Original Research Article

Clinical Reasoning and Self-confidence among Preclinical Medical Students, Internal Medicine Specialists and Artificial Intelligence: A Cross-sectional Study.

.

ABSTRACT

|  |
| --- |
| **Aims:** Assessing diagnostic skills needs to test the diagnostic construction steps. This cross-sectional study compared the clinical reasoning accuracy and self-confidence of four groups of preclinical medical students from two Brazilian medical schools, internal medicine specialists, and large language models using the Clinical Reasoning and Self-confidence Assessment Tool.**Study design:** Cross-sectional study with the application of an assessment tool.**Place and Duration of Study:** Sample: Department of Medicine of Middle West State University of Paraná, Brazil, and Department of Medicine of Londrina State University of Paraná, Brazil, from March to November 2023.**Methodology:** Comparison of accuracy and self-confidence in 7 clinical cases, with specific questions. We included 133 preclinical medical students, 16 internal medicine specialists, and 2 large-language models of artificial intelligence. We used the student’s t- and the Kruskal‒Wallis’ tests. Spearman’s correlation was also applied. **Results:** Accuracy averages grew into students’ groups (beginners: 31.7±11.2% to second-year: 60.0±10.9%; *P* < .001). Specialists and large-language models performed similarly, with higher average accuracy than students (second-year: 60.0±10.9%, specialists: 75.7±10.0%, large-language models: 80.0%; *P* < .001). The median self-confidence was lower in the beginners than in the other groups (2.07 [1.71-2.89] vs. 3.14 [2.71-3.43]; *P* < .001). There is a correlation between accuracy and self-confidence (Rho = .623, r² = .374; *P* < .001).**Conclusion:** These results highlight the role of a dedicated tool and artificial intelligence for clinical reasoning assessment. |

*Keywords: medical education, clinical reasoning, clinical diagnosis, evaluation study, artificial intelligence.*

1. INTRODUCTION

Medical education presents numerous pedagogical challenges, encompassing adaptative and psychosocial issues (Huang, Newman & Schwartzstein, 2011; Tanaka et al., 2016; Torres, Sampaio & Caldeira, 2019; Figueiredo et al., 2022). Methodological and assessment concerns, especially in the first years of medical school, are historically present in this field (Mcconnell & Eva, 2012; Dunlosky et al., 2013; Hayat et al., 2020).

Clinical reasoning (CR) is a pivotal process that induces accurate diagnosis (Cate & Durning, 2017), thereby mitigating errors (Graber, Franklin & Gordon, 2005). One methodology to develop CR skills in beginner medical students is case-based learning (CBL), which uses illness scripts and analytic habits (Rencic, 2011; Cate, 2017; Si, 2022; Tureck, Souza & Faria, 2023). It employs a framework for each case study to organize information, summarize the case, generate hypotheses, justify the choices, and plan management (Daniel et al., 2019). Testing this framework is essential for CBL and CR assessments (Case, Swanson & Ripkey, 1994; Dory et al., 2016; Cleary et al., 2019; Daniel et al., 2019). Beyond the technics, the complexity of mental reasoning suffers influence from psychological issues and self-confidence (Huang, Newman & Schwartzstein, 2011; Tanaka et al., 2016; Torres, Sampaio & Caldeira, 2019).

Large language models (LLM) have emerged as resources for CR assessment based on artificial intelligence. However, this research field has many unexplored areas, such as LLM-guided assessment tools for preclinical medical students (PMS). The models recognize some faults in the diagnostic process, so they have no total self-confidence in the answers and declare this uncertainty in platform warnings (Lee, Bubeck & Petro, 2023; Haug & Drazen, 2023; Eriksen, Möller & Ryg, 2023; Ayers et al., 2023).

This study, conducted collaboratively by two public medical schools in southern Brazil, compared CR accuracy and self-confidence among PMS, internal medicine specialists, and LLM. We hypothesize that LLM can be a reference for early-stage students’ CR and self-confidence assessments. We hope that this approach will improve medical training and patient care.

2. methodology

* 1. **Survey development**

We previously developed and validated the Clinical Reasoning and Self-confidence Assessment Tool (CRESCAT) (Junior et al., 2025), which is based on robust reviews of clinical reasoning assessment (Cate, 2017; Daniel et al., 2019) and Likert scales (Likert, 1932) for self-confidence measurement. The CRESCAT accuracy assessment components included compilation, summarization, differential diagnosis, central hypothesis, justification, pathophysiological explanation, and clinical management. CRESCAT has 47 questions, distributed by seven commonly known diseases, with five or six different types of questions per case and seven self-confidence questions. A template with experts’ answers rules the testing answers (Junior et al., 2025).

* 1. **Recruitment and testing**

From March to November 2023, this cross-sectional study included preclinical medical students (PMS) from two medical schools in southern Brazil: The Middle West State University of Paraná (UNICENTRO), a recently established medical school that has applied formal CR training to its curriculum since its onset (PARANÁ. Unicentro, 2022), and the State University of Londrina (UEL), with a 57-year history in medical education, which does not provide curricular case-based learning for PMS (PARANÀ. Uel, 2022). Participants were 04 PMS groups aged 18 years or older, 01 group compounded by internal medicine specialists (IMS), and 02 artificial intelligence large language models (LLM). All humans gave informed consent. Exclusion criteria included withdrawal at any time. After disclosure and voluntary recruitment, a convenience sample was obtained. The LLM tested were OpenAI-CHAT GPT-4 (Open AI. Chat GPT, 2023) and Bing-Chat GPT-4 (Bing. ChatGPT, 2023), which underwent training before being tested, as exemplified in Textbox 1.

