

# The Journal of Education in Perioperative Medicine

ORIGINAL RESEARCH

## A Comparative Analysis of Human and AI-Augmented Feedback in Anesthesiology Education

DANIEL J. ROSENKRANS, MD  
KATHERINE B. OWENSBY, MPhil  
JUSTIN C. MAGIN, BS

MORGAN N. MCCAIN, BS  
EMILY G. TEETER, MD

SAMUEL N. BLACKER, MD  
FEI CHEN, PhD, MED, MSTAT

### INTRODUCTION

Feedback remains an important element of graduate medical education with increasing attention as a critical component of competency-based growth.<sup>1-3</sup> Despite developing various models and strategies to enhance feedback delivery, providing consistent, high-quality feedback remains challenging.<sup>4-7</sup> Multiple factors complicate the process, including discomfort providing constructive feedback, limited training in effective techniques, and inherent predispositions such as gender biases.<sup>8-11</sup> These challenges create opportunities to explore novel methods of enhancing feedback, potentially through innovative technological tools that mitigate these barriers.

Generative artificial intelligence (AI), particularly consumer large language models such as ChatGPT, is an emerging area of exploration in medical education.<sup>12-15</sup> Considering its accessibility and user-friendly interface, AI presents a promising avenue for assisting faculty in structuring and composing feedback and has demonstrated the ability to create feedback for residents on par with experienced physicians.<sup>16</sup> Whereas this provides evidence of AI's ability to generate feedback *de novo*, its role in supporting the delivery of feedback—specifically, how faculty might engage with

AI as part of preparing or structuring their own feedback—remains unexplored.

This study examined the impact of reviewing AI feedback examples on the quality of attending anesthesiologist feedback as assessed by anesthesiology resident evaluators. The research team hypothesized that the feedback from attendings would be rated higher after reading the AI feedback examples. Additionally, attending anesthesiologists' perspectives on using feedback generated from AI were explored, shedding light on its feasibility and utility in clinical practice.

### MATERIALS AND METHODS

The Institutional Review Board at the University of North Carolina reviewed and determined to exempt this study due to its minimal risk and focus on educational practices (Study # 23-3024). Written informed consent was obtained for all participants.

Recruitment for the study began and finished in January 2024. Attending anesthesiologists and residents were recruited through direct communication within the anesthesiology residency program at the University of North Carolina. The study employed a crossover (within-subjects) design, in which each participant provided a feedback sample without AI to serve as control (human-only) and a sample after reading an AI feedback example (AI-

augmented). Accordingly, the sample size was determined using an *a priori* power analysis for a paired samples *t*-test. Based on our hypothesis of an average score difference of 5 and assuming a conservative correlation between the 2 conditions, the analysis indicated that a sample size of 14 participants was required to achieve 90% statistical power ( $\alpha = 0.05$ , 2-tailed). Our final sample size was 30 attending anesthesiologists. Six resident anesthesiologists participated in the study to evaluate the feedback samples. Attending demographic information was obtained including age, gender, years in practice, subspecialty training, practice setting, comfort level with feedback, familiarity with AI, and prior formal feedback training. Resident demographic information obtained included age, gender, level of training, prior formal training in receiving or providing feedback, experience with AI technologies in medical education or practice, familiarity with AI, and preferred ways to receive feedback.

Two clinical vignettes were developed to reflect scenarios commonly encountered in resident training. The vignettes were developed by the research team, which had significant experience in resident education, assessment, and remediation. Each vignette depicted a written example of an observation of resident performance.

*continued on next page*

*continued from previous page*

One vignette focused on an unprepared resident, meant to prompt more objective feedback (preparedness vignette), whereas the other focused on a resident with professionalism issues requiring more subjective feedback (professionalism vignette). Each vignette underwent iterative changes to enhance accuracy and relevance while minimizing potential sources of bias. Specifically, extraneous details that could cue gender, race, or other demographic assumptions were removed, and all descriptions were limited to the performance issues being observed. Names and personal attributes were intentionally excluded to maintain neutrality and ensure that the focus remained solely on the resident's performance.

