

# Harnessing Artificial Intelligence (AI) for Surgical Education: Comparing AI- and Surgeon-Generated Feedback in Informed Consent Training

Catherine Gbekie

cg3381@cumc.columbia.edu

Columbia University Vagelos College of Physicians and Surgeons <https://orcid.org/0000-0002-1684-0552>

Brianna M. Peet

Columbia University Vagelos College of Physicians and Surgeons

Mary Grace Whelan

Columbia University Vagelos College of Physicians and Surgeons

Grace B. Simmons

Columbia University Vagelos College of Physicians and Surgeons

Daniel P. Pacella

Columbia University Vagelos College of Physicians and Surgeons

Viemma Nwigwe

Columbia University Irving Medical Center

Angel Rosario

Columbia University Irving Medical Center

---

## Research Article

### Keywords:

**Posted Date:** October 19th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11120355/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# Abstract

**Purpose:** Informed consent (IC) is a complex communication competency requiring procedural knowledge, patient-centered decision-making, and nuanced risk-benefit communication. Large language models (LLMs) offer scalable opportunities for simulated IC training and individualized feedback; however, their ability to accurately evaluate learner performance remains insufficiently validated. We compared AI- and surgeon-generated evaluations of simulated IC encounters to assess performance score alignment and narrative feedback quality.

**Methods:** Six medical students completed an LLM-based simulated appendectomy consent incorporating a culturally sensitive scenario. Each transcript received one AI- and four surgeon-generated evaluations using a 17-item IC rubric. Scoring alignment was defined as the difference between AI-assigned and surgeon-assigned scores. Feedback quality was evaluated using an 8-item modified Completed Clinical Evaluation Report Rating (mCCERR) framework assessing comprehensiveness and balance, specificity and justification, and educational utility and constructiveness.

## Results

AI assigned higher performance scores than surgeons, with significant differences in addressing patient emotions, explaining the procedure, risks, and alternatives, and avoiding coercion (all  $P = 0.03$ ). The largest absolute differences included explaining contraindications, assessing capacity, and integrating patient preferences. AI-generated feedback received significantly higher quality ratings, with a higher median overall mCCERR score than surgeon-generated feedback (38.5 vs 24.75,  $P = 0.04$ ).

## Conclusions

LLMs generated high-quality, comprehensive feedback but demonstrated systematic discordance with surgeon assessment, consistently assigning more favorable performance scores. These findings demonstrate that high feedback quality should not be viewed as a proxy for greater evaluative validity. LLMs may be well-suited for repeated simulated practice and individualized formative feedback but should be paired with human oversight for nuanced assessment of IC proficiency.

# 1 | INTRODUCTION

High-quality feedback is critical to the development of clinical acumen and effective communication skills among surgical trainees, but providing timely, individualized feedback remains challenging. At large academic teaching hospitals, the responsibility for delivering meaningful feedback and guidance often falls on surgical faculty, who must balance educational responsibilities with unpredictable clinical workloads, research, and other demands of medical practice. These constraints are frequently cited as barriers to effective teaching and feedback.<sup>1-3</sup> Consequently, trainee assessments are often incomplete or delayed, with one study finding that 58% of orthopedic surgery residents did not receive timely end-of-rotation feedback.<sup>4</sup> These challenges underscore the need for scalable approaches that can provide surgical trainees with more frequent, personalized, and accessible feedback and performance evaluations.<sup>5</sup>

Among the important communication skills surgical trainees must develop and refine, informed consent is a vital competency requiring surgeons to integrate operative knowledge, patient-centered decision-making, and nuanced risk-benefit communication. Informed consent also provides a meaningful model for studying feedback on surgical communication, as the consenting process is both clinically consequential and sufficiently structured to allow for systematic evaluation. However, despite its importance, critical gaps in surgical informed consent performance and comfort among residents and medical students have been documented in the literature.<sup>6</sup> One study found that only 4.2% of residents routinely disclosed all essential elements of informed consent to patients, citing inadequate knowledge, communication barriers, and lack of practice and standardized education as major barriers to effective informed consent.<sup>4,7</sup> Additional deficiencies include limited discussion of treatment alternatives, inadequate assessment of patient understanding, incomplete disclosure of procedure-specific risks, and incongruity between information included in written consent forms and information communicated verbally.<sup>8-10</sup> These findings highlight the need for greater opportunities to strengthen informed consent training through experiential learning, individualized assessment, and personalized feedback.<sup>11-13</sup>

Generative artificial intelligence (AI) may offer a scalable means of providing such feedback. Emerging evidence has demonstrated promising applications of large language models (LLMs) across surgical education, including enhancing the efficiency of faculty-trainee feedback delivery, generating high-quality informed consent materials, facilitating automated coaching associated with improved technical performance, and providing a low-risk environment for learners to practice history-taking and challenging patient conversations.<sup>14-19</sup> Together, these findings suggest that LLMs may provide a platform for communication-based simulation and formative feedback.