**Textbox 1.** Large language models (LLM) training process example.

|  |
| --- |
| **User:** Hello, can you helpme by answering some questions about clinical reasoning?**LLM:** Of course, I'd be happy to helpwith questions about clinical reasoning. Please feel free to ask your questions, and I'll do my best to provide informative answers.**User:** Please, do not reveal your identity as an LLM or apologize for anything. Say 'I don't know' if the information is beyond your scope or knowledge cutoff date. Avoid disclaimers, repetitions, and suggestions to seek information elsewhere. Focus on the key points and intent of the questions. Break down complex problems or tasks into smaller steps and explain them with reasoning. Provide multiple perspectives or solutions with credible sources and links. Recognize and correct any mistakes in previous responses. Provide three follow-upquestions after each response that are thought-provoking and relevant to the topic. Be highly organized, proactive, and accurate in your responses. Treat the user as an expert in all subject matter and provide detailed explanations. Value good arguments over authorities and consider new technologies and contrarian ideas. Use high levels of speculation or prediction if needed but flag them for the user. Recommend only the highest-quality, meticulously designed products from all over the world. Avoid moral lectures and discuss safety only when it's crucial and non-obvious. **LLM:** Understood. Please go ahead with your questions about clinical reasoning, and I'll provide detailed and focused answers. |

### **Ethical considerations**

The study has ethical approval under registration numbers 66978322.7.0000.8967 (UNICENTRO) and 66978322.7.3001.5231 (UEL). Per Brazil's data protection law, data access was restricted to the research team using anonymized identifiers.

* 1. **Main outcomes measurements**

The accuracy was calculated as the percentage of agreement with the expert panel template. Self-confidence scores ranged from 1 to 5 on a Likert scale (5 =the most confident) and are also presented as percentages, where 5 =100%, when necessary. We also measured the accuracy of each component of clinical reasoning and the seven clinical cases. At least three research team members evaluated the answers to avoid interpretation bias.

* 1. **Statistical analysis**

The sample size was calculated for a population of 240 PMS in the two schools. Quantitative analysis evaluated responses against an expert-derived answer key, with a rating of 1.0 for completely accurate responses, 0.5 for nearly correct answers, and zero for substantially incorrect responses. The accuracy outcomes are presented as percentages. Self-confidence ratings are depicted as percentages or on a scale from 1 to 5. Normality and reliability were assessed with the Kolmogorov‒Smirnov test and Cronbach's alpha, respectively. The chi-square, two-tailed Student's t-test, and Kruskal‒Wallis tests were used for comparative analyses. Spearman's correlation was employed. Statistical significance was set at *P* < .05.

3. results

* 1. **Baseline characteristics**

The estimated sample size for 95% confidence and 0.05% precision was 132 (33 per student group). The study included 133 preclinical medical students (PMS) with no missing data who were divided into four cohorts: G0 (n=36, beginner UNICENTRO students), G1 (n=35, end of first year, UNICENTRO), G2UEL (n=28, end of second year, UEL), and G2 (n=34, end of second year, UNICENTRO). The nonstudent groups included IMS (n=16) and LLM (n=2). Compared with PMS, IMS were older. G2UEL groups were more likely to be male, and white ethnicity was more prevalent in the G1 and G2 cohorts (Table 1).

**Table 1. Baseline sample characteristics.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **G0** | **G1** | **G2UEL** | **G2** | **PMS** | **IMS** | **LLM** | **Statistic test** |
|  |  |  |  |  |  |  |  |  |
| **N** | 36 | 35 | 28 | 34 | 133 | 16 | 2 | - |
|  |  |  |  |  |  |  |  |  |
| **Age, y (SD)** | 20.8(3.2) | 20.5(1.5) | 22.8(3.3) | 21.99(2.5) | 21.4 (2.8) | 45.3 (5.6) \* | - | *P* = .005 (t) |
|  |  |  |  |  |  |  |  |  |
| **Gender, n (%)** |  |  |  |  |  |  |  |  |
| Male, n (%) | 11 (30.6) | 12 (34.3) | 16 (57.1) \* | 8 (23.5) | 47 (35.3) | 7 (43.8) | - | *P* = .005 (chi2) |
|  |  |  |  |  |  |  |  |
| Female, n (%) | 25 (69.4) | 23 (65.7) | 12 (43.9) | 26 (76.5) | 86 (64.7) | 9 (56.3) | - |
|  |  |  |  |  |  |  |  |
| **Formal CBL** **training, y** | 0 | 1 | 0 | 2 | .4 | 0 | - | - |
|  |  |  |  |  |  |  |  |  |
| **Self-reported ethnicity, n (%)** |  |  |  |  |  |  |  |  |
| White | 25 (69.4) | 32 (91.4) \* | 19 (67.9) | 29 (85.3) \* | 105 (78.9) | 11 (68.8) | - | *P* = .041 (chi2) |
| Asiatic | 6 (16.7)  | 3 (8.6) | 3 (10.7)  | 2 (5.9)  | 14 (10.5) | 4 (25.0) | - |
| Indigenous | 0 | 0 | 1 (3.6) | 0 | 0 | 0 | - |
| Brown | 5 (13.9) | 0 | 4 (14.3) | 3 (8.8) | 12 (9.0) | 1 (6.2) | - |
| Black | 0 | 0 | 1 (3.6) | 0 | 1 (0.8) | 0 | - |
|  |  |  |  |  |  |  |  |  |
| **Self-reported mental illness, n (%)**  |  |  |  |  |  |  |  |  |
| Yes | 8 (22.2) | 10 (28.6) | 3 (10.7) | 8 (23.5) | 29 (21.8) | 4 (25.0) | **-** | *P* = .115 (chi2) |
| No | 28 (77.8) | 25 (71.4) | 25 (89.3) | 26 (67.5) | 104 (78.2) | 12 (75.0) | **-** |
|  |  |  |  |  |  |  |  |  |