The vignettes were inserted verbatim as prompts to the AI model (ChatGPT, version 4.0), and the model's responses recorded for the AI feedback examples. This approach limited variation in AI-generated responses that could be attributed to differences in prompting and standardized feedback examples each attending received. The AI feedback examples were reviewed by the research team to ensure accuracy and relevance to the vignettes, minimizing the risk of factual hallucination from the AI model.

Attending feedback collection took place from February to April 2024 using virtual interviews. Members of the research team conducted the same procedure for each interview. Attending anesthesiologists were randomly assigned to 1 of 2 series of vignettes: 1 group was asked to provide feedback on the preparedness vignette first, followed by the professionalism vignette (series A), whereas the other group provided feedback on the professionalism vignette, followed by the preparedness vignette (series B). Figure 1 depicts the vignette assignment of the attendings. Randomizing attendings to different sequences allowed for an even distribution of feedback with and without AI prompts for both vignettes, thereby minimizing any effect of the order by which the attendings provided feedback on the vignettes.

During the interview, attendings read the vignette and provided feedback as if speaking directly to a resident. To

standardize the encounter, interviewers were instructed to listen to the feedback silently without interruptions and limit any nonverbal cues. The feedback from this encounter was labeled "human-only" in that the feedback generated was without any AI example for assistance. Before delivering feedback on the second vignette, each attending was given a written AI feedback example corresponding to the next vignette and allowed to incorporate it into the attending's feedback provision. This was done to simulate a naturalistic use of AI whereby an attending prompts an AI model for feedback suggestions prior to the feedback encounter with the resident. As with the first vignette, attendings were asked to provide feedback as if speaking directly to a resident using any notes deemed necessary. The feedback from the second encounter was labeled "AI-augmented" in that the feedback generated came after attendings read the AI example. The interview concluded by asking each attending 2 questions to assess their experience with the AI feedback:

1. Do you feel like the AI-generated feedback helped you, or do you feel like it didn't make an impact?
2. Do you feel like this is a tool you might use in the future if it were easily available?

Interviews were recorded for data extraction. Interview recordings were securely stored and assigned a unique identifier to protect the attending's identity. Written transcripts of the recordings were created using Zoom Video Communications integrated transcript software. All transcripts were reviewed for consistency with the original recorded interview, and any identified discrepancies were reconciled. Minor grammatical adjustments were allowed to enhance clarity for the written review.

Six volunteer residents were randomly assigned to evaluate 20 attending feedback transcripts for a total of 120 evaluations. Each attending feedback received 2 evaluations from 2 different residents to assess the interrater reliability. The residents also rated both AI feedback examples. Residents were blinded to the source of the feedback, including whether the attending was exposed to the AI example prior to generating the

attending's feedback. This design allowed for a comparison of ratings between the attending feedback before reading the AI example (human-only) and attending feedback after reading the AI example (AI-augmented).

Residents rated the feedback using a standardized grading rubric. The rubric included assessment domains adapted from a validated tool for providing feedback to residents in anesthesiology.<sup>17,18</sup> Each domain was rated using a 4-point Likert scale from 0 to 3, where 0 was *poor*, and 3 was *excellent*. Residents also rated the overall effectiveness of the feedback, serving as a holistic assessment, using a similar Likert scale from 0 to 3. Finally, they indicated the perceived origin of the feedback: human, AI, or unsure.

Completed evaluations were stored electronically with unique identifiers assigned to the resident evaluators to protect their identity.