However, despite growing evidence supporting LLMs as simulated patients and feedback tools, their ability to evaluate learner performance with similar or superior fidelity to surgeons has been unexplored. Accordingly, we compared LLM- and surgeon-generated evaluations of medical students completing simulated informed consent encounters. We assessed both alignment in performance scoring and the quality of the resulting narrative feedback to characterize the potential role and limitations of generative AI as an assessment and feedback tool for informed consent training and surgical education more broadly.

## 2 | METHODS

### 2.1 | Study Design and Simulation

Six medical students completed a simulated informed consent encounter for laparoscopic appendectomy using a programmed large language model (LLM)-based chatbot. All participants had completed a core surgery rotation at a large academic medical center and received introductory instruction on the principles of surgical informed consent prior to the simulation session. Students engaged in a one-on-one, voice- and text-enabled encounter with a synthetic patient developed using

ChatGPT (GPT-4; OpenAI, San Francisco, CA). The patient encounter incorporated a culturally sensitive scenario involving religious beliefs and restrictions, designed to assess communication skills beyond procedural knowledge and the core elements of informed consent. Following the conclusion of the patient encounter, participants received an AI-generated feedback summary from the same LLM, which was programmed to evaluate performance using a standardized informed consent rubric comprising 17 scored items and one non-scored interpreter-use item (**Supplemental Fig. 1**). De-identified encounter transcripts were collected from October through December 2025.

## 2.2 | Performance Evaluation

Each learner transcript received one AI-generated evaluation and was independently evaluated by four surgeon evaluators, including surgical residents and attending surgeons. The LLM and surgeon evaluators utilized the same predefined 17-item rubric, adapted from the New Mexico Clinical Communication Scale (NMCCS), a framework capturing core elements of patient-centered communication during clinical encounters, and prior literature evaluating informed consent.<sup>20–24</sup> The informed consent rubric assessed key components of the informed consent process, including acknowledgment of patient emotions, assessment of illness knowledge, explanation of the indication for surgery, contraindications, nature of the procedure, risks, benefits, and alternatives. Each item was scored from 0 to 3, yielding a maximum total score of 51. AI and surgeon evaluators were also instructed to provide narrative feedback, including justification for their scores and recommendations for improvement.

## 2.3 | Outcomes and Statistical Analysis

The primary outcome was scoring alignment between AI and surgeon evaluators. For each transcript and informed consent rubric item, the median score across the four surgeon evaluators was calculated and compared with the corresponding AI-generated score. Scoring alignment was defined as the difference between the AI-assigned score and the median surgeon-assigned score (AI – surgeon), such that positive values indicated higher performance ratings assigned by AI and negative values indicated lower ratings assigned by AI. Differences approaching zero indicated greater alignment between AI and surgeon evaluators. Differences between AI- and surgeon-assigned performance scores were analyzed using a Wilcoxon rank-sum test.

The secondary outcome was the quality of the narrative feedback generated by AI and surgeon evaluators, assessed using an adapted Completed Clinical Evaluation Report Rating (CCERR) scale. The original CCERR is a validated 9-item instrument developed to assess the quality of written clinical performance evaluations.<sup>25</sup> A previously published modified CCERR (mCCERR) adapted the instrument for individual feedback encounters by removing an item assessing trainee responsiveness to feedback over time.<sup>26</sup> A similar 8-item instrument was used in this study, with each item rated from 0 to 5, yielding a maximum score of 40 (**Supplemental Fig. 2**). Rubric items were grouped into three key dimensions of feedback quality: Comprehensiveness and Balance (Q1, Q2, Q8), Specificity and Justification (Q3–Q5), and Educational Utility and Constructiveness (Q6–Q7). Each AI-generated (one per student transcript)

and surgeon-generated narrative feedback report (four per student transcript) was scored by two independent reviewers (BP, CG). For each feedback report, the median score across the two reviewers was calculated. For surgeon-generated feedback, these scores were then aggregated by calculating the median across the four surgeon evaluations for each transcript. This yielded one AI-generated and one surgeon-generated feedback-quality score per student transcript, resulting in six paired observations for analysis. Transcript-level feedback-quality scores were summarized using descriptive statistics. Differences in feedback-quality scores between AI- and surgeon-generated evaluations were assessed using the Wilcoxon rank-sum test. Statistical significance was defined as  $p < 0.05$ .

