, doi: 10.1080/10400419.2024.2354622.

[48] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera, "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Information Fusion*, vol. 99, p. 101896, 2023/11/01/ 2023, doi: 10.1016/j.inffus.2023.101896.

[49] J. Yuzwa, "Speaking of the Written Word: Socrates' Critique of Writing in Plato's Phaedrus," *Crossings*, no. 3, pp. 121-138, 2019.

[50] J. M. Burkhardt, "History of fake news," *Library Technology Reports*, vol. 53, no. 8, pp. 5-9, 2017.

[51] A. Ateeq, M. Alzoraiki, M. Milhem, and R. Ateeq, "Artificial intelligence in education: implications for academic integrity and the shift toward holistic assessment," *Frontiers in Education*, vol. 9, p. 1470979, 10/01 2024, doi: 10.3389/feduc.2024.1470979.

[52] Q. Liu, N. Wald, C. Daskon, and T. Harland, "Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers," *Innovations in Education and Teaching International*, vol. 61, no. 4, pp. 802-814, 2024/07/03 2024, doi: 10.1080/14703297.2023.2222715.

[53] M. Sallam, M. Barakat, and M. Sallam, "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," (in eng), *Interact J Med Res*, vol. 13, p. e54704, Feb 15 2024, doi: 10.2196/54704.

[54] OpenAI. "GPT-3.5." https://openai.com/ (accessed 27 November 2023, 2023).

[55] Microsoft and OpenAI. "Bing is your AI-powered copilot for the web." https://www.bing.com/search?q=Bing+AI&showconv=1&FORM=hpcodx (accessed 27 November 2023, 2023).

[56] Google. "Bard." https://bard.google.com/chat (accessed 27 November 2023, 2023).

[57] A. Liu *et al.*, "Deepseek-v3 technical report," *arXiv*, 2024, doi: 10.48550/arXiv.2412.19437.

[58] DeepSeek. "DeepSeek." https://www.deepseek.com/ (accessed 2025, 2025).

[59] Alibaba Cloud. "Qwen 2.5-Max: A large language model for advanced reasoning and domain-specific applications." https://chat.qwenlm.ai/ (accessed 2025, 2025).

[60] C. A. Burtis, E. R. Ashwood, D. E. Bruns, and N. W. Tietz, *Tietz textbook of clinical chemistry and molecular diagnostics*, 5th ed. St. Louis, Mo.: Saunders, 2013, pp. xviii, 2,238 p.

[61] M. L. Bishop, E. P. Fody, and L. E. Schoeff, *Clinical chemistry : principles, techniques, and correlations*, Eighth edition. ed. Philadelphia: Wolters Kluwer, 2018, pp. xxviii, 736 pages.

[62] R. A. McPherson and M. R. Pincus, *Henry's clinical diagnosis and management by laboratory methods*, 24. ed. Philadelphia: Elsevier, 2021, p. pages cm.

[63] G. Lobo. "ChatGPT in the Public Sector – overhyped or overlooked?" https://interoperable-europe.ec.europa.eu/collection/public-sector-tech-watch/news/chatgpt-public-sector-overhyped-or-overlooked (accessed 2025, 2025).

[64] S. Sawamura, K. Kohiyama, T. Takenaka, T. Sera, T. Inoue, and T. Nagai, "An Evaluation of the Performance of OpenAI-o1 and GPT-4o in the Japanese National Examination for Physical Therapists," (in eng), *Cureus*, vol. 17, no. 1, p. e76989, Jan 2025, doi: 10.7759/cureus.76989.

[65] K. Ishida and E. Hanada, "ChatGPT (GPT-4V) Performance on the Healthcare Information Technologist Examination in Japan," (in eng), *Cureus*, vol. 17, no. 1, p. e76775, Jan 2025, doi: 10.7759/cureus.76775.

[66] Y. Chen *et al.*, "Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study," *BMC Medical Education*, vol. 24, no. 1, p. 1372, 2024/11/26 2024, doi: 10.1186/s12909-024-06309-x.

[67] Y. Chang, C. Y. Su, and Y. C. Liu, "Assessing the Performance of Chatbots on the Taiwan Psychiatry Licensing Examination Using the Rasch Model," (in eng), *Healthcare (Basel)*, vol. 12, no. 22, p. 2305, Nov 18 2024, doi: 10.3390/healthcare12222305.