## STATISTICAL ANALYSIS

Descriptive statistics were calculated for both attending and resident demographics, including frequencies, percentages, means, and standard deviations, as appropriate. Multiple measures from each attending were combined using the mean score. To assess the impact of AI augmentation on the quality of feedback provided, a Wilcoxon signed rank test with continuity correction compared resident ratings of feedback quality between human-only and AI-augmented feedback. The quality of feedback was measured by ratings on the following domains: detail, specificity, behavior focused, constructiveness, actionability, and overall effectiveness.<sup>17,18</sup> To examine associations between the perceived origin of feedback, a chi-square test of independence was used. Analysis of variance or *t*-test was used to investigate the associations between demographic covariates and resident feedback ratings for categorical variables (gender and prior feedback training). Pearson correlation was applied to continuous covariates (age and years as attending). Interrater reliability was assessed to evaluate the consistency of resident ratings across the standardized grading rubric. A 2-way random-effects intraclass correlation coefficient (ICC)

*continued on next page*

*continued from previous page*

was calculated for absolute agreement for each feedback dimension. Statistical significance was set at a threshold of  $\alpha < .05$ . The analysis was completed using R 4.4.0 (R Core Team, 2024).

## QUALITATIVE ANALYSIS

The qualitative analysis aimed to provide a more nuanced understanding of how and in what circumstances the participating attendings considered AI helpful. To this end, a content analysis approach to systematically code and identify themes within the transcript data was employed.<sup>19</sup> Blinded to the resident ratings, 2 authors reviewed deidentified transcript responses in aggregate. After initial memoing to record trends, the data was uploaded into Dedoose (version 9.0.17, 2024) for coding and analysis. An iterative approach, including inductive and deductive coding, was implemented to allow themes to emerge.<sup>20</sup> Three authors revisited the data, creating and assigning codes while rereading each transcript response. After completing an initial codebook based on all responses, each code definition was reviewed to ensure it captured distinctive data elements. Codes were then reorganized to better reflect themes that had emerged through memoing. To verify that all codes had been applied consistently, 2 additional authors independently reviewed the codebook, transcript responses, and the associated codes. Disagreements were discussed and resolved. Results emerging from the code co-occurrence table in Dedoose were recorded.

## RESULTS

Attending participant demographics are summarized in Table 1. Thirty attending anesthesiologists participated in the study with a median age of 42 years. The median years of experience as an attending anesthesiologist was 7.5 years. Whereas 40% had prior training in providing feedback, only 23.3% reported being “comfortable” with giving feedback with a notable percentage reporting some discomfort. Familiarity with AI was limited with 60% possessing only basic awareness of AI applications and 93.3%

not previously using AI for providing feedback.

Resident participant demographics are summarized in Table 2. Six resident anesthesiologists evaluated the feedback with a median age of 30.5 years. None had prior formal training in receiving or providing feedback or exposure to AI in their medical education. Most residents expressed a preference for written feedback though verbal, electronic, and in-person feedback methods were also commonly endorsed.

## COMPARATIVE ANALYSIS

The results comparing human-only versus AI-augmented feedback ratings are summarized in Table 3. Scores were similar between human-only and AI-augmented feedback across all domains assessed. Comparing the feedback ratings on the different types of vignettes, preparedness feedback received significantly higher ratings than professionalism feedback for multiple domains, including detail ( $p = .01$ ), specificity ( $p = .01$ ), constructiveness ( $p = .02$ ), actionability ( $p = .01$ ), and overall effectiveness ( $p = .02$ ). These results are summarized in Table 4.

Interrater reliability for resident ratings varied by domain with moderate reliability for detail (ICC = .502,  $p = .004$ ) and overall effectiveness (ICC = .514,  $p = .004$ ), but lower reliability for specificity (ICC = .361,  $p = .046$ ), behavior focused (ICC = .160,  $p = .26$ ), constructiveness (ICC = .384,  $p = .03$ ), and actionability (ICC = .443,  $p = .01$ ).

Residents could not reliably discern the source of feedback with no significant association between perceived feedback origin and the use of AI augmentation ( $\chi^2 = .92$ ,  $p = .63$ ).

