Comparing AI- and Human-Based Assessments of Medical Interview Transcripts Using a Generative AI Simulated Patient System.

Hiromizu Takahashi, Kiyoshi Shikino, Takeshi Kondo, Yuji Yamada, Yoshitaka Tomoda, Minoru Kishi, Yuki Aiyama, Sho Nagai, Akiko Enomoto, Yoshinori Tokushima, Takahiro Shinohara, Fumiaki Sano, Takeshi Matsuura, Rikiya Watanabe, Toshio Naito

Submitted to: JMIR Medical Education
on: August 21, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Comparing AI- and Human-Based Assessments of Medical Interview Transcripts Using a Generative AI Simulated Patient System.

Hiromizu Takahashi[1] MD, PhD; Kiyoshi Shikino[2] MD, MHPE, PhD; Takeshi Kondo[3] MD, MHPE, PhD; Yuji Yamada[4] MD, PhD; Yoshitaka Tomoda[5] MD; Minoru Kishi[6] MD; Yuki Aiyama[7] MD; Sho Nagai[8]; Akiko Enomoto[8]; Yoshinori Tokushima[9] MD, PhD; Takahiro Shinohara[10] MD; Fumiaki Sano[1] MD, PhD; Takeshi Matsuura[11] MD; Rikiya Watanabe[12] MD; Toshio Naito[1] MD, PhD, MBA

[1]Department of General Medicine Faculty of Medicine Juntendo University Tokyo JP
[2]Department of Community-Oriented Medical Education Graduate School of Medicine Chiba University Chiba JP
[3] Center for Postgraduate Clinical Training and Career Development Nagoya University Hospital Nagoya JP
[4]Brookdale Department of Geriatrics and Palliative Medicine Icahn School of Medicine Mount Sinai New York US
[5] Department of General Internal Medicine Itabashi Chuo Medical Center Tokyo JP
[6] Nishiwaki Municipal Hospital Hyogo JP
[7] Tenri Hospital Nara JP
[8]Department of Nursing School of Nursing University of Human Environments Aichi JP
[9] Department of General Medicine Saga University Hospital Saga JP
[10]Department of General Medicine Graduate School of Medical and Dental Sciences, Institute of Science Tokyo Tokyo JP
[11] Department of General Medicine Bibai City Hospital Hokkaido JP
[12] Department of General Internal Medicine Kita-Harima Medical Center Hyogo JP

Corresponding Author:
Hiromizu Takahashi MD, PhD
Department of General Medicine
Faculty of Medicine
Juntendo University
3-1-3 Hongo, Bunkyo Tokyo, 1130033 Japan
Tokyo
JP

Abstract

Background: Generative AI is increasingly used in medical education, including the use of AI-based virtual patients to improve interview skills. However, it remains unclear how much AI-based assessment (ABA) differs from those of Human-based assessment (HBA).

Objective: This study aimed to compare the quality of clinical interview assessments generated by an ABA using a virtual patient with those provided by a HBA conducted by clinical instructors. Additionally, it evaluated whether the use of AI could lead to a measurable reduction in evaluation time, and examined the level of agreement across participants with differing levels of clinical experience.

Methods: A standardized leg-weakness case was implemented in an AI based virtual patient. Seven participants—two medical students, three resident physicians, and two attending physicians—each conducted an interview, and transcripts were scored with the Master Interview Rating Scale (MIRS; 25?items, 0–5?scale; total?0–125).
Two evaluation strategies were compared. (1) ChatGPT o1-Pro scored each transcript five times with different random seeds to assess case specificity; total runtime for the five scores was automatically logged. (2) Five blinded clinical instructors , after a preparatory webinar reviewing the rubric and practicing on sample transcripts, each rated every transcript once and recorded clock time per rating. Because the five AI outputs are replicates of the same algorithm, intraclass correlation coefficients (ICC) were used to quantify repeatability rather than inter rater reliability. For human raters, we calculated ICC (2,1).
Mean scores from both methods were compared, and agreement was quantified with Pearson’s r, Lin’s concordance correlation coefficient (?c), Bland–Altman limits of agreement (LoA), internal consistency (Cronbach’s??), and ICC. Time efficiency was expressed as mean minutes per transcript and the relative percentage reduction achieved by AI scoring.

Results: Mean interview scores were similar for ABA and HBA (52.1?±?6.9 vs?53.7?±?6.8). Agreement was strong (r?=?0.92;

?c?=?0.92) with minimal bias (+0.4?points; LoA??4.9?to?+5.7). ABA showed higher internal consistency (??=?0.936 vs 0.863) and greater inter rater reliability (ICC?=?0.77 vs?0.38). The coefficient of variation for ABA scores was roughly half that of HBA scores (6.6?% vs?13.9?%). In addition, ChatGPT completed each five run analysis in 4.3?±?1.7?minutes compared with 10.3?±?3.3?minutes for physicians, representing a 58?% reduction in assessment time.

Conclusions: ABA scores that closely matched HBA scores while demonstrating superior consistency and reliability. In the setting of virtual clinical interview transcripts, these preliminary findings suggest that ABA shows potential as a valid, rapid, and scalable alternative to HBA. When applied strategically, it could potentially furnish timely formative feedback, quantify efficiency gains, and reduce faculty workload without compromising assessment quality. Further research is needed to determine whether this can be achieved without compromising assessment quality.

