

Original Paper

# What Platform Scores Miss: Multidimensional Evaluation of AI Teaching Agents in Medical Education

Hui Zhang<sup>1</sup>, MM; Lihui Qu<sup>2</sup>, PHD; Jianmin Zheng<sup>1</sup>, MMgmt; Yi Xiong<sup>1</sup>, MSc; Hongbo Bai<sup>3</sup>, MSc; Ruiying Ji<sup>3</sup>, MSc; Guohui Liu<sup>3</sup>, MSc; Wanling Chen<sup>4</sup>, MD; Zirui Cheng<sup>5</sup>, MD; Youbang Chen<sup>3</sup>, MSc; Chun-tao Yang<sup>3</sup>, PhD

<sup>1</sup>Office of Academic Affairs, Guangzhou Medical University, Guangzhou, China

<sup>2</sup>Department of Physiology, School of Basic Medical Sciences, The Fourth Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

<sup>3</sup>Guangzhou Municipal and Guangdong Provincial Key Laboratory of Protein Modification and Disease, Department of Physiology, School of Basic Medical Sciences, Guangzhou Medical University, Guangzhou, Guangdong, China

<sup>4</sup>Department of Endocrinology, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

<sup>5</sup>Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China

**Corresponding Author:**

Chun-tao Yang, PhD

Guangzhou Municipal and Guangdong Provincial Key Laboratory of Protein Modification and Disease

Department of Physiology, School of Basic Medical Sciences, Guangzhou Medical University

1 Xinzao Road

Guangzhou, Guangdong 511436

China

Phone: 86 20-37103213

Email: [cyang@gzhmu.edu.cn](mailto:cyang@gzhmu.edu.cn)

## Abstract

**Background:** Large language model (LLM)-based AI teaching agents are increasingly used in medical education, yet their pedagogical quality is typically judged by platform-generated scores whose scoring criteria are undisclosed and may not reflect the teaching quality of the agent.

**Objective:** This study aimed to develop and validate a multidimensional rubric for evaluating AI teaching agents and to examine the correspondence between platform scores and rubric-based teaching quality.

**Methods:** Eight AI teaching agents covering an endocrinology curriculum were deployed across 4 role-play paradigms (patient, student, expert, and family). Twenty-two fourth-year medical students generated 167 dialogues, which were scored both by the platform and by an independently applied 8-dimension rubric (100 points, covering knowledge accuracy, pedagogical guidance, knowledge coverage, role-play quality, difficulty calibration, medical safety, student engagement, and feedback quality). Each dialogue was scored 4 times by a primary evaluator (Claude Opus 4.8; mean within-model SD 0.36), with 2 additional LLMs as robustness checks; 40 dialogues spanning all agents were rescored by a medical-education expert for validation.

**Results:** Platform and rubric rankings diverged for most agents: the agent ranked third by the platform ranked last on rubric-based quality, and the platform's fourth-ranked agent ranked first. Agents differed most on knowledge-related dimensions (knowledge coverage coefficient of variation=27.3%) and least on role-play quality (coefficient of variation=5.7%), while difficulty calibration was a shared weakness. In a case-level observation, one agent revised specifically to strengthen empathy attained high role-play quality yet the lowest knowledge coverage of all agents. AI scores agreed with expert ratings at the total-score level (intraclass correlation coefficient=0.51) and on cognitive-process dimensions, but agreement was low for the more subjective dimensions. Student gender showed no detectable effect, though this analysis was underpowered.

**Conclusions:** In this exploratory study, platform-generated scores reflected a construct different from agent teaching quality and should be used as a complement rather than as the sole quality indicator. The 8-dimension rubric provides a transparent, standardized alternative that reveals differences missed by platform scores, including a lack of association between empathy and knowledge coverage that warrants attention in future agent design.

*JMIR Med Educ* 2026;12:e96819; doi: [10.2196/96819](https://doi.org/10.2196/96819)

**Keywords:** AI teaching agents; educational data science; educational evaluation; endocrinology; large language models; LLMs; multidimensional rubric; precision education; role-play

## Introduction

Bridging the gap between theoretical knowledge and clinical competence remains a central challenge in medical education, particularly in disciplines requiring multisystem integration such as endocrinology [1,2]. Simulated patient encounters are among the most effective approaches for developing clinical reasoning and communication skills, but their scalability is constrained by the cost and limited availability of standardized human patients [3-7]. Large language model (LLM)-based AI agents have emerged as a promising alternative, achieving clinical fidelity comparable to standardized human patients at substantially lower cost [8-13], and they are now being adopted in medical curricula at a pace that outstrips our ability to evaluate them. Yet nearly all existing implementations adopt a single role-play paradigm in which the AI portrays a patient and the student acts as a physician, without examining whether alternative role configurations produce different teaching outcomes, or how the quality of any such agent should be judged in the first place.

These developments expose a more fundamental gap: we lack a validated way to judge whether an AI teaching agent teaches well. Existing evaluation frameworks address clinical intervention outcomes or clinical task accuracy [14-16] but not the teaching-specific dimensions that determine educational value, such as adaptive difficulty calibration, knowledge coverage, and formative feedback quality. In practice, quality is often inferred from the aggregate scores generated by commercial teaching platforms, yet these scores are produced by undisclosed criteria that cannot be independently verified, are not decomposable by dimension, and are not comparable across agents, leaving it unclear what they actually measure and whether they reflect the teaching quality of the agent or merely the surface characteristics of a student's responses. This evaluation gap has a direct consequence: without a trustworthy measure of teaching quality, the design variables that might improve these agents, including role-play configuration, content domain, and learner characteristics, such as gender, cannot be studied systematically and remain largely unexplored within a single controlled context [6,17].

Endocrinology is well suited to investigating these questions, spanning molecular mechanisms to chronic disease management within a single curricular module, yet remaining heavily reliant on traditional didactics, with simulation-based approaches underrepresented [18-20]. Against this background, we deployed 8 AI teaching agents covering a complete endocrinology curriculum, each employing a distinct role-play paradigm, and analyzed 167 student-agent dialogues. Because the platform's built-in scores proved uninterpretable as measures of teaching quality, we developed a purpose-built 8-dimension rubric, validated against multiple independent LLM evaluators and a human expert, as a transparent external benchmark. To our knowledge, this is the first study to (1) propose and validate a multidimensional

framework for evaluating AI teaching agents in medical education, (2) examine how role-play design, content domain, and learner gender relate to teaching quality within a unified curriculum, and (3) compare platform-generated scores against a transparent external standard to clarify what each actually captures. Beyond the specific agents studied, this work offers a reusable approach to a problem the field will increasingly face: how to see what an AI teaching agent is actually teaching.

## Methods

### Study Design and Participants

This cross-sectional observational study was conducted during the autumn semester of 2025 at Guangzhou Medical University with 22 fourth-year undergraduate students majoring in basic medical sciences, who participated as part of their endocrinology curriculum. The cohort comprised 15 female and 7 male students. Each student interacted with all 8 AI teaching agents sequentially over the course of the semester. Owing to individual absences, the number of completed dialogues per agent ranged from 19 to 22, yielding 167 valid dialogues in total.