[68] S. Zare, S. Vafaeian, M. Amini, K. Farhadi, M. Vali, and A. Golestani, "Comparing the performance of ChatGPT-3.5-Turbo, ChatGPT-4, and Google Bard with Iranian students in pre-internship comprehensive exams," *Scientific Reports*, vol. 14, no. 1, p. 28456, 2024/11/18 2024, doi: 10.1038/s41598-024-79335-w.

[69] J. Huwiler, L. Oechslin, P. Biaggi, F. C. Tanner, and C. A. Wyss, "Experimental assessment of the performance of artificial intelligence in solving multiple-choice board exams in cardiology," *Swiss Medical Weekly*, vol. 154, no. 10, p. 3547, 10/02 2024, doi: 10.57187/s.3547.

[70] M. Sallam *et al.*, "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*, vol. 17, no. 1, p. 247, 2024/09/03 2024, doi: 10.1186/s13104-024-06920-7.

[71] M. Rodrigues Alessi, H. A. Gomes, M. Lopes de Castro, and C. Terumy Okamoto, "Performance of ChatGPT in Solving Questions From the Progress Test (Brazilian National Medical Exam): A Potential Artificial Intelligence Tool in Medical Practice," (in eng), *Cureus*, vol. 16, no. 7, p. e64924, Jul 2024, doi: 10.7759/cureus.64924.

[72] C.-Y. Tsai, S.-J. Hsieh, H.-H. Huang, J.-H. Deng, Y.-Y. Huang, and P.-Y. Cheng, "Performance of ChatGPT on the Taiwan urology board examination: insights into current strengths and shortcomings," *World Journal of Urology*, vol. 42, no. 1, p. 250, 2024/04/23 2024, doi: 10.1007/s00345-024-04957-8.

[73] L. Morjaria *et al.*, "Examining the Efficacy of ChatGPT in Marking Short-Answer Assessments in an Undergraduate Medical Program," *International Medical Education*, vol. 3, no. 1, pp. 32-43, 2024, doi: 10.3390/ime3010004.

[74] T. H. Kung *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, p. e0000198, 2023, doi: 10.1371/journal.pdig.0000198.

[75] P. Newton and M. Xiromeriti, "ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review," *Assessment & Evaluation in Higher Education*, vol. 49, no. 6, pp. 781-798, 2024/08/17 2024, doi: 10.1080/02602938.2023.2299059.

[76] S. Khanal, H. Zhang, and A. Taeihagh, "Development of New Generation of Artificial Intelligence in China: When Beijing's Global Ambitions Meet Local Realities," *Journal of Contemporary China*, vol. 34, no. 151, pp. 19-42, 2025/01/02 2025, doi: 10.1080/10670564.2024.2333492.

[77] S. Reddy, "Generative AI in healthcare: an implementation science informed translational path on application, integration and governance," (in eng), *Implement Sci*, vol. 19, no. 1, p. 27, Mar 15 2024, doi: 10.1186/s13012-024-01357-9.

[78] I. Kavathatzopoulos, "Artificial intelligence and the sustainability of thinking: How AI may destroy us, or help us," in *Ethics and Sustainability in Digital Cultures*: Routledge, 2024, pp. 19-30.

[79] S. Wartman and C. Combs, "Medical Education Must Move From the Information Age to the Age of Artificial Intelligence," *Academic Medicine*, vol. 93, p. 1, 11/01 2017, doi: 10.1097/ACM.0000000000002044.

[80] E. M. Hille, P. Hummel, and M. Braun, "Meaningful Human Control over AI for Health? A Review," (in eng), *J Med Ethics*, Sep 20 2023, doi: 10.1136/jme-2023-109095.

[81] F. Funer and U. Wiesing, "Physician's autonomy in the face of AI support: walking the ethical tightrope," (in eng), *Front Med (Lausanne)*, vol. 11, p. 1324963, 2024, doi: 10.3389/fmed.2024.1324963.

[82] H. Kempt and S. K. Nagel, "Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts," (in eng), *J Med Ethics*, vol. 48, no. 4, pp. 222-229, Apr 2022, doi: 10.1136/medethics-2021-107440.

[83] M. Mijwil, K. Hiran, R. Doshi, M. Dadhich, A.-H. Al-Mistarehi, and I. Bala, "ChatGPT and the Future of Academic Integrity in the Artificial Intelligence Era: A New Frontier," *Al-Salam Journal for Engineering and Technology,* vol. 2, pp. 116-127, 04/13 2023, doi: 10.55145/ajest.2023.02.02.015.