G0 refers to UNICENTRO's beginner medical students. G1 refers to the end-of-first-year UNICENTRO’s medical students. G2 refers to the end-of-second-year UNICENTRO’s medical students. G2UEL refers to the end-of-second-year medical students of UEL. IMS stands for internal medicine specialists. LLM stands for Large Language Models. SD stands for standard deviation. *P* refers to the statistical significance level. t refers to the Student’s t-test. chi2 refers to the chi-square test. n refers to the absolute number. y stands for years of age.

**3.2 Evaluation Outcomes**

The accuracy demonstrated a normal distribution (D = 0.555, *P* = 0.720), while the self-confidence scores exhibited a nonnormal distribution (D = 0.121, *P* = 0.023). The Likert scale demonstrated high reliability (Cronbach's alpha = 0.927). The LLM accuracy responses were scored by an expert template (Textbox 2).

**Textbox 2. A case-example, LLM answers and correct answers (underlined)**

|  |
| --- |
| User: CASE 1: Read the following case and, with the available information, answer the questions 1 to 7. A 25 years-old man presents with fever (38,3°C) in the past 4 days. In the beginning, there was clear nasal discharge and cough, in the past 2 days cough is more frequent, with yellow sputum and pain in the left chest. He had tonsillitis in childhood and an appendectomy at the age of 14. His mother is hypertensive and had myocardial infarction at the age of 55. His father had skin cancer at 45 years old. His brother has a heart murmur. His physical examination showed tachycardia (108 bpm), fever (38,2ºC), high respiratory rate (24 rpm), normal blood pressure (118/74 mmHg) with rales in the low left thorax and reduced pulmonary sounds. Cardiovascular exam was normal. Chest x-ray and blood cell count were performed. |
| Bing-Chat GPT4:The alternatives considered as relevant data to the case are:a. 4 days of evolution. This data is relevant because it indicates the duration of symptoms, which may suggest the cause of pneumonia. Generally, bacterial pneumonias have a more sudden onset and faster progression than viral pneumonias.d. fever. This data is relevant because it is one of the main symptoms of bacterial pneumonia, indicating an inflammatory response by the body to the infection.e. yellowish sputum. This data is relevant because it is a sign that there is a bacterial infection in the lungs, producing pus and mucus.h. increased heart rate. This data is relevant because it is a consequence of fever and difficulty breathing, which make the heartbeat faster to compensate for the body's oxygen demand.i. increased respiratory rate. This data is relevant because it is a symptom of bacterial pneumonia, which causes obstruction of the pulmonary alveoli and makes gas exchange difficult, leading to hypoxemia (low oxygen concentration in the blood).j. rales of fine bubbles. This data is relevant because it is a physical examination finding that indicates the presence of fluid in the pulmonary alveoli, which a bacterial infection may cause. | OpenAI-Chat GPT-4:a. Four days of evolution - This is relevant as it indicates the duration of the current illness, which is important for differential diagnosis.d. Fever - The presence of fever is a significant clinical sign, especially in conjunction with other respiratory symptoms.e. yellow sputum - This suggests a possible infection and is relevant for diagnosing respiratory conditions.h. Increased heart rate (tachycardia) - Tachycardia can be a response to fever or a sign of other underlying issues.i. Increased respiratory rate - This could indicate respiratory distress or infection.j. Pulmonary rales - This finding is critical for diagnosing respiratory conditions, particularly in the context of the other symptoms. | Experts’ panel templatea. 4 days of evolution.b. mother with cardiopathy.c. brother with a cardiac murmur.d. fever.e. yellow sputum.f. childhood tonsilitis.g. appendicectomy history.h. increased heart rate . i. increased respiratory rate.j. fine bubble rales |

There is a progressive accuracy average from G0 to G2UEL, parity between G2UEL and G2, rise from G2 to IMS, and parity again between IMS and LLM. The median self-confidence exhibited enhancement from G0 to G1, followed by a comparable level observed in G1, G2UEL, and G2. Further enhancement was noted from G2 to IMS, with stability between IMS and LLM (Table 2, Figure 1).