### Ethical Considerations

All participants were informed of the study purpose, and the study was approved by the institutional review board of Guangzhou Medical University (202607001).

### AI Teaching Agent Design and Deployment

Eight AI teaching agents were constructed and deployed on the Chaoxing e-Learning platform (Beijing Century Chaoxing Information Technology Development Co, Ltd), a widely used commercial platform in Chinese higher education. Each agent covered 1 chapter of the endocrinology curriculum. Two chapters addressed foundational content (a general introduction and organ morphology), and 6 addressed clinical or disease-oriented content (hypothalamic-pituitary diseases, thyroid diseases, adrenal diseases, glucose metabolism disorders, lipid metabolism disorders, and calcium-phosphorus metabolism disorders).

Each agent employed a distinct role-play scenario assigned to 1 of 4 role paradigms: patient (A1, A4, A5), student (A2, A3), expert (A6, A7), and family member (A8); full configurations are provided in Table 1. The 4 paradigms were chosen to represent 4 prototypical interaction scenarios in clinical medicine (doctor and patient, teacher and student, expert consultation, and family communication), thereby sampling a range of communicative and pedagogical demands rather than instantiating a single learning theory a priori. Where relevant, the resulting behaviors could be related post hoc to established educational concepts; for example, the student paradigm elicits scaffolding within the learner's zone

of proximal development [20,21]. We emphasize that because each agent combined 1 role paradigm with 1 content chapter, role and content were confounded by design.

Key implementation differences among the agents lay in the assigned AI persona and the reciprocal role assigned to the student, which together determined the communicative register and the pedagogical stance of each dialogue (Table 1). For example, within the expert paradigm, agent A6 positioned the student as a learner in a mentorship model, whereas agent A7 positioned the student as a patient, providing both a physician and a patient perspective. Agent

A8 was deliberately revised after the initial deployment of the first 7 agents, when their empathic engagement was judged to be limited; only its prompt was modified, while all other settings were held constant, to strengthen an empathic family-member persona. For each agent, the research team specified the instructional content, learning objectives, knowledge sources, role-play scenario, and safety and behavioral constraints; the complete per-agent configuration, including prompts, is provided in Multimedia Appendix 1.

**Table 1.** Overview of AI teaching agent configurations.

| Agent | Content topic                           | Role paradigm | AI role                                         | Student role                                     |
|-------|-----------------------------------------|---------------|-------------------------------------------------|--------------------------------------------------|
| A1    | General introduction to endocrinology   | Patient       | Patient (endocrinology inpatient)               | Physician (endocrinology intern)                 |
| A2    | Organ morphology                        | Student       | Learner (first-year medical student)            | Instructor (anatomy and histology teacher)       |
| A3    | Hypothalamic-pituitary diseases         | Student       | Learner (average-performing student)            | Instructor (top-performing class representative) |
| A4    | Thyroid diseases                        | Patient       | Patient (inquisitive, detail-seeking patient)   | Physician (patient's physician)                  |
| A5    | Adrenal diseases                        | Patient       | Patient (Cushing syndrome patient)              | Physician (endocrinologist)                      |
| A6    | Glucose metabolism disorders            | Expert        | Expert (senior diabetes clinician-researcher)   | Learner (basic medical science student)          |
| A7    | Lipid metabolism disorders              | Expert        | Expert (attending physician)                    | Patient (patient with obesity)                   |
| A8    | Calcium-phosphorus metabolism disorders | Family        | Family member (older patient with osteoporosis) | Caregiver (grandchild)                           |

## Data Collection

Two types of data were collected from each student-agent dialogue. First, the platform's built-in scoring system generated a single aggregate score per dialogue as an automated assessment of student performance. The platform does not disclose how these scores are generated, and its scoring mechanism could not be independently obtained or verified; only aggregate scores were available to instructors, with no dimensional breakdown. Second, complete dialogue transcripts were exported from the platform for independent external evaluation.

## Eight-Dimension Pedagogical Evaluation Rubric

To enable standardized cross-agent comparison of teaching quality, we developed an 8-dimension evaluation rubric (Table 2) informed by existing frameworks for AI

conversational agents in health care [14], AI agent evaluation in clinical medicine [16], and educational agent effectiveness research [17], supplemented by expert discussion within the research team. Unlike platform-generated scoring, the rubric was designed to assess agent teaching quality rather than student response quality. The complete scoring prompt is provided in Multimedia Appendix 2.

Dimension weights were assigned by the research team on conceptual grounds: medical knowledge accuracy and pedagogical guidance, as the core teaching functions, received the highest weight (20 points each), whereas student engagement elicitation and feedback quality, as supportive dimensions, received the lowest (5 points each). We note as a limitation that these weights were set without formal student or educator input, and that empirically derived weights are a direction for future work.

**Table 2.** Eight-dimension pedagogical evaluation rubric for AI teaching agents.

| Dimension                           | Points (total=100) | Scoring focus                                                   |
|-------------------------------------|--------------------|-----------------------------------------------------------------|
| Medical knowledge accuracy          | 20                 | Factual correctness and absence of medical errors               |
| Pedagogical guidance ability        | 20                 | Use of questioning, scaffolding, and structured guidance        |
| Knowledge-point coverage efficiency | 15                 | Breadth and completeness of curriculum content addressed        |
| Role-play quality                   | 15                 | Consistency, naturalness, and appropriateness of role portrayal |
| Adaptive difficulty calibration     | 10                 | Adjustment of complexity in response to student level           |
| Medical safety boundary             | 10                 | Avoidance of harmful advice and appropriate referral behavior   |
| Student engagement elicitation      | 5                  | Prompting of active student participation and critical thinking |
| Assessment and feedback quality     | 5                  | Specificity and constructiveness of formative feedback          |

## Rubric-Based Scoring and Validation

All 167 dialogue transcripts were scored across the 8 dimensions using LLMs operating under standardized instructions [15]. To assess the robustness and reliability of the scoring, 3 LLMs from different developers were used as evaluators: Claude Opus 4.8 (Anthropic), representing a leading international model, and DeepSeek (DeepSeek V4 Artificial Intelligence Basic Technology Research Co, Ltd) and Qwen 2.5 (Alibaba Cloud), two widely used Chinese LLMs selected for their strong Chinese-language capability given that all dialogues used Chinese. DeepSeek and Qwen were added at the recommendation of a reviewer and on the basis of our recent study of mainstream domestic and international LLMs for medical assessment tasks [22].