Attending demographics and characteristics were analyzed for associations with resident ratings. There were no significant differences in resident feedback ratings based on the attending's age, gender, and prior feedback training across all evaluated domains. Years of experience as an attending anesthesiologist showed no significant correlation with resident ratings for human-only feedback. Interestingly, for AI-augmented feedback, a weak negative correlation between

years of experience and evaluations were observed in the domains of constructiveness (correlation =  $-.40$ ,  $p = .03$ ) and actionability (correlation =  $-.39$ ,  $p = .03$ ), indicating that more experienced attendings were likely to perform worse with AI-augmentation in these domains.

## QUALITATIVE ANALYSIS

The transcript data from the attendings demonstrated an overall positive perception toward AI augmentation with many participants conveying enthusiasm for AI-generated feedback (13 instances). Some participants even expressed pleasant surprise at the implementation of AI during the study (5 instances). These sentiments can be summarized by attending 0.454, who remarked, “I’m actually really surprised. I didn’t realize it was going to be that insightful. I think it was a huge help.”

Within the qualitative data, 30 mentions of the benefits of AI-generated feedback were made. Most commonly, participants commented on its created content (19 instances), the framework that AI provided (11 instances), and the overall quality of its output (8 instances). Under the umbrella of created content, participants further identified AI’s examples, specific phrases, and checklists as relevant and helpful features.

In contrast, only 12 mentions of the drawbacks of AI augmentation appeared in the transcript data. In these responses, participants most often discussed the incongruity between the AI-generated content and the feedback that they would have given. For example, 1 participant (attending 0.055) commented, “It was far too sugary sweet and did not recognize the gravity of the deviation from a standard of care and the potential danger that puts the patient in.” Other individuals commenting on the drawbacks of AI-generated feedback considered it to be too detailed and/or disingenuous for face-to-face interactions.

Twenty-eight of the 30 participants would use AI for feedback if it were easily available (93.3%). Participant responses indicated a clear preference pattern for AI augmentation in the professionalism

*continued on next page*

*continued from previous page*

vignettes instead of preparedness vignettes. Here, the co-occurrence of the codes “professionalism” or “preparedness” and “when to use” or “when not to use” was investigated.

Attendings noted that they would likely use AI as a point of reference and for content generation before giving feedback to a resident. They remarked that consulting AI would be most helpful for preplanned conversations (7 co-occurrences of “when to use” × “planned conversation”) and for challenging feedback situations (6 co-occurrences of “when to use” × “challenging situation”). These trends closely align with subjective conversations surrounding a resident’s professionalism as exhibited by a response from attending 0.252: “Something like professionalism or interpersonal issues, where you have to be really careful about how you approach it, it [AI-generated feedback] may be a good rubric for sort of how to format that discussion, at least give you some ideas to get you started on thinking how you want to format the feedback.” A complete list of the codebook used in this analysis, including representative quotes and themes, is available online as supplemental material from the corresponding author.

## DISCUSSION

In this randomized crossover study, we found that reviewing an AI-generated feedback sample prior to providing feedback did not lead to a statistically significant improvement in feedback quality as rated by blinded resident evaluators. However, the findings offer several important insights into the evolving role of AI in anesthesiology education. Notably, our domain-level analysis suggests that AI performed particularly well in areas emphasizing behavior-focused and constructivist feedback (Table 3). These domains represent critical components of high-quality feedback yet are often challenging for faculty to implement consistently. AI’s strength in producing structured, actionable phrasing may help translate faculty observations into learner-centered coaching statements. This finding underscores the potential role of AI as

both a scaffold for individual feedback encounters and as a resource for faculty development, modeling how to move from general impressions toward constructive language.

The intervention was intentionally designed to replicate a real-world workflow, whereby an attending anesthesiologist seeking guidance on giving feedback to a resident referenced a consumer generative AI model prior to speaking with the trainee. Many participants reported that the AI-generated examples were helpful, especially the content and structure the model provided. Moreover, the participants identified specific cases in which using the AI model would be beneficial, including articulating feedback in sensitive or ambiguous areas. Whereas attendings may have benefited from reviewing feedback from more experienced colleagues, the consumer AI model offers a democratized experience that any anesthesiologist can access regardless of their institutional resources.