A qualitative comparison of the AI- and surgeon-generated narrative feedback reports was performed to contextualize quantitative differences in feedback quality. Representative quotations were extracted to illustrate differences across the three feedback-quality dimensions, and surgeon-generated feedback with higher and lower mCCERR scores were selected for comparison. All analyses were conducted using R version 4.3.0. Given the educational nature of the project, the study received a non-human subjects research determination from the institutional review board.

## 3 | RESULTS

### 3.1 | AI-Surgeon Alignment in Informed Consent Performance Scoring

Six medical students participated in the simulation and were evaluated. Differences between AI and surgeon evaluators emerged at the individual rubric-item level. AI assigned significantly higher scores than surgeon evaluators for the following rubric items: “Addressing Patient Emotions” (median [IQR], 3 [3–3] vs. 2 [2–2.375]), “Explains Nature of Procedure” (3 [3–3] vs. 1.5 [1–2]), “Explains Risks” (3 [3–3] vs. 1.75 [1.125–2]), “Explains Alternatives” (3 [3–3] vs. 0.75 [0.5–1.75]), and “Avoids Coercion” (3 [3–3] vs. 2.5 [2.125–2.5]; all  $P = 0.03$ ; Table 1). The median absolute difference between AI- and surgeon-generated scores was largest for “Explains Contraindications” (1.75 points), followed by “Assesses Capacity” (1.5 points), “Explains Nature of Procedure” (1.5 points), and “Integrates Patient Preferences” (1.25 points; Fig. 1).

Table 1  
Comparison of AI- and Surgeon-Generated Informed Consent Performance Scores

| Informed Consent Rubric Item                                                                                                                                                                                               | Performance scores |                    | <i>P</i> value |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|--------------------|----------------|
|                                                                                                                                                                                                                            | AI                 | Surgeon            |                |
| Introduction to encounter                                                                                                                                                                                                  | 3 (3, 3)           | 2 (2, 2)           | 0.05           |
| Elicits patient concerns                                                                                                                                                                                                   | 3 (3, 3)           | 2.5 (2.125, 2.875) | 0.1            |
| Addresses patient concerns                                                                                                                                                                                                 | 3 (3, 3)           | 2.5 (2.125, 2.875) | 0.1            |
| Integrating patient preferences                                                                                                                                                                                            | 3 (3, 3)           | 1.75 (1.5, 2.375)  | 0.06           |
| Addressing patient emotions                                                                                                                                                                                                | 3 (3, 3)           | 2 (2, 2.375)       | <b>0.03</b>    |
| Information sharing practices                                                                                                                                                                                              | 3 (3, 3)           | 2.25 (2, 2.5)      | 0.05           |
| Word choice                                                                                                                                                                                                                | 3 (3, 3)           | 2.25 (2, 2.875)    | 0.09           |
| Illness knowledge                                                                                                                                                                                                          | 3 (3, 3)           | 2 (2, 2.375)       | 0.09           |
| Explains need for surgery                                                                                                                                                                                                  | 3 (3, 3)           | 2.25 (2, 2.5)      | 0.05           |
| Explains contraindications                                                                                                                                                                                                 | 2 (2, 2)           | 0 (0, 0.375)       | 0.06           |
| Explains nature of procedure                                                                                                                                                                                               | 3 (3, 3)           | 1.5 (1, 2)         | <b>0.03</b>    |
| Explains risks                                                                                                                                                                                                             | 3 (3, 3)           | 1.75 (1.125, 2)    | <b>0.03</b>    |
| Explains benefits                                                                                                                                                                                                          | 3 (3, 3)           | 2 (1.25, 2.375)    | 0.06           |
| Explains alternatives                                                                                                                                                                                                      | 3 (3, 3)           | 0.75 (0.5, 1.75)   | <b>0.03</b>    |
| Encourages questions                                                                                                                                                                                                       | 3 (3, 3)           | 2.25 (2, 2.875)    | 0.09           |
| Avoids coercion                                                                                                                                                                                                            | 3 (3, 3)           | 2.5 (2.125, 2.5)   | <b>0.03</b>    |
| Assesses capacity                                                                                                                                                                                                          | 3 (2.25, 3)        | 1.25 (1, 1.5)      | 0.09           |
| Values are presented as median (IQR). Surgeon performance scores represent the median score across four independent surgeon evaluators for each learner. <i>P</i> values were calculated using the Wilcoxon rank-sum test. |                    |                    |                |

## 3.2 | Quality of AI- and Surgeon-Generated Narrative Feedback

AI-generated feedback received significantly higher quality scores across all three dimensions: Comprehensiveness and Balance ( $P = 0.04$ ), Specificity and Justification ( $P = 0.04$ ), and Educational Utility and Constructiveness ( $P = 0.03$ ; Table 2). Consistent with these dimension-level differences, the median overall mCCERR score was 38.5 out of 40 for AI-generated feedback compared with 24.75 for surgeon-generated feedback ( $P = 0.04$ ; Fig. 2).