All scoring was performed programmatically through each model's application programming interface under a fixed request configuration, with an intercall interval of 1.0 second and up to 3 retries per call. To promote deterministic, reproducible outputs, DeepSeek and Qwen were queried with temperatures of 0 and top-p of 0.01; Claude Opus 4.8 was queried at its default setting. Each model independently applied the identical 8-dimension rubric and scored every dialogue 4 times, and the mean of the 4 runs was used as that model's final score. This repeated-scoring design followed directly from our recent finding that the reliability of LLM-based scoring is better captured by repeated runs and cross-run variance than by single-run performance [22]. Within-model reproducibility across the 4 runs was quantified by the SD of the repeated scores; reproducibility was high, indicating that scores were stable rather than single, unrepeatable judgments [23].

To calibrate the AI-based scoring against human judgment, a medical-education expert independently rescored a stratified subset of 40 dialogues using the identical 8-dimension rubric. The subset was stratified by Claude total score to span high, medium, and low quality and to cover all 8 agents, and the expert was blinded to the AI scores. Agreement between AI and expert ratings was quantified by the Pearson correlation and the intraclass correlation coefficient (ICC), computed at the total-score level and per dimension [24]. Because rescoring was performed by a single expert, expert-to-expert reliability could not be computed; this is addressed in the Limitations section.

## Statistical Analysis

Given nonnormal score distributions and heterogeneous variances across agents (Shapiro-Wilk and Levene tests),

nonparametric methods were used throughout. Platform-generated scores were compared across agents using the Kruskal-Wallis  $H$  test, with post hoc Dunn tests and Bonferroni correction for pairwise comparisons. Effect sizes were reported as  $\eta^2$  for omnibus tests and Cohen  $d$  for selected contrasts. Between-agent variation in each rubric dimension was summarized using the coefficient of variation (CV).

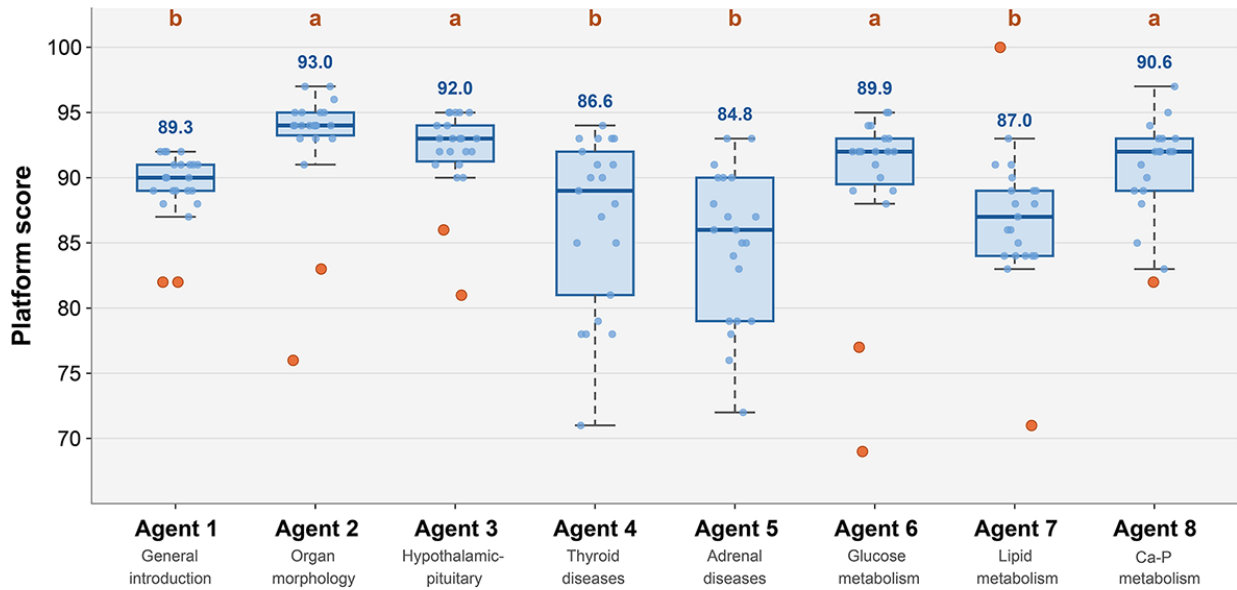
Comparisons across role paradigms were treated as descriptive, because the independent unit was the agent and each role group comprised only 1 to 3 agents; formal inferential testing was therefore not performed at the agent level. The divergence between platform-based and rubric-based agent rankings was characterized through concrete rank comparisons rather than a single agent-level correlation coefficient, which would be unstable at this sample size. Learner gender differences were examined across dimensions using Mann-Whitney  $U$  tests, with results interpreted in light of the small and imbalanced sample. The correspondence between rubric-based agent quality and students' end-of-course chapter examination performance was examined descriptively. All analyses were conducted in Python (version 3.11), with a 2-sided significance threshold of  $P<.05$ .

## Results

### Platform-Generated Scores Varied Across Agents

Chaoxing platform-generated scores from 167 student-agent dialogues differed significantly across the 8 agents (Kruskal-Wallis  $H_7=57.981$ ;  $P<.001$ ;  $\eta^2=0.321$ ; Figure 1), with 9 significant pairwise differences identified by post hoc Dunn tests (Bonferroni-corrected  $P<.05$ ). Basic science agents (A2, A3) received the highest scores, while clinical disease agents (A4, A5, A7) received the lowest. While these differences indicate that agent design influenced student performance under platform scoring, the platform does not disclose how its scores are generated, and its scoring mechanism could not be independently verified, precluding any interpretation of what drove them or whether they reflect genuine agent teaching quality. Therefore, the 8-dimension rubric was applied as an independent external audit, reported in the following sections.

**Figure 1.** Distribution of Chaoxing platform-generated scores across 8 agents (22 students; 167 valid dialogues). Box plots show median, IQR, and 1.5×IQR whiskers; individual scores are jittered dots; outliers (>1.5×IQR) are in orange. Blue numerals above boxes denote means. Agents not sharing a common letter (a vs b) differ significantly (Bonferroni-corrected  $P<.05$ ).



**LLM Evaluators Differed in Scoring Discrimination**

To evaluate the robustness of the rubric-based scoring, the 8-dimension rubric was independently applied by 3 LLM

evaluators from different developers (Claude, DeepSeek, and Qwen), each scoring all 167 dialogues 4 times, with high within-model reproducibility (Table 3; mean within-model  $SD\leq0.36$ ).

**Table 3.** Comparison of 3 LLM evaluators on the 8-dimension rubric.

| Metric                                        | Claude (Opus 4.8) | DeepSeek V4 | Qwen 2.5 |
|-----------------------------------------------|-------------------|-------------|----------|
| Dialogues scored (×4 runs), n                 | 167               | 167         | 167      |
| Intrarater SD (score stability) <sup>a</sup>  | 0.36              | 0.06        | 0.17     |
| Agent-total score range                       | 61-92             | 82-95       | 75-100   |
| Agent-total SD (discrimination) <sup>b</sup>  | 11.6              | 4.0         | 7.3      |
| Overall ceiling rate <sup>c</sup> , %         | 17                | 30          | 51       |
| Mean achievement rate, %                      | 82                | 90          | 92       |
| Rank correlation with Claude (ρ) <sup>d</sup> | — <sup>e</sup>    | 0.64        | 0.71     |

<sup>a</sup>Intrarater SD: mean within-model SD across the 4 runs (lower=more stable).  
<sup>b</sup>Agent-total SD: SD of the 8 agent-level totals (higher=better discrimination).  
<sup>c</sup>Overall ceiling rate: proportion of dialogue×dimension cells scored at the maximum (higher=stronger ceiling effect).  
<sup>d</sup>Rank correlation: Spearman ρ of agent rankings with Claude (both  $P<.05$ ).  
<sup>e</sup>Not applicable.