(JMIR Preprints 21/08/2025:81673)
DOI: https://doi.org/10.2196/preprints.81673

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original Paper

Comparing AI- and Human-Based Assessments of Medical Interview Transcripts Using a Generative AI Simulated Patient System

Takahashi Hiromizu, MD, PhD, Shikino Kiyoshi, MD, MHPE, PhD, Kondo Takeshi, MD, MHPE, PhD, Yamada Yuji, MD, Yoshitaka Tomoda, MD, Minoru Kishi, MD, Yuki Aiyama, MD, Akiko Enomoto, Sho Nagai, Yoshinori Tokushima, MD, Takeshi Matsuura, MD, Fumiaki Sano, MD, PhD, Yoshihiro Shinohara, MD, Rikiya Watanabe, MD, Naito Toshio, MD, PhD

Corresponding Author:

Takahashi Hiromizu, MD, PhD,
Juntendo University Faculty of Medicine Department of General Medicine
3-1-3 Hongo, Bunkyo
Tokyo, 1130033
Japan
Phone: 81 3 3813 3111
Email: hrtakaha@juntendo.ac.j

Title

Comparing AI- and Human-Based Assessments of Medical Interview Transcripts Using a Generative AI Simulated Patient System

Abstract

Background: Generative artificial intelligence is increasingly used in medical education, including the use of AI-based virtual patients to improve interview skills. However, how much AI-based assessment (ABA) differs from human-based assessment (HBA) remains unclear.

Objective: This study aimed to compare the quality of clinical interview assessments generated by an ABA (ChatGPT-o1 Pro (ABA-o1) and ChatGPT-5 Pro(ABA-5)) with those provided by an HBA conducted by clinical instructors in an AI-based virtual patient setting. In addition, whether the use of AI could lead to a measurable reduction in evaluation time was evaluated, and the level of agreement across participants with differing levels of clinical experience was examined.

Methods: A standardized case of leg weakness was implemented in an AI-based virtual patient. Seven participants, including two medical students, three resident physicians, and two attending physicians, each conducted an interview with the AI-patient, and the transcripts were scored with the Master Interview Rating Scale (25 items, 0–5 scale; total 0–125).

Three evaluation strategies were compared. In the first approach, ChatGPT-o1 Pro and ChatGPT-5 Pro were used to score each transcript five times with different random seeds to test case specificity. The processing time required to generate all five scores was automatically logged, measuring the full runtime from start to finish. In the second approach, five blinded clinical instructors independently rated each transcript once using the same rubric. Before conducting their evaluations, the instructors participated in a preparatory webinar during which they reviewed the rubric and practiced using sample transcripts to standardize their scoring approach. Following the study interviews, each instructor recorded the clock time required for each rating session. Because the five AI outputs were replicates of the same algorithm, intraclass correlation coefficients (ICCs) were used to quantify repeatability rather than inter-rater reliability. For human raters, ICC (2,1) was calculated.

Mean scores from both methods were compared, and agreement was quantified with Pearson's *r*, Lin's concordance correlation coefficient (CCC), Bland-Altman limit of agreement (LoA), internal consistency (Cronbach's α), and ICC. Time-efficiency was expressed as mean minutes per transcript and the relative percentage reduction achieved by AI scoring.

Results: Mean interview scores for ABA were similar to HBA (ABA-o1=52.1 ± 6.9, ABA-5=53.2 ± 6.8, HBA=53.7 ± 6.8). Agreement was strong (r: ABA-o1=0.90, ABA-5=0.87; CCC: ABA-o1=0.88, ABA-5=0.86) with minimal bias (mean bias: ABA-o1=0.4, ABA-o5=1.54; LoA: ABA-o1= −4.9 to +5.7, ABA-5=−8.60 to 11.68.). ABA showed higher internal consistency (Cronbach's α: ABA-o1=0.81, ABA-5=0.83, HBA=0.80) and greater inter-rater reliability (ICC (3.1): ABA-o1=0.77, ABA-5=0.82; ICC (2.1): HBA=0.38). The coefficient of variation for ABA scores was roughly half that of HBA scores (ABA-o1=6.6%, ABA-5=6.6%, HBA=13.9%). In addition, ChatGPT completed each five-run analysis in 4 m 19 s (ABA-o1) and 3 m 20 s (ABA-5) compared with 10 m 16 s for physicians, representing a 58% and 67.6% reduction in assessment time.

Conclusions: ABA-o1 and ABA-5 produced scores that closely matched HBA while demonstrating superior internal consistency and inter-rater reliability. In the setting of virtual clinical interview transcripts, these preliminary findings suggest that ABA-o1 and ABA-5 show potential as a valid, rapid, and scalable alternative to HBA, reducing per-encounter assessment time by 58% (ABA-o1) and 67.6% (ABA-5), respectively. When applied strategically, they could potentially furnish timely formative feedback, quantify efficiency gains, and reduce faculty workload without compromising assessment quality. Further research is needed to determine whether this can be achieved without compromising assessment quality.