Table 2  
Comparison of AI- and Surgeon-Generated Narrative Feedback Quality

| Feedback Quality Item                                                                                   | Quality scores   |                     | <i>P</i> value |
|---------------------------------------------------------------------------------------------------------|------------------|---------------------|----------------|
|                                                                                                         | AI               | Surgeon             |                |
| Comprehensiveness and Balance                                                                           |                  |                     |                |
| Q1. Domain Coverage and Performance Differentiation                                                     | 5 (5, 5)         | 4 (4, 4.75)         | 0.1            |
| Q2. Balance of Strengths and Improvements                                                               | 5 (4.625, 5)     | 3.25 (2.625, 3.875) | 0.04           |
| Q8. Clarity and Comprehensiveness                                                                       | 5 (5, 5)         | 2.25 (2, 2.875)     | 0.04           |
| Specificity and Justification                                                                           |                  |                     |                |
| Q3. Justification of Judgments or Scoring                                                               | 4.5 (4.125, 4.5) | 3.25 (2.25, 3.5)    | 0.04           |
| Q4. Specific Examples of Strengths                                                                      | 5 (4.625, 5)     | 3 (2.25, 3.375)     | 0.04           |
| Q5. Specific Examples of Weaknesses                                                                     | 4.5 (4.5, 4.5)   | 3 (3, 3.375)        | 0.04           |
| Educational Utility and Constructiveness                                                                |                  |                     |                |
| Q6. Actionable Recommendations                                                                          | 5 (4.625, 5)     | 3 (2.625, 3)        | 0.03           |
| Q7. Supportive Professional Tone                                                                        | 5 (4.625, 5)     | 2.75 (2.5, 3)       | 0.03           |
| Values are presented as median (IQR). <i>P</i> values were calculated using the Wilcoxon rank-sum test. |                  |                     |                |

Representative excerpts from AI- and surgeon-generated feedback were selected for qualitative comparison across the three feedback-quality dimensions (Table 3). Within each comparison, excerpts were drawn from the same student transcript to minimize heterogeneity attributable to differences in student performance. Generally, surgeon-generated feedback was shorter and more concise than AI-generated feedback. Surgeon-generated feedback with higher mCCERR scores, as assessed by the mCCERR, highlighted learner strengths and areas for improvement and provided specific guidance regarding next steps. In contrast, surgeon-generated feedback with lower mCCERR scores was more limited in scope and less balanced, primarily focusing on identifying areas for improvement and errors rather than consistently acknowledging learner strengths.

Table 3  
Representative Examples of AI- and Surgeon-Generated Narrative Feedback by Feedback Quality Dimension

| Feedback Quality Dimension                                                                                                                                                                                                                                                                                                              | Higher Quality Surgeon Feedback                                                                                                                                                                                                                                       | Lower Quality Surgeon Feedback                                                        | AI-Generated Feedback Excerpt                                                                                                                                                                                                                                                                                                   |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Comprehensiveness and Balance</b>                                                                                                                                                                                                                                                                                                    | <i>"Introduction is complete but did not get to be warm enough. Could be more approachable and ask the patient in advance if it's a good time to talk, find the best place and best time." [T5]</i>                                                                   | <i>"Very brief introduction, difficult to assess tone/warmth via text." [T5]</i>      | <i>"You clearly introduced yourself, stated your role as a medical student, and set the agenda respectfully. This created a professional and patient-centered tone. You could enhance rapport further by using a more personalized opening (e.g., brief acknowledgment of how overwhelming the day has been for him)." [T5]</i> |
| <b>Specificity and Justification</b>                                                                                                                                                                                                                                                                                                    | <i>"[Illness knowledge was] superficial but touches on patient's understanding of illness and gives brief explanation of consequences of untreated appendicitis." [T3]</i>                                                                                            | <i>"Needs to do better about exploring patients' perspective." [T3]</i>               | <i>"Justification: You described appendicitis and sepsis progression in accessible, urgent, but non-alarming terms. Excellent work; maintain clear link between condition and need for intervention." [T3]</i>                                                                                                                  |
| <b>Educational Utility and Constructiveness</b>                                                                                                                                                                                                                                                                                         | <i>"Using the word 'options' to emphasize to patient that the choice to proceed or not is ultimately theirs. Could emphasize this more by asking the patient at the end something like 'Now that you have this information, how would you like to proceed?'" [T6]</i> | <i>"Allows a lot of time/silence for patient to reach their own conclusion." [T6]</i> | <i>"You emphasized respect for autonomy and religious belief throughout; your tone was supportive, not pressuring. Improvement advice: Maintain that patient-led framing even under time pressure." [T6]</i>                                                                                                                    |
| Higher and lower quality surgeon feedback refers to representative excerpts from the highest- and lowest-scoring surgeon evaluations, respectively, for the same student transcript within each feedback quality dimension, as assessed using the mCCERR. <b>T3</b> , Transcript 3; <b>T5</b> , Transcript 5; <b>T6</b> , Transcript 6. |                                                                                                                                                                                                                                                                       |                                                                                       |                                                                                                                                                                                                                                                                                                                                 |