The 3 models agreed broadly on the relative ranking of agents (Spearman  $\rho=0.64$  and  $0.71$  for Claude vs DeepSeek and Qwen) but differed substantially in discrimination. DeepSeek and Qwen showed pronounced leniency, scoring large proportions of dialogues at the ceiling (Figure 2; Qwen: above 90% on medical safety boundary, student engagement elicitation, and assessment and feedback quality; DeepSeek: 60-77% on the same dimensions). This compressed between-agent differences, yielding narrow score ranges (82-95 and 75-100) and low discrimination (SD of agent total scores was 4.0 and 7.3). Claude showed a minimal ceiling effects except on the two 5-point dimensions (student engagement

elicitation, assessment and feedback quality), with the widest range (61-92) and greatest discrimination (SD 11.6).  
Because the analysis aimed to resolve differences in teaching quality among agents, an evaluator able to discriminate between them was required; Claude was therefore used as the primary evaluator, with its agreement against human expert judgment examined separately. The leniency of the other evaluators is itself a relevant finding, indicating that ceiling effects should be anticipated when LLMs serve as evaluators of educational quality.

**Figure 2.** Ceiling effects across the 3 large language model evaluators on the 8 scoring dimensions. Each cell shows the proportion of the 167 dialogues scored at the maximum for that dimension; darker red indicates a stronger ceiling effect (less discrimination). ADC: adaptive difficulty calibration; AFQ: assessment and feedback quality; KCE: knowledge-point coverage efficiency; MKA: medical knowledge accuracy; MSB: medical safety boundary; PGA: pedagogical guidance ability; RPQ: role-play quality; SEE: student engagement elicitation.

|          | MKA | PGA | KCE | RPQ | ADC | MSB | SEE | AFQ |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|
| Claude   | 0%  | 0%  | 4%  | 7%  | 2%  | 1%  | 60% | 59% |
| DeepSeek | 2%  | 2%  | 3%  | 19% | 2%  | 60% | 77% | 77% |
| Qwen     | 9%  | 17% | 22% | 50% | 20% | 93% | 99% | 95% |

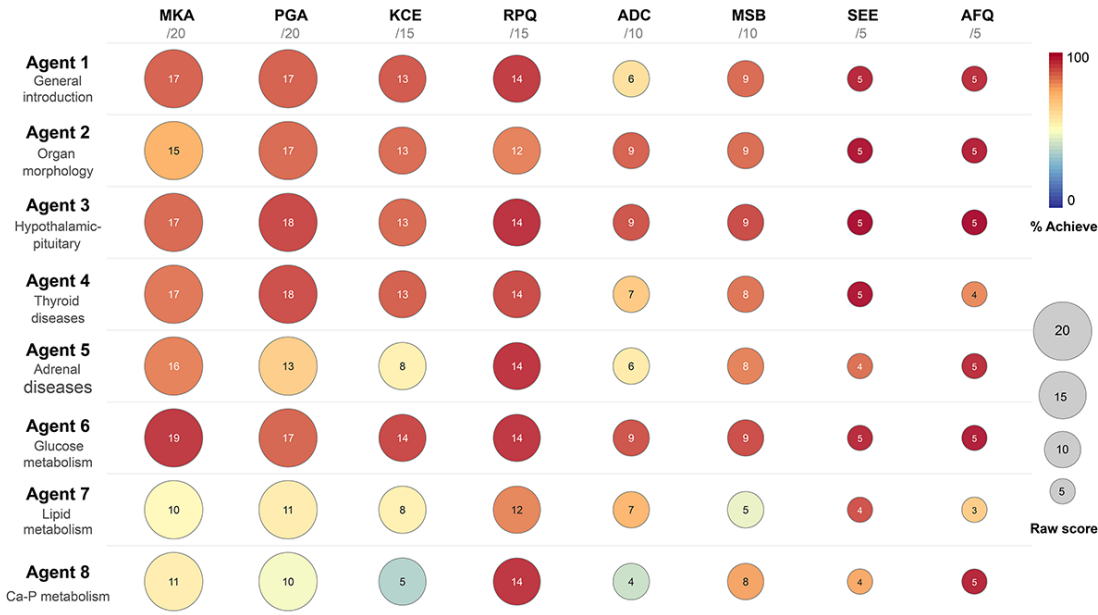


Agents Differed Mainly in Knowledge, Not in Role Performance

Applying the 8-dimension rubric, agent total scores ranged widely (60.6-91.6; Figure 3), with A6 and A3 scoring highest and A7 and A8 lowest. This variation was not uniform across dimensions. It was largest on the 2 knowledge-related dimensions, namely knowledge-point coverage (CV=27.3%) and medical knowledge accuracy (CV=18.5%), indicating

that agents differed most in what and how accurately they taught. In contrast, role-play quality was the most uniform dimension (CV=5.7%), with all agents sustaining their assigned personas comparably well. Adaptive difficulty control was low across nearly all agents, emerging as a shared weakness rather than a source of between-agent variation. Thus, the agents were broadly comparable in role enactment but diverged mainly in the knowledge content they delivered.

**Figure 3.** Eight-dimension rubric profiles of the 8 AI teaching agents. Scores are the mean of 4 independent runs by Claude Opus 4.8. Bubble size indicates the dimension's maximum score, color indicates the achievement rate; and number indicates the raw score. ADC: adaptive difficulty calibration; AFQ: assessment and feedback quality; KCE: knowledge-point coverage efficiency; MKA: medical knowledge accuracy; MSB: medical safety boundary; PGA: pedagogical guidance ability; RPQ: role-play quality; SEE: student engagement elicitation.