Keywords: Medical Education, Artificial Intelligence, Virtual Patient, Clinical Interview, ChatGPT, Simulation-Based Learning, Assessment

Introduction

Background

Effective clinical interviewing is essential for making correct diagnoses and building strong relationships with patients [1]. Traditionally, students learn these skills through supervised practice with real or standardized patients and feedback from faculty [1]. However, this apprenticeship-style approach is time-intensive and limits opportunities for deliberate practice [2].

The assessment component itself also consumes substantial faculty and resident physician time. In competency-based medical education, faculty complete numerous workplace-based assessment forms; one Canadian study found a mean of 3 min 6 s per Entrustable Professional Activity form, adding approximately 18 min of extra documentation time for each staff member every four-week block [3]. Multi-program qualitative work further confirms that the cumulative “assessment burden” is now viewed as a major threat to sustainability, prompting programs to redesign processes to reduce administrative load [4].

Recently, generative artificial intelligence (AI) using large language models (LLMs) has enabled the creation of AI-based virtual patients (AI-patients) that both converse with learners and automatically evaluate performance [2,5]. Empirical studies have shown promising results for AI assessment in free-text clinical documentation [6], script concordance testing [7], and Objective Structured Clinical Examination (OSCE) history-taking stations [8]. Many of these systems use validated rubrics such as the Master Interview Rating Scale (MIRS) to structure feedback [9]. Nevertheless, the reliability and validity of AI-generated ratings remain under-studied; establishing concordance with expert evaluations is therefore a prerequisite for educational or licensure use.

Objective

This study compared AI-based assessment (ABA) scores of clinical interview performance using ChatGPT-o1 Pro (ABA-o1) and ChatGPT-5 Pro (ABA-5) with human-based assessment (HBA) scores. We hypothesized that ABA scores and HBA scores demonstrate strong concordance and that ABA scoring serves as a substitute for HBA scoring. We also hypothesized that AI completes evaluations more rapidly, reducing the assessment burden on clinicians. A secondary aim was to evaluate agreement across participants with differing clinical experience, and to evaluate whether the use of AI could lead to a measurable reduction in evaluation time, thereby contributing to overall efficiency in assessment processes.

Methods

Study Design and Setting

A cross-sectional validation study was conducted. The study involved two medical students (MSs), three resident physicians (RPs), and two attending physicians (APs) who participated in standardized clinical scenarios.

Virtual-Patient Scenario

A 27-year-old man presenting with progressive bilateral leg weakness, particularly proximal, was scripted based on a published case of thyrotoxic periodic paralysis. The scenario, created by a general internal medicine specialist with extensive educational experience, drawing directly on prior literature, included relevant clinical history (e.g., recent myalgias, tremors, diarrhea, insomnia), red flag cues (e.g., acute onset, muscle weakness, hypokalemia), and psychosocial factors (e.g., recent immigration, use of herbal supplements). The case represented a classic presentation of thyrotoxic periodic paralysis caused by hyperthyroidism. The patient was implemented as an AI-simulated character using ChatGPT’s Custom GPTs.

Participants

The participants were recruited through convenience sampling, complemented by snowball sampling.

- Medical students: a third-year student and a fifth-year medical student.
- Resident physicians: three postgraduate year-1 residents.
- Attending physicians: two board-certified physicians in internal medicine or general internal medicine in Japan, each with ≥5 years of clinical teaching experience.

Each participant conducted a history-taking encounter by speaking with an AI-patient. All conversations were recorded and transcribed verbatim.

Scoring Instrument

The MIRS from the University of Tennessee was originally designed to assess 27 items. In this study, 25 of these items were evaluated based on the available conversational recordings. Each item was rated on a 0–5 scale (total possible score: 0–125), covering domains such as information gathering, organization, empathy, and patient-centered communication. The excluded items were nonverbal behavior and pace and flow of the interview, which require audiovisual input to evaluate.

Assessment Methods

Main outcome: Comparison of MIRS scores

1. AI-based assessment (ABA-o1): Each transcript was submitted separately to ChatGPT-o1 Pro with a base prompt directing it to rate the encounter using MIRS and justify each score. This process was repeated five times per transcript, and item-level and total scores were averaged across runs.
2. AI-based assessment (ABA-5): Using the same base prompt, the seven transcripts were scored in two batch submissions rather than individually: Run 1 included MS1, MS2, RP1, and RP2, and Run 2 included RP3, AP1, and AP2. In each batch, the prompt explicitly stated that it contained four interview transcripts (Run 1) or three interview transcripts (Run 2). For each participant within a batch, item-level and total MIRS scores were extracted from the model's output.
3. Human-based assessment: Five blinded clinical instructors independently rated each transcript using the same MIRS rubric. All assessors were board-certified in general internal medicine or general practice in Japan, actively involved in medical education, and co-authors of this study (Drs. K.T, Y.S, R.W, T.M, and F.S). Item-level and total scores were averaged across the five raters.

Secondary outcome: Comparison of assessment time

1. Physician scoring time (HBA): A stopwatch measured the time from transcript review to completion of scoring.
2. AI scoring time (ABA-o1): The elapsed time was automatically recorded for each of the seven individual submissions from prompt submission to receipt of the complete output.
3. AI scoring time (ABA-5): The elapsed time was automatically recorded for each of the two batch submissions, from prompt submission to receipt of the complete output.