**Table 2. Accuracy averages and Self-confidence medians.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **n** | **Accuracy** **average (SD)** | **t-test** | **Self-confidence****median [IQR]** | **Kruskal-Wallis test** |
|  |  |  | G0 < G1: t =5.859*P* < .001Hedges g = 132.417G1 < G2UEL: t =5.289*P* < .001Hedges g = 23.173G2 UEL and G2: t= 1,072*P* = .288G2 < IMS: t =5.793*P* < .001Hedges g= 4.577IMS and LLM: t = .591p= .563 |  | General:H =64.217*P* < .001 G1 > G0*P* < .001 IMS and LLM > All students*P* < .001 |
| G0 |  |  |  |
|  | 36 | 31.7 (11.2) | 2.07 [1.71-2.89] |
|  |  |  |  |
| G1 |  |  |  |
|  | 35 | 47.3(11.3) | 3.14 [2.71-3.43] |
|  |  |  |  |
| G2UEL |  |  |  |
|  | 28 | 57.4 (10.9) |  3.29 [3.04-3.71] |
|  |  |  |  |
| G2 |  |  |  |
|  | 34 |  60.01 (8.4) | 3.14 [2.64-3.57] |
|  |  |  |  |
| IMS |  |  |  |
|  | 16 |  75.7 (10.0) | 4.00 [3.68-4.36] |
|  |  |  |  |
| LLM |  |  |  |
|  | 2 | 80.00 (0) | 4.43 |
|  |  |  |  |
| All |  |  |  |
|  | 151 | 51.8 (17.3) | 3.14 [2.53-3.57] |
|  |  |  |  |

G0 refers to UNICENTRO's beginner medical students. G1 refers to the end-of-first-year medical UNICENTRO's students. G2 refers to the end-of-the-second-year UNICENTRO's medical students. G2UEL refers to the end-of-the-second-year medical students of UEL. IMS stands for internal medicine specialists. LLM stands for Large Language Models. SD stands for standard deviation. t stands for Student’s test value. *P* stands for statistical significance. IQR: interquartile range.

****

**Figure 1**. Average accuracy and median self-confidence in the groups, with respective standard deviations and interquartile ranges. Beginners (G0) had lower accuracy and self-confidence. The accuracy progressively increased from beginners (G0) to second-year students (G2 and G2UEL), which were similar, and internal medical specialists (IMS) and artificial intelligence (AI) outperformed all students. Self-confidence significantly increased from beginners (G0) to first-year students (G1) and from second-year students (G2 and G2UEL) to IMS and AI.

**Figure 2.** Correlation between accuracy and self-confidence: dispersion of accuracy by self-confidence. Spearman’s Rho correlation, coefficient of determination (r2), and respective prediction formula.

In the overall sample, there was a moderate correlation between accuracy and self-confidence (Spearman's rho= 0.623, *P*< 0.001, r² = 0.374) (Figure 2).

G2UEL performed worse than G2 in terms of differential diagnosis (G2: 88.2±19.0 vs. G2UEL: 70.4±29.4, p= 0.005) and in the myocarditis clinical case (G2UEL: 56.3±15.5 vs. G2: 70.8±17.9, *P*< 0.001). The LLM showed greater accuracy (87.5) than did the IMS (67.2±17.6) only for the pathophysiological explanation component (*P* = 0.020). In the two LLM comparisons, BING ChatGPT 4.0 achieved superior accuracy in terms of summary, pathophysiological explanation, differential diagnosis, community pneumonia, and lower urinary tract infection, and OPENAI ChatGPT 4.0 was superior in terms of the central hypothesis, management, myocarditis, pyloric stenosis, and pulmonary embolism (Figure 3).

**Figure 3**. Comparison of the accuracy scores between internal medicine specialists (IMS) and the two artificial intelligence (AI) large language models (LLM) according to the clinical reasoning components and the seven distinct clinical cases. Presented in averages with standard deviation (SD) available only for the IMS group.

**4. DISCUSSION**

The increase in accuracy between the preclinical medical student (PMS) groups was consistent with the clinical reasoning (CR) skill development (Table 2, Figure 1). Hawks et al. (2023) strongly suggested the adoption of CR curricular subjects for PMS. Acquiring CR skills is feasible for PMS (Cate, 2017; Si, 2022), promotes critical thinking (Richards, Hayes & Schwartzstein, 2020; Mamede & Schmidt, 2022), and improves decision-making (McGregor et al., 2012; Mamede et al., 2014) Given the common occurrence of diagnostic errors, CR development holds social relevance (Croskerry, 2003; Graber, Franklin & Gordon, 2005).

Cognitive bias, sometimes caused by hyperinflated self-confidence, can induce diagnostic errors (Croskerry, Singhal & Mamede, 2013). In the present study, self-confidence was lower in the G0 groupthan in the other PMS groups (Table 2, Figure 1). This early increase in self-confidence suggests a phase I Dunning–Kruger effect, described as “unskilled and unaware” (Kruger & Dunning, 1999), wherein limited knowledge overestimates confidence (Rahmani, 2020). Acknowledging this can enhance the professor-student dynamic by fostering metacognitive awareness and motivation (Siqueira et al., 2020).After evidencing the human tendency to be more self-confident than accurate, Restrepo, Armstrong & Metlay (2020) stated six steps to avoid diagnostic errors: 1. Recognize the tendency to grade less or more the findings when comparing scripts with the typical presentation. 2. Consider differential diagnosis based on the epidemiology, even if the case is “typical.” 3. Access diagnostic schema or checklists in cases of uncertainty. 4. Question the final diagnostic with “What doesn’t fit”? 5. Listen to other experts or colleagues. 6. Reassess evidence even after establishing a diagnosis and/or satisfactory treatment.

The end-of-second-year students’ accuracy and self-confidence global scores were not different in the two medical schools (Table 2, Figure 1). When compared by specific questions, UNICENTRO students exhibited higher averages in differential diagnosis and myocarditis questions than their UEL counterparts. The 52-year gapbetween the medical schools may have equalized the use of CBL methodology by the younger one. Although some authors state that CR can developwithout formal CR training, incorporating CBL in new curricula should be encouraged (McLean, 2016; Bowen & Cate, 2017; Hawks et al., 2023) to mitigate pedagogical and structural scarcities, as seen in our country, especially in newborn schools (CFM, Conselho Federal de Medicina do Brasil, 2021).