Role Paradigms Shaped Dimension Profiles, but Empathy Did Not Ensure Knowledge Coverage

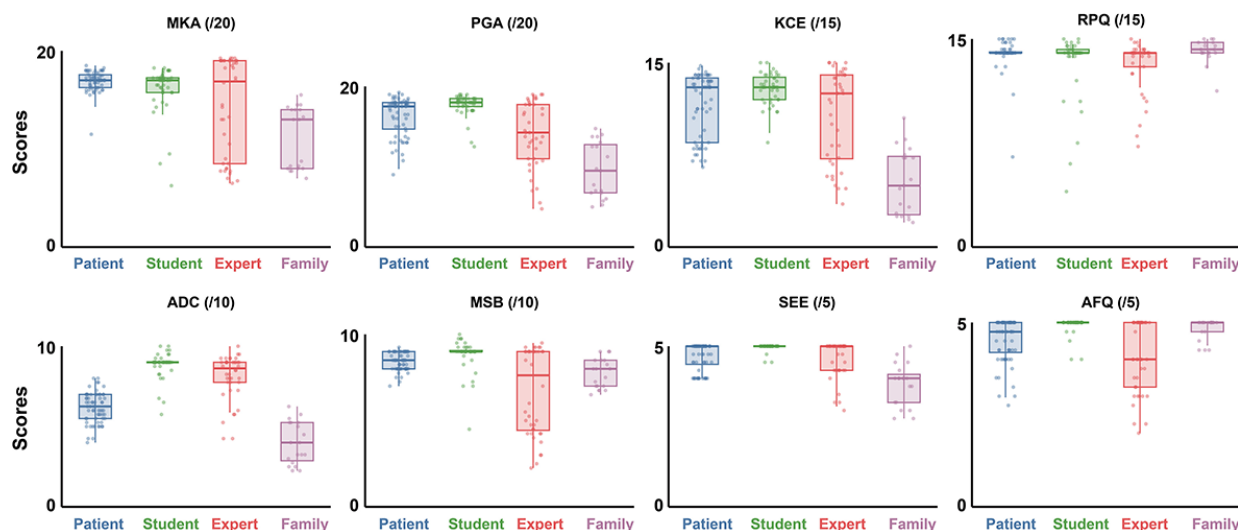
Grouping agents by role paradigm revealed distinct dimension profiles (Figure 4). At the group level, student-type agents showed the strongest pedagogical guidance (group-mean pedagogical guidance ability was 17.7), whereas expert-type agents showed the lowest medical safety

(group-mean medical safety boundary was 6.7). The family-member agent illustrated a notable dissociation between empathy and knowledge: it attained one of the highest role-play scores (role-play quality=14.1) yet the lowest knowledge-point coverage of all agents (knowledge-point coverage efficiency=5.3) and the lowest total score (60.6). Notably, this agent had been deliberately revised after the first 7 agents showed generally limited empathic engagement in student use; only its prompt was adjusted—while all other

settings were held constant—to strengthen the empathic, family-member persona. Despite this targeted enhancement of empathy, it remained weakest in knowledge coverage, indicating that strengthening empathy did not translate into broader knowledge delivery. How to balance affective

engagement and knowledge coverage within a single agent thus remains an open question. Because each role group comprised only 2 to 3 agents, these between-group comparisons are descriptive; formal inferential testing was not performed at the agent level.

**Figure 4.** Eight-dimension score distributions across the 4 role paradigms: patient (A1, A4, A5; n=64 dialogues), student (A2, A3; n=44), expert (A6, A7; n=40), and family member (A8; n=19). Each panel shows 1 dimension (scored out of its maximum, indicated in parentheses). Boxes show median and IQR; points are individual dialogue scores. Between-group differences are descriptive; formal inferential testing was not performed at the agent level. ADC: adaptive difficulty calibration; AFQ: assessment and feedback quality; KCE: knowledge-point coverage efficiency; MKA: medical knowledge accuracy; MSB: medical safety boundary; PGA: pedagogical guidance ability; RPQ: role-play quality; SEE: student engagement elicitation.



### Student Gender Showed No Detectable Effect on Teaching Quality Scores

Male (n=7) and female (n=15) students were compared across all 8 dimensions using Mann-Whitney *U* tests. No significant difference was detected in any dimension or in total score (79.1 vs 80.1;  $P=.82$ ; all dimension-level  $P>.05$ ; Cohen  $d=0.07$  for total score). Given the small, imbalanced sample and the very small observed effect size, this analysis was substantially underpowered; the absence of a detectable difference should therefore not be interpreted as evidence of gender-equitable delivery, as a type II error cannot be excluded.

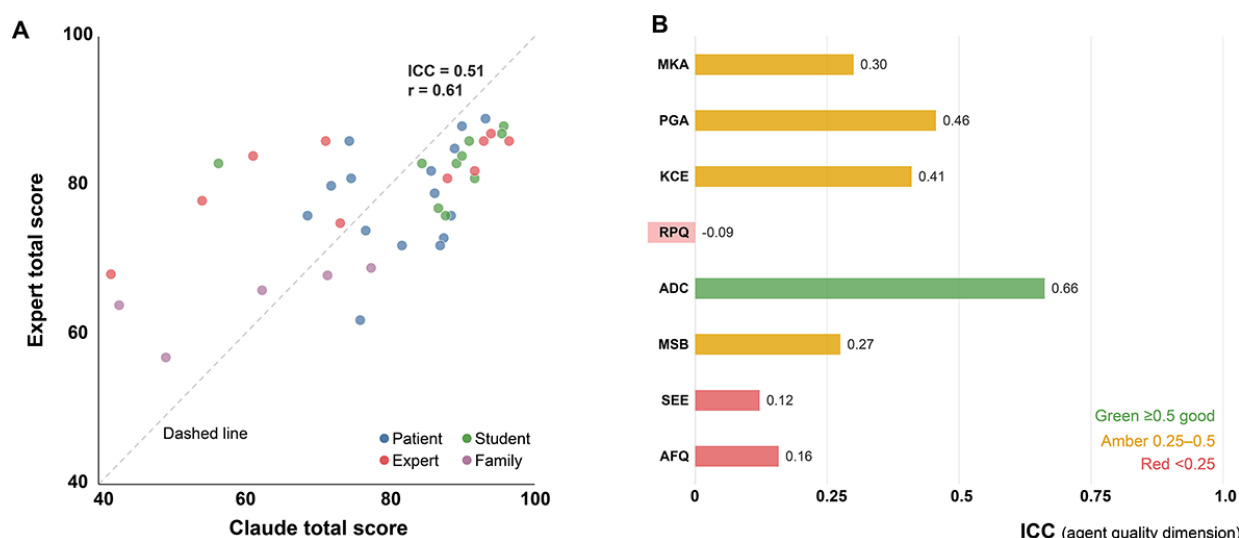
### AI Scores Agreed With Expert Ratings on Overall and Cognitive-Process Dimensions

To validate the rubric-based AI scoring, 40 dialogues spanning all 8 agents and the full quality range were independently rescored by a medical-education expert using an identical rubric, blinded to the AI scores (Figure 5A). AI and expert ratings were closely aligned in magnitude (mean totals 79.1 vs 78.5) and were positively correlated at

the total-score level ( $r=0.61$ ; ICC=0.51). Because the two derive from a highly consistent automated rater and a single subjective human expert, close but not exact agreement was expected; the aim was to confirm convergence on the same construct, not identity.