Yet the perceived benefits of using AI were not reflected in the quality ratings of the residents. Several factors may explain this discrepancy. We did not assess the degree to which attendings incorporated the AI feedback examples into their final provision, which likely varied among the participants. It is also possible that reading a fixed AI response may not capture AI’s potential effect on feedback quality. Instead, it may be more beneficial to interact with the model, allowing for personalized prompt engineering, clarification requests, and rehearsing feedback prior to interacting with the trainee. In doing so, attendings may better incorporate the AI feedback into their own provision. Altogether, these findings underscore that, whereas AI shows potential as a tool for supporting feedback, it remains in its early stages. Further refinement using AI, including interactive and tailored approaches, are necessary to optimize its role in feedback delivery.

The difference in feedback quality between vignette types, higher for preparedness than for professionalism, may highlight a persistent challenge of delivering feedback in subjective domains. Professionalism feedback involves nuanced interpersonal dynamics, often making it more difficult

to construct and deliver. Although our qualitative analysis suggests that AI may offer useful language for these types of challenging conversations, we did not observe a significant interaction between AI use and improved professionalism feedback. Further investigation is warranted to determine the optimal strategies for integrating AI-generated feedback into clinical education.

Our data indicate that residents could not accurately discern AI-only (verbatim responses from the AI model) and AI-augmented feedback from human-only feedback, suggesting that AI can be integrated without compromising the perceived legitimacy of feedback. This finding is essential for the acceptability of AI in educational settings as it indicates that attendings may use AI without disrupting the traditional attending–resident dyad.

This study enlisted a diverse group of attendings with a broad range of ages and experience. Nearly half of the attendings had prior training in providing feedback, yet the training did not correlate with higher feedback ratings. Interestingly, our data also revealed that greater years of experience as an attending were weakly but significantly correlated with lower rating in constructiveness and actionability when AI augmentation was used. On interpretation is that experienced faculty, with established heuristics and ingrained feedback habits, may be less inclined to integrate AI phrasing, whereas junior faculty—still refining their feedback repertoire—may be more receptive to adopting structured AI-generated suggestions. This suggests AI assistance may be particularly beneficial for early career faculty as a scaffold, whereas for more experienced attendings, its utility may lie in reflective comparison, prompting reconsideration of long-standing feedback styles.

Looking ahead, we envision AI as an adjunct rather than a replacement for faculty feedback. A practical model would involve attendings entering brief observations or structured notes into an AI platform, receiving a draft feedback summary, and then tailoring the output before sharing it in a dialogue with the

*continued on next page*

*continued from previous page*

resident. We do not advocate handing learners an unedited AI report. Effective feedback depends on many things, including timing and relational trust. Instead, AI may best serve as a rapid drafting tool that accelerates the creation of high-quality feedback while preserving the educational alliance through faculty-delivered conversations.

This study has several limitations that warrant consideration. The sample size was relatively small, particularly among resident raters. Future studies with larger cohorts would provide a more comprehensive evaluation of AI's impact on feedback quality. Our intervention of reading an AI example prior to providing feedback may not fully capture AI's role in feedback enhancement. Additionally, there was no formal process to assess how much AI-generated feedback was incorporated into the attending feedback provision. Future studies that allow for interaction with the AI-model and measure its incorporation would provide a clearer understanding of AI's impact on feedback quality. The interrater reliability was generally low across most feedback domains despite being adapted from previously validated frameworks. This was likely due to the absence of formal rater training and calibration on the assessment rubric. The feedback collected in this study was provided in a written format without the natural interaction present in a typical dialogic encounter. In clinical settings, feedback often incorporates verbal and nonverbal cues that can influence its reception. The absence of these interactive elements may have impacted the perception and quality of feedback. Our questions at the end of the attending interviews used in the qualitative analysis lacked a negative response option, and this may have limited participants' ability to express difficulties or frustration with the technology. Finally, this study was conducted in a single academic center, and

this may limit the generalizability of its results to other institutions or specialties. AI's role in feedback provision may vary across different educational environments, necessitating multicenter and multidisciplinary studies to understand its broader applicability.