   For all three methods, mean assessment time (SD) was calculated, and absolute and relative time savings vs HBA were reported.

Statistical Analysis

All analyses used R version 4.3.1 (R Foundation for Statistical Computing, Vienna, Austria). Descriptive statistics (mean ± SD, coefficient of variation (CV)) summarized scores. Agreement was assessed with:

- Pearson's correlation coefficient (r) for linear associations;
- Lin's concordance correlation (CCC) for both correlations and bias; and
- Bland-Altman analysis for bias and LoA.

Reliability metrics included Cronbach's α for internal consistency, and intraclass correlation coefficients (ICCs) were calculated to quantify (i) repeatability across the five independent ChatGPT-o1Pro and ChatGPT-5 Pro runs and (ii) inter-rater reliability across the five physician raters. A two-sided α < .05 denoted significance.

Ethical Considerations

Ethical approval was obtained from the Juntendo University Institutional Review Board (Approval No. E24-0314-U02). All participants provided written, informed consent prior to participation.

Results

Participant Scores

Table 1 summarizes the interview scores produced by ABA-o1, ABA-5, and HBA. Across all seven participants, group-level means (mean ± SD across participants) were HBA = 53.7 ± 6.8, ABA-5 = 53.2 ± 9.2, and ABA-o1 = 52.1 ± 6.9. Within-participant variability (mean CV%) was similar for the two automated methods (ABA-o1 = 6.6%, ABA-5 = 6.6%) and higher for HBA (13.9%). Individual-level differences were generally small, though notable divergences appeared for RP2 when comparing HBA vs. ABA-o1 (46.8 vs. 53.4; Δ = 6.6) and for AP2 when comparing ABA-5 vs. HBA (67.8 vs. 58.8; Δ = 9.0) and ABA-5 vs. ABA-o1 (67.8 vs. 55.6; Δ = 12.2).

Participant	HBA	ABA-o1	ABA-5
MS1	48.0 ± 8.9 (18.5%)	46.4 ± 2.4 (5.2%)	46.0 ± 1.9 (4.1%)
MS2	65.0 ± 9.7 (15.0%)	63.6 ± 5.1 (8.1%)	64.6 ± 4.2 (6.5%)
RP1	47.0 ± 2.9 (6.2%)	46.8 ± 2.9 (6.1%)	50.0 ± 2.6 (5.3%)
RP2	53.4 ± 7.2 (13.4%)	46.8 ± 3.3 (7.2%)	51.0 ± 7.1 (14.0%)
RP3	47.2 ± 3.6 (7.6%)	47.6 ± 2.7 (5.7%)	44.0 ± 1.0 (2.3%)
AP1	56.4 ± 9.4 (16.7%)	58.0 ± 5.4 (9.3%)	49.2 ± 2.6 (5.3%)
AP2	58.8 ± 11.7 (19.8%)	55.6 ± 2.7 (4.9%)	67.8 ± 6.2 (9.1%)
All (n = 7)	**53.7 ± 6.8 (13.9%)**	**52.1 ± 6.9 (6.6%)**	**53.2 ± 9.2 (6.6%)**

Table 1 presents mean ± SD (CV %) scores by method and participant. *Abbreviations:* ABA-o1 = AI-based assessment using ChatGPT-o1 Pro; ABA-5 = AI-based assessment using ChatGPT-5 Pro; HBA = human-based assessment.

Agreement and Reliability Across ABA-o1, ABA-5, and HBA

Agreement and reliability were evaluated across the three rating methods (ABA-o1, ABA-5, and HBA). Pairwise concordance with HBA was high for both AI variants: ABA-o1 vs HBA showed a Pearson's correlation coefficient (*r*) of 0.90 (95% CI 0.78–0.96) and CCC of 0.88; ABA-5 vs HBA showed *r*=0.87 (95% CI 0.72–0.94) and CCC=0.86. Concordance between the two AI pipelines was the highest (ABA-o1 vs ABA-5: *r*=0.98, 95% CI 0.95–0.99; CCC=0.99), indicating near-interchangeability of the AI variants (Table 2). Internal consistency followed the same pattern: Cronbach's α was 0.81 for ABA-o1, 0.83 for ABA-5, and 0.80 for HBA.

All correlations were significant (two-sided *P*<.001). Bland-Altman analyses comparing each ABA with HBA showed small positive mean biases (ABA-o1 vs HBA: +0.43; ABA-5 vs HBA: +1.54), with 95% limits of agreement (LoAs) of −4.87 to 5.72 and −8.60 to 11.68, respectively; no proportional bias was observed in either comparison (Figure 1A,B).

Repeatability was assessed using the ICC. ABA-o1 showed substantial repeatability across five independent runs (ICC (3,1) = 0.77; ICC (3,5) = 0.94), and ABA-5 likewise showed substantial repeatability (ICC (3,1) = 0.82; ICC (3,5) = 0.96). In contrast, interrater reliability among the five HBA physician raters was only fair on single measures (ICC (2,1) = 0.38) and improved when averaging five raters (ICC (2,5) = 0.75). Overall, both AI-based approaches yielded more stable ratings across repeated evaluations than HBA, with ABA-5 slightly more stable than ABA-o1.