Agreement was strongest on the cognitive-process dimensions, ranging from moderate for knowledge accuracy to highest for adaptive difficulty control. For role-play quality and the two 5-point dimensions, the ICC was near zero or negative, but this reflects a ceiling effect rather than rater disagreement: both AI and expert rated these dimensions uniformly high (for role-play quality, 13.4 vs 13.3 out of 15), leaving minimal between-agent variance (Figure 5B). Because the ICC indexes relative discrimination, it is driven toward zero when there is little variance to rank, and its low value here does not indicate that the raters disagreed; on the contrary, they consistently converged on high scores [24]. The moderate agreement on knowledge accuracy likewise does not undermine the study's key knowledge-related finding, which rests on a rank-level difference too large to depend on rating precision: A8 scored markedly lowest on knowledge coverage under both AI and expert scoring.

**Figure 5.** Agreement between Claude AI and expert ratings on 40 dialogues, independently scored by a medical-education expert using the same 8-dimension rubric. (A) Total score agreement: expert versus Claude total scores (the dashed line indicates perfect agreement). (B) Per-dimension agreement represented by intraclass correlation coefficient (ICC) per dimension (green  $\geq 0.5$ ; amber 0.25–0.5; red  $< 0.25$ ). ADC: adaptive difficulty calibration; AFQ: assessment and feedback quality; KCE: knowledge-point coverage efficiency; MKA: medical knowledge accuracy; MSB: medical safety boundary; PGA: pedagogical guidance ability; RPQ: role-play quality; SEE: student engagement elicitation.

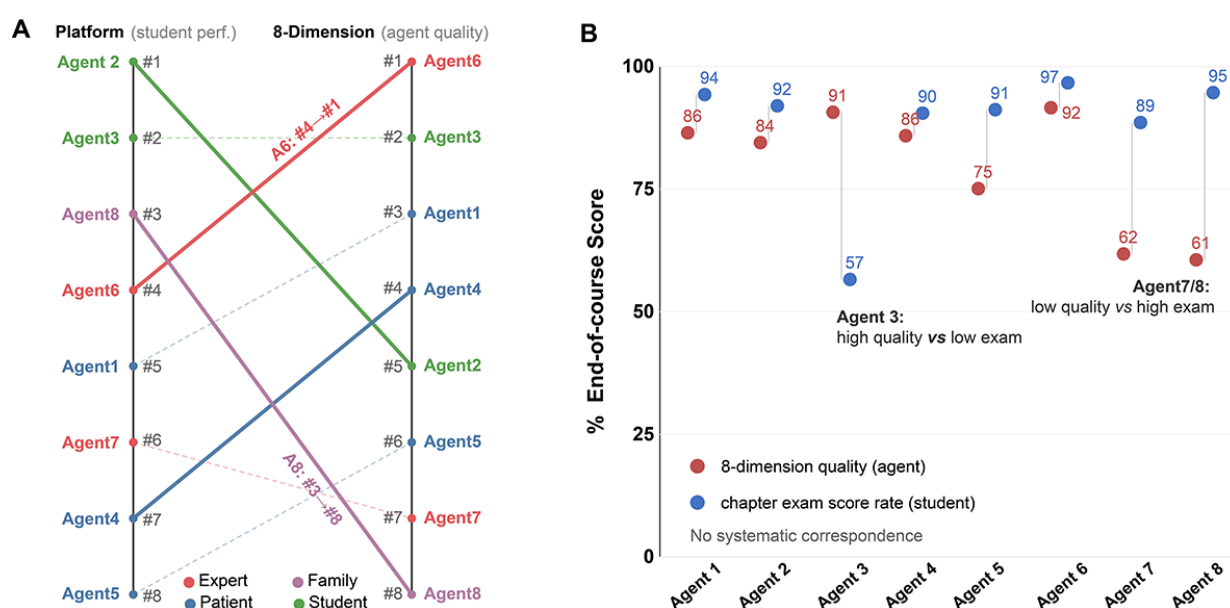


### Platform and Examination Performance Captured Constructs Distinct From Rubric-Based Agent Quality

Agent rankings from platform scores and from the 8-dimension rubric diverged substantially (Figure 6A). Two agents illustrate this most clearly: A8 ranked third by platform score but last by rubric quality, and A6 rose from fourth to first. Rather than indicating that either measure is invalid, this divergence reflects that the two capture different constructs—platform scores index student performance during the interaction, whereas the rubric indexes the teaching quality of the agent itself. The two are therefore complementary rather than interchangeable.

There was a similar lack of association between rubric quality and end-of-course chapter examination performance (Figure 6B): higher agent quality did not correspond to higher chapter examination scores (eg, A3 showed high quality but the lowest examination score rate, whereas A7 and A8 showed low overall quality but high examination scores). This absence of a systematic association should be interpreted cautiously, given the small number of agents, the near-ceiling examination scores, and the lack of individual baseline data; it indicates only that agent teaching quality and this particular examination did not track one another, not that better teaching is ineffective.

**Figure 6.** Platform and examination performance versus rubric-based agent quality. (A) Rank divergence: agent rankings by platform score (student performance) vs by 8-dimension quality (agent quality); 1=best. (B) Eight-dimension agent quality versus end-of-course chapter examination score rate.



## Discussion

This study set out to test whether platform-generated scores, the default metric for AI teaching agents in many educational platforms, adequately capture teaching quality, and to ask what a transparent, multidimensional alternative would reveal. The findings speak less to any single agent than to how the field should evaluate, and design, conversational AI tutors.

Evaluation is not a solved problem, and the default metric may mislead. The most consequential implication is that the metric an educator happens to have at hand can point in the wrong direction. When a platform score and a criterion-based rubric rank the same agents in nearly opposite orders, the choice of metric is not a technical detail but a determinant of which agents get adopted, refined, or discarded. Crucially, this does not make either metric wrong. A platform score that reflects student performance during an interaction and a rubric that reflects the agent's teaching behavior are answering different questions; the problem arises only when one is silently substituted for the other. As AI tutors proliferate faster than the frameworks to evaluate them [11, 14], this argues for treating evaluation itself as a first-class research object, reporting what a metric measures, and pairing convenient platform analytics with transparent, education-specific criteria rather than trusting either alone [17].

Automated evaluation is promising but must be bounded. Using LLMs to score pedagogical quality is attractive because it scales, but our results caution against naive trust [25,26]. Independently developed models converged on the ordering of agents yet differed markedly in leniency, showing that agreement on ranking can coexist with disagreement on absolute standards, and that some models are too lenient to discriminate at all [13,23]. Human calibration therefore remains indispensable [27], and it is not uniform across constructs: automated and expert judgments aligned well on cognitive-process dimensions but poorly on more subjective, affective ones [27]. This suggests a pragmatic division of labor, in which automated scoring is trusted most for the dimensions where it is demonstrably calibrated, while human judgment is reserved for those where meaning is contested [12].

