“Comparing AI against Human Experts in Evaluating the Communication of Simulated Doctor-Patient Interactions”

Kylan Steele BS[1], Bradley Grimm MS[3], John Gunaldo BS[1], Peter DeBlieux MD[1,2], Rachel Fiore PhD CCC-SLP[1].

[1]Standardized Patient Project, Louisiana State University Health Sciences Center
[2]Office of Medical Education, Louisiana State University Health Sciences Center
[3]Videra Health Inc.

Introduction

Background: Effective doctor–patient communication is fundamental in promoting patient understanding, trust, and adherence. When physicians connect empathically, patients experience better outcomes and satisfaction. Medical education places strong emphasis on teaching these skills yet evaluating them remains challenging.

Traditional assessments such as Objective Structured Clinical Examinations (OSCEs) rely on human raters using itemized checklists, an approach that is valuable but also resource-intensive, time-consuming, and subject to human variability. These challenges highlight the need for more scalable, objective methods to evaluate communication skills in medical education.

Recent advances in artificial intelligence (AI) offer potential for standardized communication assessment; however, AI’s ability to accurately evaluate nuanced aspects of patient interaction, particularly those involving nonverbal cues or affective tone, remains unclear.

Objective: To examine the alignment, discrepancies, and consistency between AI- and human expert–based evaluations of clinical communication performance.

Methods

Design: Cross-sectional, mixed-methods study comparing AI and human-expert evaluations of physician communication in standardized patient encounters.
Videos: Four simulated primary care visits were filmed at LSUHSC. Each scenario included a low-skill and high-skill communication version, separated by a brief reflection segment.
AI Model: Videra Health’s multimodal system analyzed visual, audio, and transcribed dialogue inputs using natural language processing and machine learning to produce 5-point Likert ratings across 12 communication domains.
Human Raters: 17 members of LSUHSC’s Standardized Patient Working Group (clinicians, faculty, students, educators) rated the same 12 domains and answered 4 open-ended questions assessing perceived trust in the encounter.
Procedure: AI evaluated all four videos independent of human input. Human raters were then shown the videos separately and submitted surveys through Microsoft Forms. Scores were aggregated for comparison across empathy and teach-back domains.
Analysis:

- Quantitative: Agreement within ±1 and ±0.5 Likert points, Mean Absolute Deviation (MAD), and Spearman correlation coefficients.
- Qualitative: Qualitative review of open-ended responses on trust-related behaviors was compared to AI-generated qualitative feedback to identify patterns or recurring themes that may contextualize the quantitative data.

Overall Agreement Between AI and Human Raters

Video	Communication Quality	Human Average	AI Average	Agreement Rate
Video 1	Poor	1.58	2.08	91.7%
Video 2	Good	3.74	3.33	91.7%
Video 3	Poor	2.19	3.08	58.3%
Video 4	Good	4.15	4.17	100%

Table 1. AI and human average communication ratings across four videos, with agreement rates based on scores within one Likert point. Videos were categorized based on the intended communication skill level (poor or good).

Human Rater Variability Across Videos

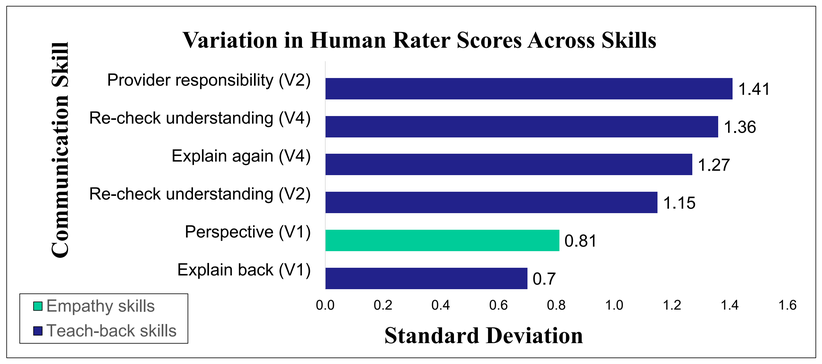

Figure 1. Standard deviation (SD) of human rater scores for communication skills across videos. Lower SD values indicate strong rater agreement, while higher values indicate greater variability. Skills included in the graph were selected based on notably high or low SD values to demonstrate the range of inter-rater consistency. Green bars represent skills in the “teach-back” domain, while blue bars represent skills in the “empathy” domain.