[84] A. H. Sam, S. Hameed, J. Harris, and K. Meeran, "Validity of very short answer versus single best answer questions for undergraduate assessment," (in eng), *BMC Med Educ,* vol. 16, no. 1, p. 266, Oct 13 2016, doi: 10.1186/s12909-016-0793-z.

[85] P. Parekh and V. Bahadoor, "The Utility of Multiple-Choice Assessment in Current Medical Education: A Critical Review," (in eng), *Cureus,* vol. 16, no. 5, p. e59778, May 2024, doi: 10.7759/cureus.59778.

[86] M. Iñarrairaegui *et al.*, "Evaluation of the quality of multiple-choice questions according to the students' academic level," *BMC Medical Education,* vol. 22, no. 1, p. 779, 2022/11/11 2022, doi: 10.1186/s12909-022-03844-3.

[87] A. B. Mbakwe, I. Lourentzou, L. A. Celi, O. J. Mechanic, and A. Dagan, "ChatGPT passing USMLE shines a spotlight on the flaws of medical education," *PLOS Digital Health,* vol. 2, no. 2, p. e0000205, 2023, doi: 10.1371/journal.pdig.0000205.

[88] S. French, A. Dickerson, and R. A. Mulder, "A review of the benefits and drawbacks of high-stakes final examinations in higher education," *Higher Education,* vol. 88, no. 3, pp. 893-918, 2024/09/01 2024, doi: 10.1007/s10734-023-01148-z.

[89] B. Charlin, J. Tardif, and H. P. Boshuizen, "Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research," (in eng), *Acad Med,* vol. 75, no. 2, pp. 182-90, Feb 2000, doi: 10.1097/00001888-200002000-00020.

[90] S. A. Tabish, "Assessment methods in medical education," (in eng), *Int J Health Sci (Qassim),* vol. 2, no. 2, pp. 3-7, Jul 2008.

[91] Organisation for Economic Co-operation and Development (OECD). "Artificial Intelligence and the health workforce." https://www.oecd.org/en/publications/artificial-intelligence-and-the-health-workforce_9a31d8af-en.html (accessed 2025, 2025).

[92] S. Koos and S. Wachsmann, "Navigating the Impact of ChatGPT/GPT4 on Legal Academic Examinations: Challenges, Opportunities and Recommendations," *Media Iuris,* vol. 6, no. 2, pp. 255-270, 06/20 2023, doi: 10.20473/mi.v6i2.45270.

[93] L. Sun, C. Yin, Q. Xu, and W. Zhao, "Artificial intelligence for healthcare and medical education: a systematic review," (in eng), *Am J Transl Res,* vol. 15, no. 7, pp. 4820-4828, 2023.

[94] M. Mijwil, A. L. I. Guma, and E. Sadıkoğlu, "The Evolving Role of Artificial Intelligence in the Future of Distance Learning: Exploring the Next Frontier," *Mesopotamian Journal of Computer Science,* vol. 2023, pp. 98-105, 05/21 2023, doi: 10.58496/MJCSC/2023/012.

[95] P. M. Newton *et al.*, "Can ChatGPT-4o Really Pass Medical Science Exams? A Pragmatic Analysis Using Novel Questions," *Medical Science Educator,* 2025/02/04 2025, doi: 10.1007/s40670-025-02293-z.

[96] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?," *Journal of applied learning and teaching,* vol. 6, no. 1, pp. 342-363, 2023, doi: 10.37074/jalt.2023.6.1.9.

[97] Y. Sun, D. Sheng, Z. Zhou, and Y. Wu, "AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content," *Humanities and Social Sciences Communications,* vol. 11, no. 1, p. 1278, 2024/09/27 2024, doi: 10.1057/s41599-024-03811-x.

[98] J. C. Vázquez Parra, C. González González, J. Amézquita, A. Cotino Arbelo, S. Palomino-Gámez, and M. Cruz-Sandoval, "Complex thinking and adopting artificial intelligence tools: a study of university students," *Frontiers in Education,* vol. 9, p. 1377553, 09/09 2024, doi: 10.3389/feduc.2024.1377553.

[99] M. Garcia *et al.*, "Effective Integration of Artificial Intelligence in Medical Education: Practical Tips and Actionable Insights," 2024, pp. 1-19.

[100] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," (in eng), *Future Healthc J,* vol. 8, no. 2, pp. e188-e194, Jul 2021, doi: 10.7861/fhj.2021-0095.