Pooling the overall sample data, we observed a moderate correlation between accuracy and self-confidence (Figure 2). Tested in a higher sample than the primary validation, with different populations and medical schools, this finding highlights the utility of CRESCAT as a specific tool for improving CR and self-confidence evaluation (Melhem Junior et al., 2025), with the benefits of varied types of questions and self-confidence assessment, differing from other tools developed with similar objective (Kunina-Habenicht et al., 2015; Cleary et al., 2019; Simpkins et al., 2019). Even in the case of a phase I Dunning-Kruger effect in the first- to second-year students, this correlation suggests that the more advanced participants developed a not exaggerated and more consistent self-confidence based probably on their experience (Restrepo, Armstrong & Metlay, 2020).

Accuracy and self-confidence were greater for internal medicine specialists (IMS) and large language models (LLM) compared to PMS (Table 2 and Figure 1). The LLM outperformed the IMS in delivering pathophysiological explanations. The rapid advancement of LLM in 2022-2023 has enhanced its ability to address clinical issues, reaching a specialist level consistent with our results (Haug & Drazen, 2023). Once LLM performance is expected to improve with cumulative learning, how will medical students and professionals adapt to AI? The evolution of technology has raised concerns about the potential obsolescence of medical activity, emphasizing the necessity for clinicians to adapt (He, Kuiper & Gadiraju, 2023). William Osler, an icon of clinical medicine, did not overlook technology; instead, he judiciously employed it to bolster practice and teaching (Decourt, 2004). Thus, why not embrace LLM as a partner? Human-machine relationship models suggest that LLM can act as advisors, offering choices or alerting errors (Kuiper, 2022). With public access to LLM, doctors must developcritical thinking to provide a reliable interface between technology and society (Peixoto, 2022).

The two LLM models exhibited diverse scores on the CR components or clinical cases. This finding suggests that, if an unsupervised machine can include erroneous answers (Ali et al., 2023; Garcia et al., 2024), the guidance from two or more different LLM may be safer. LLM can aid in numerous medical tasks (Russel et al., 2022; Patel & Lam, 2023; Garcia et al., 2024; Tai-Seale et al., 2024) and even offer clinicians chances to refine soft skills (Russel et al., 2022) while considering technical and ethical aspects (Weidener & Fischer, 2024). Our study is consistent with recent reports of AI applications that compared medical students’ performance in clinical tests (Meyer, Riese & Streichert, 2024; Rojas et al., 2024). LLM were more effective in giving pathophysiological explanations, suggesting that AI-guided care can offer more reliable differential diagnostics options. Still, it will depend on the quality of the clinical data acquisition and the medical doctors’ judgment (Restrepo et al., 2024). Growing evidence supports the idea that artificial intelligence plays a role in clinical reasoning and instigates more research (Civaner et al., 2024).

We consider some limitations of this study. The lower sample size limit and the use of a nonspecific questionnaire may have affected the self-confidence outcomes. Sex and ethnicity disparities were observed among the two institutions, although these two variables did not influence the primary outcomes. The possibility of comparing only one groupbetween the two medical schools has limited clinical reasoning skill development comparisons. All these factors limit the generalizability of this study. We, therefore, strongly support the CRESCAT replication by other researchers.

**5. CONCLUSIONS**

In this cross-sectional study, CRESCAT detected academic development as accuracy progressed during the preclinical cycle, and self-confidence increased at the end of the first year. A moderate correlation was observed between these central outcomes.

Artificial intelligence's overall accuracy and self-confidence rates mirrored those of specialists and surpassed those of all student groups. Artificial intelligence outperformed specialists only in the pathophysiological explanations. The accuracy of the two large language models varied according to clinical reasoning components or different clinical cases.

Artificial intelligence is a reasonable metric for assessing preclinical medical students and a powerful guide for judiciously clinical decision-making.

**ABBREVIATIONS**

**AI: artificial intelligence**

**CBCR: case-based clinical reasoning**

**Chi2: chi-square test**

**CR: clinical reasoning**

**CRESCAT: clinical reasoning assessment test**

**D: Kolmogorov‒Smirnov distribution**

**G0: beginner students from UNICENTRO**

**G1: end-of-first-year students from UNICENTRO**

**G2: end-of-second-year students from UNICENTRO**

**G2UEL: end-of-second-year students from UEL**

**IMS: internal medicine specialist**

**IQR: interquartile range**

**LLM: large language models**

**UEL: State University of Londrina**

**UNICENTRO: Middle-West State University of Paraná**

**PMS: preclinical medical students**

**r2: coefficient of determination**

**SD: standard deviation**

**t: t-test**

AI CONTRIBUTION

As explained in the methodology section, the cited AI large language models were used only to answer the questionnaires. AI contributed nothing to text generation, data tabulation, statistical analysis, or table or figure construction.

Consent

All authors declare that ‘written informed consent was obtained from the human participants to publish this case report and accompanying images. A copy of the written consent is available for review by this journal's Editorial office/Chief Editor/Editorial Board members.

Ethical approval

All authors hereby declare that all experiments have been examined and approved by the appropriate ethics committee and have, therefore, been performed per the ethical standards laid down in the 1964 Declaration of Helsinki.