Comparison	n	Pearson's r (95% CI)	Lin's CCC
ABA-o1 vs HBA	25	0.90 (0.78–0.96)	0.88
ABA-5 vs HBA	25	0.87 (0.72–0.94)	0.86
ABA-o1 vs ABA-5	25	0.98 (0.95–0.99)	0.98

Table 2. Correlation, concordance, and internal consistency between AI-based and human-based assessment scores.
Values are Pearson's correlation coefficient *r* (95% CI) and Lin's concordance correlation coefficient (CCC) values; *n* = 25 items per comparison. Higher values indicate stronger association/consistency.
Abbreviations: ABA-o1 = AI-based assessment using ChatGPT o1 Pro; ABA-5 = AI-based assessment using ChatGPT 5 Pro; HBA = human-based assessment; CCC = concordance correlation coefficient.

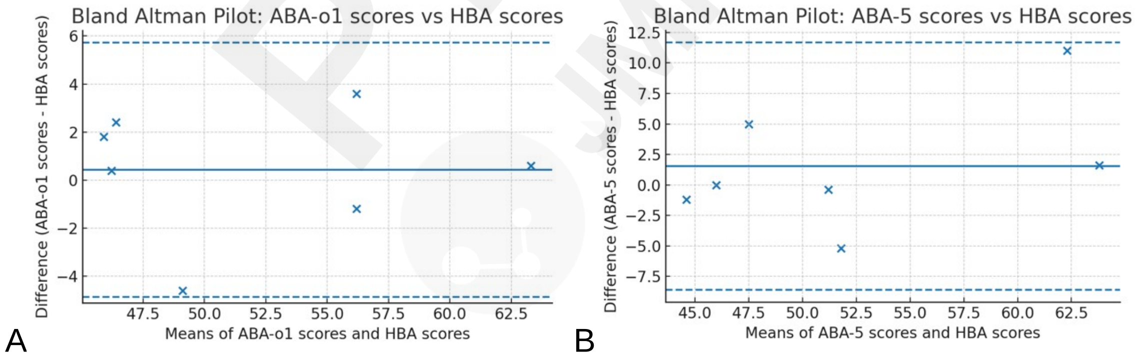


Figure 1. Bland-Altman plots comparing AI-based assessment (ABA) with human-based assessment (HBA). (A) ABA-o1 vs HBA: mean bias = 0.43; LoA = −4.87 to 5.72. (B) ABA-5 vs HBA: mean bias = 1.54; LoA = −8.60 to 11.68. Points indicate participants (×). The solid line shows the mean bias; dashed lines indicate the LoA. *Abbreviations:* ABA-o1 = AI-based assessment using ChatGPT-o1 Pro; ABA-5 = AI-based assessment using ChatGPT-5 Pro; HBA = human-based assessment.

Scores by Training Level

Table 2 summarizes mean ± SD interview scores by training level. Across methods, APs had the highest means (HBA: 57.6 ± 1.7; ABA-o1: 56.8 ± 1.7; ABA-5: 58.5 ± 13.2). MSs were next (HBA: 56.5 ± 12.0; ABA-o1: 55.0 ± 12.2; ABA-5: 55.3 ± 13.2), in some cases approximating AP performance. RPs had the lowest means (HBA: 49.2 ± 3.6; ABA-o1: 47.1 ± 0.5; ABA-5: 48.3 ± 3.8). The anticipated ordinal pattern (APs > RPs > MSs) was therefore not consistently observed, since MS means exceeded RP means across all methods.

Group	n	HBA Scores, mean ± SD	ABA-o1 Scores, mean ± SD	ABA-5 Scores, mean ± SD
Attending physicians	2	57.6 ± 1.7	56.8 ± 1.7	58.5 ± 13.2
Medical students	2	56.5 ± 12.0	55.0 ± 12.2	55.3 ± 13.2
Resident physicians	3	49.2 ± 3.6	47.1 ± 0.5	48.3 ± 3.8

Table 2. Mean interview scores (mean ± SD) by training level, as rated by human-based assessment (HBA) and AI-based assessments (ABA-o1, ABA-5).
n denotes the number of participants per group.
Abbreviations: HBA = human-based assessment; ABA-o1 = AI-based assessment using ChatGPT o1 Pro; ABA-5 = AI-based assessment using ChatGPT 5 Pro; SD = standard deviation.

Processing time (35 cases)

Total processing time was 5:59:35 for the physician benchmark, 1:56:38 for ABA-5, and 2:31:05 for ABA-o1. Average time per case was 3:19.9 for ABA-5 (batch-to-batch SD ±1:06), 4:19 for ABA-o1 (SD ±3:09), and 10:16.4 for the physician (SD ±11:09). Relative to the physician, total time was reduced by 67.6% with ABA-5 and 58.0% with ABA-o1 (Table 3).