AI-generated feedback was consistently more comprehensive, providing detailed information to guide learner improvement. Compared with surgeon-generated feedback, AI-generated feedback provided justification for its assigned scores and contextualized both areas for improvement and recommendations for future performance. In the representative excerpts, surgeon-generated feedback generally offered broad recommendations without concrete examples or detailed explanations, with one exception as shown in Table 3 (Educational Utility and Constructiveness, T6). AI-generated feedback also

consistently employed supportive and empathetic language, whereas surgeon-generated feedback was more direct and less frequently included explicit validation of learner performance.

## **4 | DISCUSSION**

### **4.1 | Quality and Educational Utility of AI-Generated Feedback**

Generative AI has attracted growing interest in medical and surgical education because of its accessibility, scalability, and capacity to provide personalized learner experiences. Prior studies have demonstrated potential applications of LLMs for simulation-based training and feedback delivery.<sup>17,27</sup> Brügge et al. developed an LLM-based patient simulation in which medical students completed simulated history-taking encounters and received personalized feedback generated by the same LLM.<sup>14</sup> Students receiving AI-generated feedback demonstrated greater improvement in clinical decision-making across successive encounters than those completing simulations without feedback. Similarly, Jarry et al. found that residents perceived AI-generated feedback following surgical scenarios as comparable to feedback from experienced surgeons in both usefulness for skill improvement and overall quality.<sup>28</sup>

Our findings extend this literature by directly comparing the quality of AI- and surgeon-generated feedback. AI-generated feedback received higher ratings than surgeon-generated feedback across all three quality domains of the mCCERR. Qualitative comparison also demonstrated that AI feedback consistently identified strengths and areas for improvement, provided explicit justification for its assessments, and offered actionable next steps. Surgeon-generated feedback was more variable in completeness and specificity, a finding that may reflect the time pressures and competing clinical demands previously identified as barriers limiting the depth, timeliness, and nuance of faculty-provided feedback. LLMs may therefore provide a scalable mechanism for delivering structured, comprehensive feedback while minimizing additional clinician burden.

Importantly, however, prior demonstrations of AI feedback utility have not established the validity of its underlying assessment. In Brügge et al., for example, learner performance was independently evaluated by human raters without direct comparison of AI- and human-generated scores. Our study addresses this gap by evaluating both the quality of LLM-generated feedback and its alignment with surgeon-generated assessment, distinguishing its potential value as a feedback tool from its capacity to serve as an evaluator of learner performance.

### **4.2 | AI-Surgeon Discordance in Performance Assessment**

Despite receiving higher feedback-quality ratings, AI demonstrated substantial discordance with surgeons in its assessment of learner performance. AI generally assigned higher performance scores across the informed consent rubric, with statistically significant differences in addressing patient

emotions, explaining the nature and risks of the procedure, explaining alternatives, and avoiding coercion. The largest absolute differences were observed in explaining contraindications and the nature of the procedure, assessing capacity, and integrating patient preferences. Most AI ratings were also at or near the maximum allotted score, suggesting limited discrimination across levels of learner performance.

The qualitative comparison further illustrated this discordance. AI characterized one student's Illness Knowledge as "excellent," while the surgeon described the same student's understanding of appendicitis as "superficial" and recommended a more detailed explanation of the consequences of untreated disease (Specificity and Justification, T3; Table 3). Thus, although AI-generated feedback was more comprehensive and structured, its performance assessments appeared less sensitive to the deficiencies in depth, nuance, and case understanding that were identified by surgeon evaluators. These findings demonstrate that while LLMs may offer useful feedback, they remain imperfectly aligned with expert, human performance assessment, which is informed by years of clinical practice.