AI–Human Agreement on Perceived Trust Across Videos

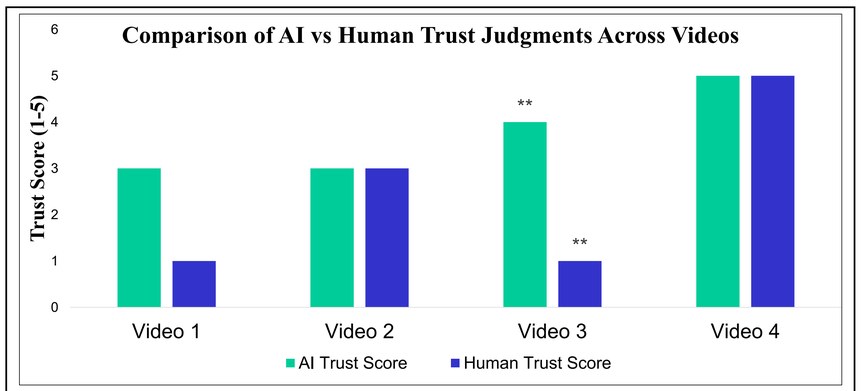

Figure 2. Comparison of AI-generated and human-perceived trust scores across four physician-patient encounter videos. Human trust scores were derived from a qualitative analysis of open-ended response data, while AI scores reflect the model’s evaluations of trust based on its clinical communication assessment framework. Asterisk marks above Video 3 indicate notable divergence between AI and human trust ratings.

Results

Overall AI–Human Alignment:

- AI and human raters demonstrated 85% agreement within one Likert point across all four standardized videos.
- AI showed strong directional accuracy, correctly distinguishing high- vs. low-quality communication, with strongest alignment in the “good” communication videos (Videos 2 and 4).
- Mean Absolute Deviation (MAD) = 0.68; AI tended to slightly overestimate communication scores (+0.34 bias).

Domain-Level Agreement:

- Agreement was highest for empathy-related skills (average 90%) and slightly lower for teach-back behaviors (82%).
- AI performed best on concrete, linguistic skills such as **“**plain language” and **“**expressing concern”.
- Lower agreement occurred on multi-step skills like “ask to explain back” and “re-check understanding.”

Human Rater Variability:

- Human raters were most consistent when evaluating plain language and empathy behaviors but greatly diverged on teach-back items.
- This variability highlights the subjectivity inherent in communication assessment and provides context for minor AI–human discrepancies.

Qualitative Trends:

- AI and human raters used similar criteria to identify trust-building behaviors, but AI was less sensitive to nonverbal cues such as tone or demeanor.
- Divergence was most evident in Video 3, where human raters cited an “inappropriate tone” as damaging trust—a nuance the AI missed.

Conclusion

- AI-generated evaluations of physician communication showed **strong alignment with expert human ratings** across simulated encounters.
- Agreement was **highest in empathy-related skills**, where linguistic markers were clear, and **lowest in teach-back behaviors**, which required sequential, interpretive judgment, suggesting that AI performs better when language markers are evident and less subject to interpretation.
- The **variability in subjective domains** demonstrated among **human raters** themselves, emphasized both the **challenge and necessity of objective tools** for communication assessment.
- These findings indicate that AI offers potential as a **scalable adjunct** to human evaluation in medical education, providing consistent feedback for structured communication skills.
- However, **human oversight remains essential** for context-dependent and emotional aspects of interaction, which are areas where nuance and empathy are best recognized by trained observers.