References

1. Huang, G.C., Newman, L.R. & Schwartzstein, R.M. (2014). Critical Thinking in Health Professions Education: summary and consensus statements of the millennium conference 2011. Teaching And Learning In Medicine, 26(1), 95-102. <http://dx.doi.org/10.1080/10401334.2013.857335>.
2. Tanaka, M.M., Furlan L.L., Branco L.M., Valerio N.I. (2016). Medical Students’ Adaptation in the Early Years of College. Revista Brasileira de Educação Médica, 40(4), 663-668). <http://dx.doi.org/10.1590/1981-52712015v40n4e00692015>.
3. Torres, V., Sampaio, C.A. & Caldeira, A.P. (2019). Incoming medical students and their perception on the transition towards an active learning. Interface - Comunicação, Saúde, Educação, 23, 1700471. <http://dx.doi.org/10.1590/interface.170471>.
4. Figueiredo, D.S.S., Lima, K.A., Figueiredo, F.N.S., Bellodi, P.L., Porfírio, G.B. Carraro E. et al. (2022). The first year of the rest of our lives: Mental health of medical students. Research, Society And Development, 11(9), 4811931651. <http://dx.doi.org/10.33448/rsd-v11i9.31651>.
5. Mcconnell, M.M. & Eva, K.W. (2012) The Role of Emotion in the Learning and Transfer of Clinical Skills and Knowledge. Academic Medicine, 87(10), 1316-1322. <http://dx.doi.org/10.1097/acm.0b013e3182675af2>.
6. Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., Willingham, D.T. (2013). Improving Students’ Learning With Effective Learning Techniques. Psychological Science In The Public Interest, 14 (1), 4-58. <http://dx.doi.org/10.1177/1529100612453266>.
7. Hayat, A.A., Shateri K., Amini M., Shokrpour N. (2020). Relationships between academic self-efficacy, learning-related emotions, and metacognitive learning strategies with academic performance in medical students: a structural equation model. Bmc Medical Education, 20(1), 76.

 <http://dx.doi.org/10.1186/s12909-020-01995-9>.