Method	Total time (hh:mm:ss)	Mean per case	Batch-to-batch SD	Time reduction vs physician
ABA-5	1:56:38	0:03:20	±1:06	67.6%
ABA-o1	2:31:05	0:04:19	±3:09	58.0%
HBA	5:59:35	0:10:16	±11:09	—

Table 3. Analysis time by method (five independent runs/raters per method).
Total time and mean time per case are reported in hh:mm:ss. Batch-to-batch SD indicates across-run variability. “Time reduction vs physician” is the percentage reduction relative to HBA. *Abbreviations:* ABA-o1 = AI-based assessment using ChatGPT o1 Pro; ABA-5 = AI-based assessment using ChatGPT 5 Pro; HBA = human-based assessment; SD = standard deviation.

Discussion

Principal Findings

In this validation study, comparing three rater groups (HBA, ABA-o1, and ABA-5), ABA-o1 and ABA-5 produced interview ratings that were statistically indistinguishable from HBA, yet showed markedly superior psychometric stability relative to HBA (Cronbach’s α: ABA-o1=0.81, ABA-5=0.86, HBA=0.80; ICC: ABA-o1=0.77, ABA-5=0.82, HBA=0.38). Cronbach’s α values ≥ 0.8 indicate good internal consistency [10], and ICC (2,1) values ≥ 0.75 denote good inter-rater reliability [11]. Agreement metrics were likewise robust as evaluative tools: CCC assesses both correlation and bias in a single index [12], whereas Bland-Altman analysis remains the standard for visualizing bias and limits of agreement [13]. ABA-5 was benchmarked against HBA using the same agreement framework.

Though the observed differences in reliability were significant, they may also have practical implications in educational settings. The consistently higher internal consistency and inter-rater reliability suggest that ABA scoring (including ABA-o1 and ABA-5) could enhance assessment efficiency and reproducibility. Depending on the context, ABA may serve not only as a scalable adjunct, but also as a viable alternative to human raters in transcript-based clinical interview evaluations, although this requires significant larger scale validation

Comparison With Prior Work

The present findings corroborate earlier work in which LLMs matched or exceeded faculty performance when scoring free-text notes [6], designing script-concordance tests [7], and evaluating OSCE encounters [8]. A recent study showed that GPT-4o can produce inpatient documentation of comparable quality while reducing charting time by >50% [14]. Consistent with ChatGPT's near-passing performance on the USMLE [15], the current study suggests that foundation models possess clinically relevant semantic competence even in spoken communication tasks. Moreover, the 58% reduction in analysis time mirrors the 2025 *Time for Class* survey, where 36% of faculty who use generative AI daily reported a measurable workload decrease [16].

Beyond efficiency, such time savings could play a decisive role in addressing the growing problem of clinician-educator burnout and faculty shortages, which are societal challenges that threaten the sustainability of competency-based medical education [17,18]. By automating labor-intensive scoring, AI can free physicians to devote more time to high-value coaching and mentorship, thereby enhancing both educator well-being and learner support [17]. Furthermore, the superior scoring consistency observed with LLMs may help curb rater drift and cognitive biases such as leniency, halo, or contrast effects, long-recognized sources of unreliability in workplace-based assessments [19]. Improved fairness and reliability in assessment would advance equity in trainee progression and ultimately foster a more competent, patient-centered workforce.

Interpretation and Educational Implications

From an educational perspective, three observations are noteworthy when framed across the three rater groups (HBA, ABA-o1, ABA-5).

1. Consistency vs nuance – The score distributions from ABA-o1 and ABA-5 suggest these models apply the rubric more consistently than HBA, likely because their underlying embeddings execute the criteria more deterministically once sampling stochasticity is averaged across runs. Consistency is a hallmark of fair assessment; however, the absence of human nuance in ABA-o1/ABA-5 could miss contextual subtleties (e.g., cultural cues, atypical communication styles) that HBA raters may detect.
2. Efficiency gains – Relative to HBA (10 m16 s per case), ABA-5 and ABA-o1 reduced analytic time to 3 m 20 s (–67.6%) and 4 m19 s (–58.0%) per case, respectively, amounting to approximately 240 and approximately 210 faculty-minutes **[Does this revision reflect your intended meaning?]** saved across 35 encounters. In throughput terms, this corresponds to an increase in throughput from approximately 6 cases per hour with HBA to 18 cases per hour with ABA-5 and 14 cases per hour with ABA-o1, supporting more timely formative feedback and enabling the reinvestment of AI-derived efficiency gains into coaching rather than grading. In addition, ABA-5 could process

data for 3–4 individuals in a single run, reducing the need for repeated prompt inputs and minimizing data-handling overhead.

3. Level-based performance – Medical students (MSs) outperformed resident physicians (RPs) on the same rubric in this cohort. This pattern may reflect (i) sampling error in a modest cohort, (ii) case specificity favoring recently studied content, and/or (iii) a rubric that emphasizes foundational communication more than advanced clinical reasoning. Replication with larger, more varied case sets and tiered rubrics evaluated across HBA, ABA-o1, and ABA-5 is warranted.

Practically, programs could deploy an “AI-first, faculty-verified” workflow in which ABA-o1 and ABA-5 provide rapid formative scores and narrative feedback immediately after an encounter; HBA then audits a random subset for quality assurance, similar to double-reading in radiology. Such hybridity leverages the speed and reliability of LLMs while retaining human oversight for high-stakes decisions.