## **4.3 | Implications for AI-Assisted Informed Consent Training**

The observed discrepancy may reflect the complexity inherent in evaluating informed consent discussions. Prior work suggests that LLMs may perform well when assessing explicitly defined clinical or technical behaviors.<sup>28</sup> Jarry et al., for example, found that AI tools could effectively recognize and respond to discrete, predefined technical errors when evaluating simulated surgical scenarios. Informed consent, by contrast, requires assessment not only of whether requisite information and communication behaviors are present, but also whether they are appropriately contextualized and responsive to the patient's needs. A learner may reference an operative risk or management alternative without adequately acknowledging a patient's cultural preferences, determining whether the patient can meaningfully engage in shared decision-making, or confirming the patient's understanding and comfort. Notably, in this study, several domains demonstrating substantial AI-surgeon discordance, including assessment of capacity and integration of patient preferences, require this type of contextual judgment.

LLMs may therefore be more effective at identifying the presence of required informed consent elements than at distinguishing superficial inclusion from meaningful execution within nuanced clinical interactions. This limitation has important educational implications. Overly favorable and inflated performance assessments could provide false reassurance and inadvertently reinforce incomplete or suboptimal consenting practices. This risk is particularly relevant to informed consent, where proficiency extends beyond procedural knowledge and inclusion of requisite elements to effective communication adapted to patient preferences, cultural considerations, and clinical context. Although LLMs may reduce barriers to repeated formative practice, their ability to independently determine whether proficiency in informed consent has been demonstrated requires further validation.

Taken together, our findings support an adjunctive model of AI-assisted training. LLMs may be particularly useful for repeated, low-stakes practice and immediate, formative feedback, but faculty

oversight remains important for identifying deficiencies that require contextual judgment and for consequential assessments of patient-facing competency. A hybrid approach could leverage the accessibility and consistency of AI-generated feedback while preserving expert clinical judgment where it is most consequential. Future studies should explore AI calibration to expert assessment across learner levels, diverse clinical scenarios, and different LLMs to determine whether reliable alignment between AI and human evaluations can be achieved through model refinement.

## 4.4 | Limitations

This study should be interpreted in the context of several limitations. First, only six medical students from a single institution were included, limiting the statistical power and generalizability across learners, clinical contexts, and communication-based competencies. Despite the involvement of multiple independent surgeon evaluators, human ratings are inherently subjective; therefore, observed discordance between AI and surgeon scores cannot be interpreted as definitive evidence of AI error. Additionally, our findings reflect the performance of ChatGPT under our specific prompting strategy, informed consent rubric, and adapted mCCERR framework and may not generalize to other LLMs or implementation approaches. Finally, we examined the cross-sectional quality of AI- and surgeon-generated feedback rather than the downstream educational implications of receiving such feedback. As such, it remains unknown whether feedback rated as “higher quality” using the mCCERR translates into improved learner performance. Evaluating this relationship represents a fruitful future avenue for further investigation.

## 5 | CONCLUSION

Generative artificial intelligence offers a scalable platform for simulated informed consent training, practice, and individualized feedback. This study highlights a critical distinction between the quality of AI-generated feedback and the validity of the performance assessment underlying it. Although generative AI provides comprehensive and constructive feedback, its evaluative discordance with surgical experts raises concerns regarding its capacity to independently assess learner performance. Accordingly, generative AI should not currently serve as a standalone performance evaluator and may instead be most valuable for repeated practice and individualized formative feedback, with faculty oversight preserved for nuanced assessment of communication skills central to surgical informed consent proficiency.

## Declarations

## Conflict of Interest Statement

The authors declare no conflicts of interest.

# Funding Statement

This research received no external funding.

# Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

# References

1. Case L, Khan I, Qato K. The past, present, and future of feedback in medical education. *JVS-Vasc Insights Open Access Publ Soc Vasc Surg*. 2024;2. 10.1016/j.jvsvi.2024.100116.
2. Wu C, Williams W, Reich AJ, Nitzschke SL. A Tale of Two Sides: Exploring Faculty and Resident Perspectives on Feedback by Qualitative Analysis. *J Surg Res*. 2025;308:233–42. 10.1016/j.jss.2025.02.023.
3. Wolpaw J, Saddawi-Konefka D, Dwivedi P, Toy S. Faculty Underestimate Resident Desire for Constructive Feedback and Overestimate Retaliation. *J Educ Perioper Med JEPM*. 2019;21(4):E634.
4. Alomar AZ. Confidence level, challenges, and obstacles faced by orthopedic residents in obtaining informed consent. *J Orthop Surg*. 2021;16:390. 10.1186/s13018-021-02531-1.
5. McKendy KM, Watanabe Y, Lee L, et al. Perioperative feedback in surgical training: A systematic review. *Am J Surg*. 2017;214(1):117–26. 10.1016/j.amjsurg.2016.12.014.
6. Braddock CH, Edwards KA, Hasenberg NM, Laidley TL, Levinson W. Informed decision making in outpatient practice: time to get back to basics. *JAMA*. 1999;282(24):2313–20. 10.1001/jama.282.24.2313.
7. Bean S, Magwood B, Abdoh AA, Chen J, Hochman J. Informed consent: exploring surgical residents' beliefs, attitudes and practices. *Healthc Q*. 2010;13(4):68–73. 10.12927/hcq.2013.22001.
8. Long KL, Ingraham AM, Wendt EM, et al. Informed Consent and Informed Decision-Making in High-Risk Surgery: A Quantitative Analysis. *J Am Coll Surg*. 2021;233(3):337–45. 10.1016/j.jamcollsurg.2021.05.029.
9. Lamb M, Woodward JM, Quaranto B, et al. Do Interns Learn On-The-Job How to Obtain Proper Informed Consent for Surgical Procedures? *J Surg Educ*. 2024;81(9):1215–21. 10.1016/j.jsurg.2024.06.005.
10. Braddock C, Hudak PL, Feldman JJ, Bereknyei S, Frankel RM, Levinson W. Surgery is certainly one good option: quality and time-efficiency of informed decision-making in surgery. *J Bone Joint Surg Am*. 2008;90(9):1830–8. 10.2106/JBJS.G.00840.
11. Thompson BM, Sparks RA, Seavey J, et al. Informed consent training improves surgery resident performance in simulated encounters with standardized patients. *Am J Surg*. 2015;210(3):578–84.

- 10.1016/j.amjsurg.2014.12.044.
12. Angelos P, DaRosa DA, Bentram D, Sherman H. Residents seeking informed consent: Are they adequately knowledgeable? *Curr Surg*. 2002;59(1):115–8. 10.1016/s0149-7944(01)00591-8.
  13. Hall DE, Hanusa BH, Fine MJ, Arnold RM. Do surgeons and patients discuss what they document on consent forms? *J Surg Res*. 2015;197(1):67–77. 10.1016/j.jss.2015.03.058.
  14. 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*. 2024;24:1391. 10.1186/s12909-024-06399-7.
  15. Vaira LA, Lechien JR, Maniaci A, et al. Evaluating AI-Generated informed consent documents in oral surgery: A comparative study of ChatGPT-4, Bard gemini advanced, and human-written consents. *J Cranio-Maxillofac Surg*. 2025;53(1):18–23. 10.1016/j.jcms.2024.10.002.
  16. Ali R, Connolly ID, Tang OY, et al. Bridging the literacy gap for surgical consents: an AI-human expert collaborative approach. *Npj Digit Med*. 2024;7(1):63. 10.1038/s41746-024-01039-2.
  17. Cvijic L, Denecke K. SurgiBot: Large Language Model-Based Patient Simulator for Surgical Informed Consent Training. *Stud Health Technol Inf*. 2026;336:2250–1. 10.3233/SHTI260661.
  18. Talwalkar JS, Chartash D, Zhang L, et al. Ambient AI Scribes to Create Educational Feedback Notes for Medical Students: Randomized Trial. *JMIR Med Educ*. 2026;12:e89996. 10.2196/89996.
  19. Wu S, Tang M, Liu J, et al. Impact of an AI-based laparoscopic cholecystectomy coaching program on the surgical performance: a randomized controlled trial. *Int J Surg Lond Engl*. 2024;110(12):7816–23. 10.1097/JS9.0000000000001798.
  20. Schneider N, VALIDITY EVIDENCE FOR THE INTERPRETATION AND USE OF ESSENTIAL ELEMENTS, OF COMMUNICATION GLOBAL RATING SCALE SCORES. *Individ Fam Community Educ ETDs*. Published online February 1, 2016. [https://digitalrepository.unm.edu/educ\\_ifce\\_etds/34](https://digitalrepository.unm.edu/educ_ifce_etds/34)
  21. Roberts LW, Mines J, Voss C, et al. Assessing medical students' competence in obtaining informed consent. *Am J Surg*. 1999;178(4):351–5. 10.1016/s0002-9610(99)00178-6.
  22. Anderson TN, Aalami LR, Lee EW, et al. Perception and confidence of medical students in informed consent: A core EPA. *Surgery*. 2020;167(4):712–6. 10.1016/j.surg.2019.11.012.
  23. Saed A. A Systematic Review of the Medical Student Feedback on Undergraduate Surgical Education During the Pandemic. *Cureus*. 2022;14. 10.7759/cureus.30440.
  24. Pang JH, Finlay E, Fortner S, Pickett B, Wang ML. Teaching Effective Informed Consent Communication Skills in the Virtual Surgical Clerkship. *J Am Coll Surg*. 2021;233(1):64–e722. 10.1016/j.jamcollsurg.2021.04.026.
  25. Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ*. 2008;42(8):816–22. 10.1111/j.1365-2923.2008.03105.x.
  26. Sheikh N, Mehta J, Shah R, Brydges R. Feedback that Lands: Exploring How Residents Receive and Judge Feedback During Entrustable Professional Activities. *Perspect Med Educ*. 2023;12(1):427–37. 10.5334/pme.1020.

27. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual Patients Using Large Language Models: Scalable, Contextualized Simulation of Clinician-Patient Dialogue With Feedback. J Med Internet Res. 2025;27:e68486. 10.2196/68486.

28. Jarry Trujillo C, Vela Ulloa J, Escalona Vivas G, et al. Surgeons vs ChatGPT: Assessment and Feedback Performance Based on Real Surgical Scenarios. J Surg Educ. 2024;81(7):960–6. 10.1016/j.jsurg.2024.03.012.

Figures

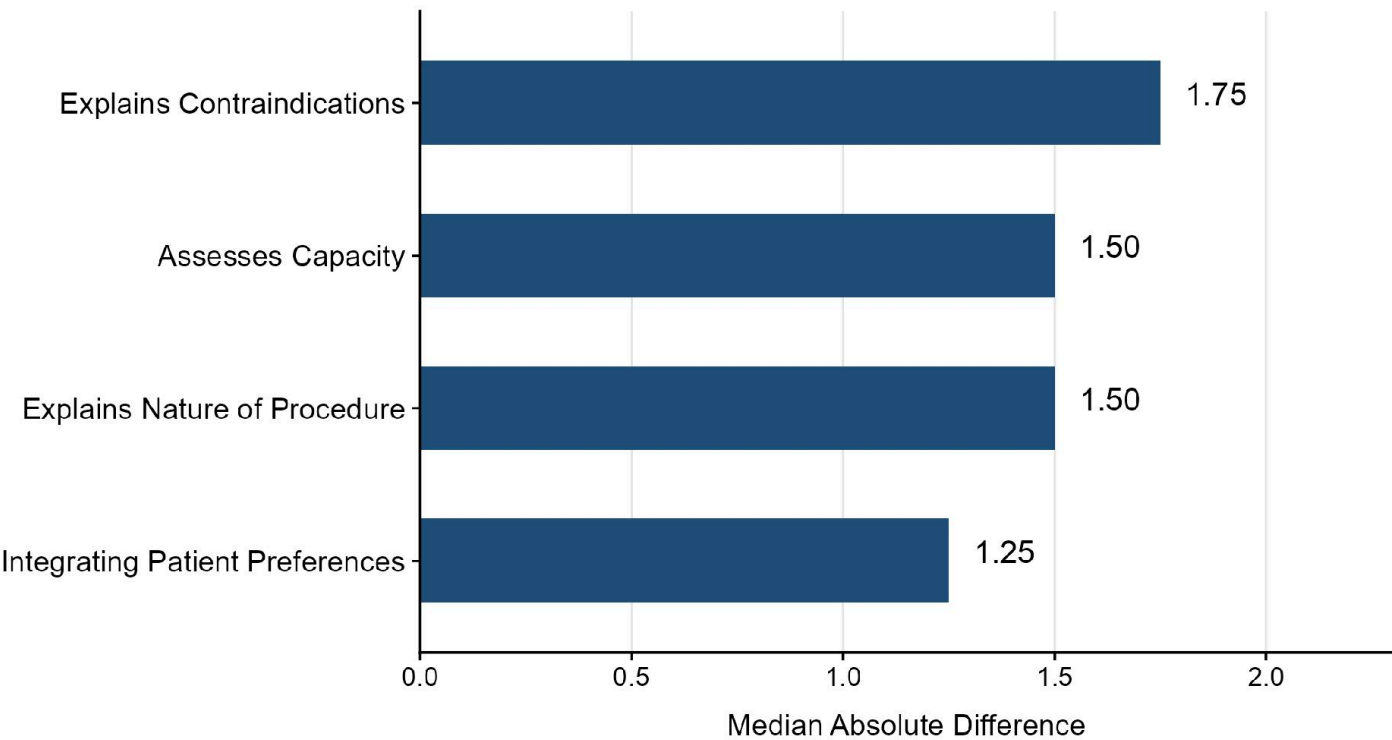


Figure 1

Median Absolute Differences in Scores for Select Informed Consent Rubric Items.

Median absolute differences between the AI-assigned score and the median score assigned by the four surgeon evaluators for selected informed consent rubric items that demonstrated the greatest discordance. Larger values indicate greater AI-surgeon discordance.