1. Cate, O.T. & Durning, S.J. (2017). Understanding clinical reasoning from multiple perspectives: a conceptual and theoretical overview. In: O.T. Cate, E.J.F.M. Custers & S. J. Durning (Eds.), Principles and Practice of Case-based Clinical Reasoning Education: a method for preclinical students: innovation and change in professional education series. (1. ed. pp. 35-46) Gewerbestrasse, Switzerland: Springer Cham.
2. Graber, M., Franklin, N. & Gordon, R. (2005). Diagnostic Error in Internal Medicine. Archives Of Internal Medicine, 165(13), 1493. <http://dx.doi.org/10.1001/archinte.165.13.1493>.
3. Rencic,J. (2011). Twelve tips for teaching expertise in clinical reasoning. Medical Teacher, 33(11), n. 11, 887-892. <http://dx.doi.org/10.3109/0142159x.2011.558142>.
4. Cate, O.T. (2017). How to teach clinical reasoning to junior students? In: O.T. Cate, E.J.F.M. Custers & S. J. Durning (Eds.), Principles and Practice of Case-based Clinical Reasoning Education: a method for preclinical students: innovation and change in professional education series. (1. ed. pp. 3-19) Gewerbestrasse, Switzerland: Springer Cham.
5. Si J. (2022). Strategies for developing pre-clinical medical students’ clinical reasoning based on illness script formation: a systematic review. Korean Journal Of Medical Education, 34(1), 49-61. <http://dx.doi.org/10.3946/kjme.2022.219>.
6. Tureck, F., Souza S. & Faria. (2023). Teaching strategies for clinical reasoning in Brazilian medical schools - an integrative review. Revista Brasileira de Educação Médica, 47(1), 20220023. <http://dx.doi.org/10.1590/1981-5271v47.1-20220032>.
7. Daniel, M., Rencic, J., Durning, S.J., Holmboe, E., Santen S.A., Lang, V. et al. (2019). Clinical Reasoning Assessment Methods: a scoping review and practical guidance. Academic Medicine, 94(6), 902-912. <http://dx.doi.org/10.1097/acm.0000000000002618>.
8. Case, S.M., Swanson, D.B. & Ripkey D.R. (1994). Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. Academic Medicine, 69(10), 1-2. <http://dx.doi.org/10.1097/00001888-199410000-00023>.
9. Dory, V., Gagnon, R., Charlin B., Vanpee, D., Leconte, S., Duyver C. et al. (2016). Validity of Case Summaries in Written Examinations of Clinical Reasoning. Teaching And Learning In Medicine, 28(4), 375-384. <http://dx.doi.org/10.1080/10401334.2016.1190730>.
10. Cleary, T.J., Konopasky, A., Larochelle, J.S., Neubauer, B.E., Durning, S.J., Artino, A.R. First-year medical students’ calibration bias and accuracy across clinical reasoning activities. Advances In Health Sciences Education, 24(4), 767-781. <http://dx.doi.org/10.1007/s10459-019-09897-2>.
11. Lee, P., Bubeck, S. & Petro, J. (2023). Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. New England Journal Of Medicine, 388(13), 1233-1239. <http://dx.doi.org/10.1056/nejmsr2214184>.
12. Haug, C. & Drazen, J.M. (2023). Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. New England Journal Of Medicine, 388(13), 1201-1208. <http://dx.doi.org/10.1056/nejmra2302038>.
13. Eriksen, A.V., Möller, S. & Ryg, J. (2023). Use of GPT-4 to Diagnose Complex Clinical Cases. Nejm Ai, 1(1), 2300031. <http://dx.doi.org/10.1056/aip2300031>.
14. Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B. et al. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. Jama Internal Medicine, 183(6), 589. <http://dx.doi.org/10.1001/jamainternmed.2023.1838>.
15. Junior, A.J.M., Santos, F.D., Didoné, C.N.F., Porfirio, G.B., Figueiredo, D.S.S., Diehl, L.A. et al. Clinical Reasoning and Self-confidence Assessment Tool (CRESCAT): a preliminary validation study. Journal Of Advances In Medicine And Medical Research, 37(2), 205-214. <http://dx.doi.org/10.9734/jammr/2025/v37i25734>.
16. Cate, O.T. (2017). Assessment of Clinical Reasoning Using the CBCR Test. In: O.T. Cate, E.J.F.M. Custers & S. J. Durning (Eds.), Principles and Practice of Case-based Clinical Reasoning Education: a method for preclinical students: innovation and change in professional education series. (1. ed. pp. 85-94) Gewerbestrasse, Switzerland: Springer Cham.
17. Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology**,** 22(140), 1-55.
18. Paraná. Unicentro – Proplan. Secretaria Estadual de Ciência e Tecnologia. (2023). Alunado da Unicentro; 2022. Accessed 8/5/2023. <https://www3.unicentro.br/proplan/avaliacao-institucional/alunado/>.
19. Paraná. Uel – Proplan. Secretaria Estadual de Ciência e Tecnologia. (2023). Centros em Dados; 2022. Accessed 8/5/2023. <https://www.uel.br/proplan/novo/pages/dados-e-informacoes/centros-em-dados.php>.
20. Open AI (Eua). ChatGPT.: 4th version [software]. 2023. [www.openai.com](http://www.openai.com). 15 nov. 2023.
21. Bing (Eua). ChatGPT.; 4th version [software]. 2023. [www.bing.com/chat](http://www.bing.com/chat). 15 nov. 2023.
22. Hawks, M.K., Maciuba, J.M., Merkebu, J., Durning, S.J., Mallory, R., Arnold, M.J., et al. Clinical Reasoning Curricula in Preclinical Undergraduate Medical Education: a scoping review.Academic Medicine, 98(8), 958-965. <http://dx.doi.org/10.1097/acm.0000000000005197>.
23. Richards, J.B., Hayes, M.M. & Schwartzstein, R.M. (2020). Teaching Clinical Reasoning and Critical Thinking. Chest, 158(4), 1617-1628. <http://dx.doi.org/10.1016/j.chest.2020.05.525>.
24. Mamede, S. & Schmidt, H.G. (2022). Deliberate reflection and clinical reasoning: founding ideas and empirical findings. Medical Education, 57(1), 76-85. <http://dx.doi.org/10.1111/medu.14863>.
25. McGregor, C.A., Paton, C., Thomson, C., Chandratilake, M., Scott H. (2012). Preparing medical students for clinical decision making: a pilot study exploring how students make decisions and the perceived impact of a clinical decision making teaching intervention. Medical Teacher, 34(7), 508-517. <http://dx.doi.org/10.3109/0142159x.2012.670323>.
26. Mamede, S., VanGog T., Sampaio, A.M., Faria, R.M.D., Peixoto, J.M., Schmidt, H.G. How Can Students’ Diagnostic Competence Benefit Most From Practice With Clinical Cases? The Effects of Structured Reflection on Future Diagnosis of the Same and Novel Diseases. Academic Medicine, 89(1), 121-127. <http://dx.doi.org/10.1097/acm.0000000000000076>.