Designing for the whole of teaching, not its most visible parts, and comparing role paradigms, rather than the field's default patient-doctor script [8,9], exposed a design tension that is easy to miss when engagement is the headline goal. An agent optimized to be more empathic became more engaging without becoming more knowledgeable, a reminder that affective and cognitive teaching functions do not automatically move together [17], and that improving the salient, likable qualities of a tutor can leave its instructional substance untouched. Rather than seeking a single best role, this points toward designing for complementarity, orchestrating different

pedagogical strengths across an interaction [16], and toward evaluation that can detect when a more engaging agent is not in fact teaching more. The most uniform finding across every agent, namely persistent weakness in adapting difficulty and providing formative feedback, reinforces the same lesson: the harder and less visible pedagogical skills are precisely those that current prompt-based designs deliver least well [9,28].

Future directions should consider 3 points. First, disentangling role from content requires a factorial design that crosses role paradigms with content domains, so that the effect of how an agent teaches can be separated from what it teaches [5]. Second, establishing educational value, rather than teaching quality alone, requires learning-outcome designs with pre- and postlearning knowledge assessment and individual-level linkage [10,19], ideally using more discriminating assessments than a near-ceiling course examination. Third, the persistent deficits in adaptive difficulty and feedback motivate designs that move beyond static prompting toward dynamic learner modeling and memory-augmented tutoring [8,16], evaluated against the dimensions on which current agents underperform. More broadly, the 8-dimension evaluation framework introduced here offers a transferable reference for the development of future teaching agents, and of medical teaching agents in particular, by making explicit the pedagogical dimensions against which such agents should be designed and assessed.

Undeniably, this is a small, single-institution study in one discipline, and its design confounds role paradigm with content domain, each agent being one role and one chapter, so role-related observations are descriptive rather than causal and no independent content-domain effect can be claimed. Because role groups contain few agents, group comparisons are descriptive and were not tested inferentially at the agent level. Expert calibration rested on a single expert, precluding expert-to-expert reliability, and was weaker for subjective dimensions. The gender comparison was underpowered and cannot support claims of equitable delivery, the rubric weights were team-assigned without formal stakeholder input, and teaching quality rather than learning gain was the outcome assessed. These constraints frame the study as an exploratory, framework-building effort whose specific estimates await confirmation in larger, factorial, outcome-linked designs.

Platform scores answer a different question than teaching quality and should complement rather than replace criterion-based evaluation. By making the dimensions of teaching explicit, and by validating those judgments against expert and cross-model agreement, a multidimensional rubric reveals distinctions that convenient metrics miss, including where more engaging agents are not more instructive and where all current agents falter. Beyond any single result, the study suggests that how we evaluate AI teaching agents is inseparable from how we improve them.

---

## Acknowledgments

The authors declare that generative AI was used in the preparation of this manuscript. Claude Opus 4.8 was employed to enhance the language and readability of the draft. The final version was thoroughly reviewed, revised, and approved by the authors, who take full responsibility for the content.

---

## Funding

This work was supported by the following grants: the 2026 Educational Science Planning and Teaching Reform Projects of Guangzhou Medical University (2026JXGG02) and the 2025 Guangdong Provincial Graduate Education Innovation Program (2025JGXM\_139).

---

## Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

---

## Conflicts of Interest

None declared.

---

## Multimedia Appendix 1

Complete configuration of the 8 AI teaching agents.

[\[PDF File \(Adobe File\), 199 KB-Multimedia Appendix 1\]](#)

---

## Multimedia Appendix 2

The complete 8-dimension scoring prompt used by all evaluators.

[\[PDF File \(Adobe File\), 124 KB-Multimedia Appendix 2\]](#)

---

## References

1. Long C, Tsai K, Napolitano JD, Khawaja RA, Leung AM. The current status of preclinical endocrine education in U.S. medical schools. *Endocr Pract.* Aug 2022;28(8):744-748. [doi: [10.1016/j.eprac.2022.04.008](#)] [Medline: [35452814](#)]
2. Hu MH, Han YT, Lyu J, Li LM. Advantages and challenges of applying artificial intelligence in medical education. *Zhonghua Liu Xing Bing Xue Za Zhi.* Feb 10, 2026;47(2):200-206. [doi: [10.3760/cma.j.cn112338-20250512-00309](#)] [Medline: [41765658](#)]
3. Jiang Y, Fu X, Wang J, et al. Enhancing medical education with chatbots: a randomized controlled trial on standardized patients for colorectal cancer. *BMC Med Educ.* Dec 20, 2024;24(1):1511. [doi: [10.1186/s12909-024-06530-8](#)] [Medline: [39707245](#)]
4. Öncü S, Torun F, Ülkü HH. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. *BMC Med Educ.* Feb 20, 2025;25(1):278. [doi: [10.1186/s12909-025-06877-6](#)] [Medline: [39979969](#)]
5. Xiao J, Fu X. Is the use of standardized patients more effective than role-playing in medical education? A meta-analysis. *Front Med (Lausanne).* Jun 18, 2025;12:1601116. [doi: [10.3389/fmed.2025.1601116](#)] [Medline: [40606465](#)]
6. Huang LJ, Huang HC, Chuang CL, et al. Role-play of real patients improves the clinical performance of medical students. *J Chin Med Assoc.* Feb 1, 2021;84(2):183-190. [doi: [10.1097/JCMA.0000000000000431](#)] [Medline: [32925298](#)]
7. Flanagan OL, Cummings KM. Standardized patients in medical education: a review of the literature. *Cureus.* Jul 17, 2023;15(7):e42027. [doi: [10.7759/cureus.42027](#)] [Medline: [37593270](#)]
8. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *Commun Med (Lond).* Dec 19, 2025;6(1):27. [doi: [10.1038/s43856-025-01283-x](#)] [Medline: [41420084](#)]
9. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Med Educ.* Nov 28, 2024;24(1):1391. [doi: [10.1186/s12909-024-06399-7](#)] [Medline: [39609823](#)]
10. Peng J, Zhang H, Tu X, et al. Effectiveness of AI-assisted medical education for Chinese undergraduate medical students: a meta-analysis. *BMC Med Educ.* Aug 27, 2025;25(1):1207. [doi: [10.1186/s12909-025-07770-y](#)] [Medline: [40866973](#)]
11. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - scoping literature review. *Med Sci Educ.* Nov 13, 2024;35(1):555-567. [doi: [10.1007/s40670-024-02206-6](#)] [Medline: [40144083](#)]

12. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare (Basel)*. Mar 10, 2025;13(6):603. [doi: [10.3390/healthcare13060603](https://doi.org/10.3390/healthcare13060603)] [Medline: [40150453](https://pubmed.ncbi.nlm.nih.gov/40150453/)]
13. Elhilali A, Ngo ASH, Reichenpfader D, Denecke K. Large language model-based patient simulation to foster communication skills in health care professionals: user-centered development and usability study. *JMIR Med Educ*. Dec 12, 2025;11:e81271. [doi: [10.2196/81271](https://doi.org/10.2196/81271)] [Medline: [41385781](https://pubmed.ncbi.nlm.nih.gov/41385781/)]
14. Ding H, Simmich J, Vaezipour A, Andrews N, Russell T. Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review. *J Am Med Inform Assoc*. Feb 16, 2024;31(3):746-761. [doi: [10.1093/jamia/ocad222](https://doi.org/10.1093/jamia/ocad222)] [Medline: [38070173](https://pubmed.ncbi.nlm.nih.gov/38070173/)]
15. Gu J, Jiang X, Shi Z, et al. A survey on LLM-as-a-judge. *Innovation (Camb)*. Jun 1, 2026;7(6):101253. [doi: [10.1016/j.xinn.2025.101253](https://doi.org/10.1016/j.xinn.2025.101253)] [Medline: [42254963](https://pubmed.ncbi.nlm.nih.gov/42254963/)]
16. Gorenshtein A, Omar M, Glicksberg BS, Nadkarni GN, Klang E. AI agents in clinical medicine: a systematic review. *medRxiv*. Aug 26, 2025:2025.08.22.25334232. [doi: [10.1101/2025.08.22.25334232](https://doi.org/10.1101/2025.08.22.25334232)] [Medline: [40909853](https://pubmed.ncbi.nlm.nih.gov/40909853/)]
17. Xu X, Cao X, Wu Q. Impact of educational agents on student's learning outcomes: a meta-analysis. *Front Psychol*. Feb 24, 2026;17:1707196. [doi: [10.3389/fpsyg.2026.1707196](https://doi.org/10.3389/fpsyg.2026.1707196)] [Medline: [41815245](https://pubmed.ncbi.nlm.nih.gov/41815245/)]
18. Alshammri F, Abdulshakour M, Chen L, et al. Pediatric endocrinology education among trainees: a scoping review. *Clin Teach*. Feb 2025;22(1):e70011. [doi: [10.1111/tct.70011](https://doi.org/10.1111/tct.70011)] [Medline: [39743233](https://pubmed.ncbi.nlm.nih.gov/39743233/)]
19. Myers SE, Bender NR, Seidel MA, Weinstock RS. Diabetes SPECIAL (Students Providing Education on Chronic Illness and Lifestyle): a novel preclinical medical student elective. *Perspect Med Educ*. Oct 2021;10(5):312-315. [doi: [10.1007/s40037-020-00641-w](https://doi.org/10.1007/s40037-020-00641-w)] [Medline: [33349906](https://pubmed.ncbi.nlm.nih.gov/33349906/)]
20. Daemicke AK, Galt NJ, Samonds KE, Bergan-Roller HE. Challenging endocrinology students with a critical-thinking workbook. *Adv Physiol Educ*. Mar 1, 2020;44(1):72-79. [doi: [10.1152/advan.00101.2019](https://doi.org/10.1152/advan.00101.2019)] [Medline: [32057267](https://pubmed.ncbi.nlm.nih.gov/32057267/)]
21. Hodge TJ, Potter LJ, Helsby CJ. Scaffolding simulation activities for medical students learning cardiopulmonary assessment: a retrospective study. *Cureus*. Apr 10, 2025;17(4):e82013. [doi: [10.7759/cureus.82013](https://doi.org/10.7759/cureus.82013)] [Medline: [40351920](https://pubmed.ncbi.nlm.nih.gov/40351920/)]
22. Zhang H, Qu L, Bai H, et al. Beyond accuracy: evaluating the reliability of large language models for medical assessment. *Front Artif Intell*. Jul 8, 2026;9:1832829. [doi: [10.3389/frai.2026.1832829](https://doi.org/10.3389/frai.2026.1832829)] [Medline: [42488377](https://pubmed.ncbi.nlm.nih.gov/42488377/)]
23. Thomas LDW, Romasanta AKG, Pujol Priego L. Jagged competencies: measuring the reliability of generative AI in academic research. *J Bus Res*. Jan 2026;203:115804. [doi: [10.1016/j.jbusres.2025.115804](https://doi.org/10.1016/j.jbusres.2025.115804)]
24. Mehta S, Bastero-Caballero RF, Sun Y, et al. Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. *Stat Med*. Aug 15, 2018;37(18):2734-2752. [doi: [10.1002/sim.7679](https://doi.org/10.1002/sim.7679)] [Medline: [29707825](https://pubmed.ncbi.nlm.nih.gov/29707825/)]
25. Croxford E, Gao Y, First E, et al. Evaluating clinical AI summaries with large language models as judges. *NPJ Digit Med*. Nov 5, 2025;8(1):640. [doi: [10.1038/s41746-025-02005-2](https://doi.org/10.1038/s41746-025-02005-2)] [Medline: [41193667](https://pubmed.ncbi.nlm.nih.gov/41193667/)]
26. Zhai G, Bar M, Cowan AJ, et al. AI for evidence-based treatment recommendation in oncology: a blinded evaluation of large language models and agentic workflows. *Front Artif Intell*. Dec 9, 2025;8:1683322. [doi: [10.3389/frai.2025.1683322](https://doi.org/10.3389/frai.2025.1683322)] [Medline: [41446897](https://pubmed.ncbi.nlm.nih.gov/41446897/)]
27. Teckwani SH, Wong AHP, Luke NV, Low ICC. Accuracy and reliability of large language models in assessing learning outcomes achievement across cognitive domains. *Adv Physiol Educ*. Dec 1, 2024;48(4):904-914. [doi: [10.1152/advan.00137.2024](https://doi.org/10.1152/advan.00137.2024)] [Medline: [39514712](https://pubmed.ncbi.nlm.nih.gov/39514712/)]
28. Bogenschutz K, Demeter J, Knoderer CA. Artificial intelligence in medical education: promoting active learning with a customized chatbot tool. *J Physician Assist Educ*. Jun 1, 2026;37(2):216-220. [doi: [10.1097/JPA.0000000000000742](https://doi.org/10.1097/JPA.0000000000000742)] [Medline: [41603621](https://pubmed.ncbi.nlm.nih.gov/41603621/)]

## Abbreviations

**CV:** coefficient of variation  
**ICC:** intraclass correlation coefficient  
**LLM:** large language model

*Edited by Waseem Jerjes; peer-reviewed by Song Liu, Yunchu Yang; submitted 01.Apr.2026; final revised version received 10.Jul.2026; accepted 16.Jul.2026; published 03.Sep.2026*

### *Please cite as:*

Zhang H, Qu L, Zheng J, Xiong Y, Bai H, Ji R, Liu G, Chen W, Cheng Z, Chen Y, Yang CT  
What Platform Scores Miss: Multidimensional Evaluation of AI Teaching Agents in Medical Education

*JMIR Med Educ* 2026;12:e96819

URL: <https://mededu.jmir.org/2026/1/e96819>

doi: [10.2196/96819](https://doi.org/10.2196/96819)

© Hui Zhang, Lihui Qu, Jianmin Zheng, Yi Xiong, Hongbo Bai, Ruiying Ji, Guohui Liu, Wanling Chen, Zirui Cheng, Youbang Chen, Chun-tao Yang. Originally published in JMIR Medical Education (<https://mededu.jmir.org>), 03.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on <https://mededu.jmir.org/>, as well as this copyright and license information must be included.