Strengths and Limitations

A key strength is the dual evaluation of accuracy (agreement) and efficiency (time), providing a more complete picture of implementation value than accuracy alone. Nonetheless, several limitations warrant caution:

1. Sample size and scope – Only seven participants and a single thyrotoxic periodic paralysis scenario were tested, limiting generalizability across learner levels, languages, and clinical contexts.
2. Convenience sampling – Self-selection may bias toward technology-friendly participants.
3. Model and prompt dependence – Results pertain to ChatGPT-o1 Pro and ChatGPT-5 Pro with a specific rubric prompt; other LLMs or prompt engineering strategies could alter performance.
4. Transcription fidelity – Speech-to-text errors were not exhaustively audited and may have influenced ratings. In addition, the evaluation was limited to transcribed textual data; nonverbal cues, vocal tone, and conversational pauses present in the actual interviews could not be assessed.
5. Potential systemic bias – High concordance does not preclude shared cognitive blind spots between AI and human raters; fairness audits across sex, accent, and cultural communication styles remain necessary.

Future Research

Future studies should (1) evaluate multiple diverse clinical scenarios, including psychosocially complex cases; (2) compare real-time vs post-encounter AI feedback; (3) examine learner outcomes such as skill acquisition and satisfaction; (4) conduct cost-effectiveness analyses at scale; and (5) explore bias-mitigation and explainability techniques to satisfy accreditation requirements.

Since the present study was limited to transcript-based assessments of simulated encounters, future work is also needed to evaluate how well ABA scores correlate with actual clinical performance, and whether AI can reduce assessor burden while maintaining fairness and reliability.

Conclusions

Within the constraints of this pilot, ChatGPT-o1 Pro and ChatGPT-5 Pro matched expert physicians in scoring simulated patient interviews, produced more reliable ratings, and delivered a substantial 67.6% and 58% reduction in analytical time. These preliminary results indicate that the LLM could potentially serve as a complementary or alternative tool to human raters for transcript-based interview assessments. This approach warrants further investigation as a means to contribute to assessment efficiency in medical education. Careful curricular design and continuous human oversight will be essential to ensure that such tools enhance, rather than compromise, the validity and equity of learner evaluations.

Acknowledgments

The authors would like to thank the clinical instructors for their time and the participants for engaging in this educational pilot. This manuscript was prepared with the assistance of generative artificial intelligence.

Conflicts of Interest

None declared.

Funding

This study was supported by JSPS KAKENHI Grant Number JP23K05953, by the Japan Medical Education Foundation FY2025 Research Grant, and by the program Creating training hubs for advanced medical personnel by the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

References

1. Talwalkar JS, Fortin AH, Morrison LJ, et al. An advanced communication skills workshop using standardized patients for senior medical students. MedEdPORTAL. 2021;17:11163. doi:10.15766/mep_2374-8265.11163 PMID:34124349

2. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach. 2025;47(1):40-42. doi:10.1080/0142159X.2024.2376879 PMID:38992981

3. Cheung K, Rogoza C, Chung AD, Kwan BYM. Analyzing the administrative burden of competency-based medical education. Can Assoc Radiol J. 2022;73(2):299-304. doi:10.1177/08465371211038963 PMID:34449283

4. Szulewski A, Braund H, Dagnone DJ, et al. The assessment burden in competency-based medical education: how programs are adapting. Acad Med. 2023;98(11):1261-1267. doi:10.1097/ACM.0000000000005305 PMID:37343164

5. Takahashi H, Shikino K, Kondo T, Komori A, Yamada Y, Saita M, Naito T. Educational utility of clinical vignettes generated in Japanese by ChatGPT-4: mixed methods study. JMIR Med Educ. 2024;10:e59133. doi:10.2196/59133 PMID:39137031

6. Burke HB, Hoang A, Lopreiato JO, et al. Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. JMIR Med Educ. 2024;10:e56342. doi:10.2196/56342 PMID:39118469

7. Hudon A, Kiepura B, Pelletier M, Phan V. Using ChatGPT in psychiatry to design script concordance tests in undergraduate medical education: mixed methods study. JMIR Med Educ. 2024;10:e54067. doi:10.2196/54067 PMID:38596832

8. Huang T-Y, Hsieh P-H, Chang Y-C. Performance comparison of junior residents and ChatGPT in the objective structured clinical examination for medical history taking and documentation: development and usability study. JMIR Med Educ. 2024;10:e59902. doi:10.2196/59902 PMID:39622713

9. Pfeiffer CA. Master interview rating scale (MIRS). Norfolk, VA: Eastern Virginia Medical School; 2003. Available from: https://www.dmu.edu/wp-content/uploads/Master-Interview-Rating-Scale.pdf [accessed Jul 23, 2025].

10. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53-55. doi:10.5116/ijme.4dfb.8dfd PMID:28029643

11. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012 PMID:27330520

12. Lin LK. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255-268. PMID:2720055

13. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-310. doi:10.1016/S0140-6736(86)90837-8 PMID:2868172

14. Lu X, Gao X, Wang X, et al. Comparison of medical history documentation efficiency and quality based on GPT-4o: a study on the comparison between residents and artificial intelligence. Front Med (Lausanne). 2025;12:1545730. doi:10.3389/fmed.2025.1545730 PMID:4043835

15. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. doi:10.2196/45312 PMID:36753318

16. Tyton Partners. Time for Class 2025 report: daily AI use linked to reduced faculty workload. Published Jun 11, 2025. Available from: https://www.d2l.com/newsroom/tyton_partners_report_examines_ai_in_higher_education/ [accessed Jul 23, 2025].

17. Banerjee G, Mitchell JD, Brzezinski M, DePorre A, Ballard HA. Burnout in academic physicians. Perm J. 2023;27(2):142-149. doi:10.7812/TPP/23.032 PMID:37309180

18. Association of American Medical Colleges. AAMC applauds introduction of bill to reduce physician shortage [press release]. AAMC. Published Jun 10, 2025 [cited Jul 29, 2025]. Available from: https://www.aamc.org/news/press-releases/aamc-applauds-introduction-bill-reduce-physician-shortage-0

19. Yeates P, McCray G. Investigating the accuracy of adjusting for examiner differences in multicentre objective structured clinical exams (OSCEs): a simulation study of video-based examiner score comparison and adjustment (VESCA). BMC Med Educ. 2024;24(1):1466. doi:10.1186/s12909-024-06462-3 PMID:39695612

Multimedia Appendix

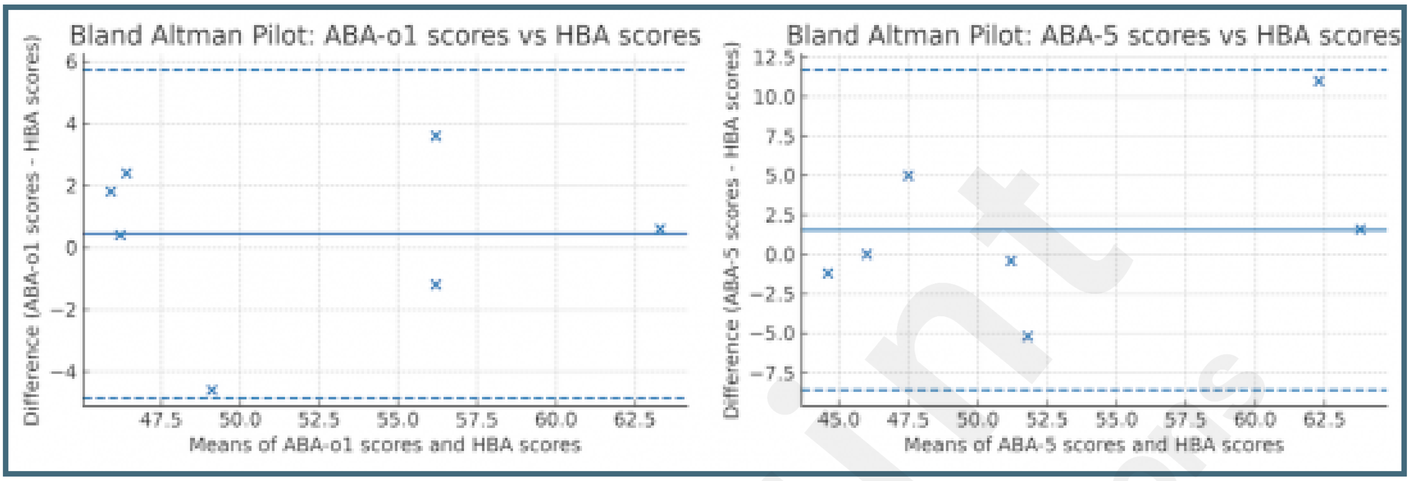
Prompt used for ChatGPT-o1 Pro scoring of medical interview transcripts.

This appendix provides the exact prompt text submitted to ChatGPT o1-Pro (OpenAI, San Francisco, California, USA) for the automated assessment of clinical interview transcripts. The prompt instructed the model to evaluate each transcript using the 25-item Master Interview Rating Scale (MIRS), assign a score for each item (1–5), and provide justifications based on specific utterances or actions within the dialogue. This standardized prompt was used across all ABA runs to ensure consistency and reproducibility.

Supplementary Files

Figures

Bland-Altman plots comparing AI-based assessment (ABA) with human-based assessment (HBA). (A) ABA-o1 vs HBA: mean bias = 0.43; LoA = ?4.87 to 5.72. (B) ABA-5 vs HBA: mean bias = 1.54; LoA = ?8.60 to 11.68. Points indicate participants (×). The solid line shows the mean bias; dashed lines indicate the LoA. Abbreviations: ABA-o1 = AI-based assessment using ChatGPT-o1 Pro; ABA-5 = AI-based assessment using ChatGPT-5 Pro; HBA = human-based assessment.



TOC/Feature image for homepages

A symbolic illustration showing a doctor and an AI robot shaking hands under dramatic lighting. The glowing spiral between their hands represents collaboration and shared evaluation, highlighting the partnership of human expertise and artificial intelligence in assessing medical interviews.