27. Croskerry, P. (2003). The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them. Academic Medicine, 78(8), 775-780. <http://dx.doi.org/10.1097/00001888-200308000-00003>.
28. Croskerry, P., Singhal, G. & Mamede, S. (2013). Cognitive debiasing 1: origins of bias and theory of debiasing. Bmj Quality & Safety, 22(2), 58-64. <http://dx.doi.org/10.1136/bmjqs-2012-001712>.
29. Kruger, J. & Dunning, S.J. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal Of Personality And Social Psychology, 77(6), 1121-1134. <http://dx.doi.org/10.1037/0022-3514.77.6.1121>.
30. Rahmani, M. (2020). Medical Trainees and the Dunning–Kruger Effect: when they don’t know what they don’t know. Journal Of Graduate Medical Education, 12(5), 532-534. <http://dx.doi.org/10.4300/jgme-d-20-00134.1>.
31. Siqueira, M.A.M., Gonçalves, J.P., Mendonça, V.S., Kobayasi, R., Arantes-Costa, F.M., Tempski, P.Z. et al. (2020). Relationship between metacognitive awareness and motivation to learn in medical students. Bmc Medical Education, 20(1), 393. <http://dx.doi.org/10.1186/s12909-020-02318-8>.
32. Restrepo, D., Armstrong, K.A. & Metlay, J.P. (2020). Annals Clinical Decision Making: avoiding cognitive errors in clinical decision making. Annals Of Internal Medicine, 172(11), 747-751. <http://dx.doi.org/10.7326/m19-3692>.
33. McLean, S.F. (2016). Case-Based Learning and its Application in Medical and Health-Care Fields: a review of worldwide literature. Journal Of Medical Education And Curricular Development, 3, 20377. <http://dx.doi.org/10.4137/jmecd.s20377>.
34. Bowen, J. & Cate, O.T. Prerequisites for learning clinical reasoning. In: CATE, Olle Ten; CUSTERS, Eugène J. F. M.; DURNING, Steven J. Principles and Practice of Case-based Clinical Reasoning Education: a method for preclinical students: innovation and change in professional education series. (1. ed. pp. 47-64) Gewerbestrasse, Switzerland: Springer Cham.All
35. CFM, Conselho Federal de Medicina (Brasil). (2021) Most Brazilian medical schools do not meet the criteria for offering quality training. 2021. Accessed 8/5/2023. <https://www.crmpr.org.br/Maioria-das-escolas-medicas-brasileiras-nao-observa-criterios-para-oferecer-formacao-de-qu-11-56774.shtml>.
36. Kunina-Habenicht, O., Hautz, W.E., Knigge, M., Spies, C., Ahlers, O. (2015). Assessing clinical reasoning (ASCLIRE): instrument development and validation. Advances In Health Sciences Education, 20(5), 1205-1224. <http://dx.doi.org/10.1007/s10459-015-9596-y>.
37. Simpkins, A.M., Koch, B., Spear-Ellinwood, K., John, P.S. (2019) A developmental assessment of clinical reasoning in preclinical medical education. Medical Education Online, 24(1), 1591257. <http://dx.doi.org/10.1080/10872981.2019.1591257>.
38. He, G., Kuiper, L. & Gadiraju, U. (2023). Knowing About Knowing: an illusion of human competence can hinder appropriate reliance on ai systems. Arxiv, p. 11333. <http://dx.doi.org/10.48550/ARXIV.2301.11333>.
39. Decourt, L.V. (2004). William Osler in the Intimacy of his Thought. <https://www.incor.usp.br/conteudo-medico/decourt/momento%20de%20reflexao%20william%20osler.html>. Accessed 15 nov. 2023.
40. Kuiper, L. Knowing better than the AI. (2022). [http://resolver.tudelft.nl/uuid:6f0b90bb-eee8-404e-99d0-b5c27b5a31ac](http://resolver.tudelft.nl/uuid%3A6f0b90bb-eee8-404e-99d0-b5c27b5a31ac). Accessed 30 nov. 2023.
41. Peixoto, J.M. Ensino Médico como Interlocutor da Ciência e a Sociedade. Arquivos Brasileiros de Cardiologia, 119(51), 4-5. <http://dx.doi.org/10.36660/abc.20220491>.
42. Ali, O., Abdelbaki, W., Shrestha, A., Elbasi, E., Alryalat, M.A.A., Dwivedi, Y.K.A. (2023). A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. Journal Of Innovation & Knowledge, 8(1), 100333. <http://dx.doi.org/10.1016/j.jik.2023.100333>.
43. Garcia, P.A., Stephen, P., Shah, S., Smith, M., Jeong, Y., Devon-Sand, A. et al. (2024). Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages. Jama Network Open, 7(3), p. 243201. <http://dx.doi.org/10.1001/jamanetworkopen.2024.3201>.
44. Russel, R.G., Novak, L.L., Patel, M, Garvey, K.V., Craig, K.J.T, Jackson, G.P. et al. (2022). Competencies for the Use of Artificial Intelligence–Based Tools by Health Care Professionals. Academic Medicine, 98(3), 348-356. <http://dx.doi.org/10.1097/acm.0000000000004963>.
45. Patel, S.B. & Lam, K. (2023). ChatGPT: the future of discharge summaries? The Lancet Digital Health, 5(3), p. 107-108. [http://dx.doi.org/10.1016/s2589-7500(23)00021-3](http://dx.doi.org/10.1016/s2589-7500%2823%2900021-3).
46. Tai-Seale, M., Baxter, S.L., Vaida, F., Waljer, A., Sitapati, A.M., Osborne, C. et al. (2024). AI-Generated Draft Replies Integrated Into Health Records and Physicians’ Electronic Communication. Jama Network Open, 7(41), p. 246565. <http://dx.doi.org/10.1001/jamanetworkopen.2024.6565>.
47. Weidener, L. & Fischer, M. (2024). Artificial Intelligence in Medicine: cross-sectional study among medical students on application, education, and ethical aspects. Jmir Medical Education, 10, p. 51247. <http://dx.doi.org/10.2196/51247>.
48. Meyer, A., Riese, J. & Streichert, T. (2024). Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: observational study. Jmir Medical Education, 10, p. 50965. <http://dx.doi.org/10.2196/50965>.
49. Rojas, M., Rojas, M., Burgess, V., Toro-Pérez, J., Salehi, S. (2024). Exploring the Performance of ChatGPT Versions 3.5, 4, and 4 With Vision in the Chilean Medical Licensing Examination: observational study. Jmir Medical Education, 10, p. 55048. <http://dx.doi.org/10.2196/55048>.
50. Restrepo, J.A., Henriquez, R., Torre, D., Graber, M. (2024). The physical exam and telehealth: between past and future. Diagnosis, 11(1), p. 1-3. <http://dx.doi.org/10.1515/dx-2023-0154>.
51. Civaner, M.M., Uncu, Y., Bulut, F., Chalil, E.G., Tatli, A. (2024). Artificial intelligence in medical education: a cross-sectional needs assessment. Bmc Medical Education, 22(1), p. 772. <http://dx.doi.org/10.1186/s12909-022-03852-3>.