In summary, this study highlights both the promise and current limitations of AI in supporting faculty feedback in anesthesiology education. Exposure to AI examples did not significantly enhance attending-provided feedback, yet AI may serve as a valuable scaffold to help faculty structure feedback, particularly in professionalism domains. Future work should examine more interactive and tailored AI approaches as well as the impact of AI-generated feedback on resident performance rather than perception to better define how AI may be effectively integrated into competency-based medical education.

### Acknowledgements

The authors thank the attending anesthesiologists and residents who generously contributed their time to submit and evaluate feedback, making this study possible. They also thank Dr Xinming An, who provided invaluable assistance with the statistical analysis.

### References

1. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366(11):1051-6.
2. Ekpenyong A, Zetkolic M, Edgar L, Holmboe ES. Reimagining feedback for the milestones era. *J Grad Med Educ*. 2021;13(suppl 2):109-12.
3. Bing-You R, Hayes V, Varaklis K, et al. Feedback for learners in medical education: what is known? A scoping review. *Acad Med*. 2017;92(9):1346-54.
4. Telio S, Ajjawi R, Regehr G. The “educational alliance” as a framework for reconceptualizing feedback in medical education. *Acad Med*. 2015;90(5):609-14.
5. Sargeant J, Lockyer JM, Mann K, et al. The R2C2 model in residency education: how does it foster coaching and promote feedback use? *Acad Med*. 2018;93(7):1055-63.
6. Yarris LM, Linden JA, Gene Hern H, et al. Attending and resident satisfaction with feedback in the emergency department. *Acad Emerg Med*. 2009;16(suppl 2):S76-81.
7. Branfield Day L, Miles A, Ginsburg S, Melvin L. Resident perceptions of assessment and feedback in competency-based medical education: a focus group study of one internal medicine residency program. *Acad Med*. 2020;95(11):1712-7.
8. Mitchell JD, Holak EJ, Tran HN, et al. Are we closing the gap in faculty development needs for feedback training? *J Clin Anesth*. 2013;25(7):560-4.
9. Mueller AS, Jenkins TM, Osborne M, et al. Gender differences in attending physicians' feedback to residents: a qualitative analysis. *J Grad Med Educ*. 2017;9(5):577-85.
10. Klein R, Ufere NN, Rao SR, et al. Association of gender with learner assessment in graduate medical education. *JAMA Netw Open*. 2020;3(7):e2010888.
11. Watling CJ, Kenyon CF, Schulz V, et al. An exploration of faculty perspectives on the in-training evaluation of residents. *Acad Med*. 2010;85(7):1157-62.
12. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. *Acad Med*. 2021;96(11S):S62-70.
13. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. *Acad Med*. 2024;99(1):22-7.
14. Patel SJ, Notarianni AP, Martin AK, et al. The year in graduate medical education: selected highlights from 2023. *J Cardiothorac Vasc Anesth*. 2024;38(12):2906-14.
15. Mangold S, Ream M. Artificial intelligence in graduate medical education applications. *J Grad Med Educ*. 2024;16(2):115-8.
16. 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.
17. Mitchell JD, Ku C, Diachun CAB, et al. Enhancing feedback on professionalism and communication skills in anesthesia residency programs. *Anesth Analg*. 2017;125(2):620-31.
18. Walsh DP, Chen MJ, Buhl LK, Neves SE, Mitchell JD. Assessing interrater reliability of a faculty-provided feedback rating instrument. *JMECD*. 2022;9:1-8.
19. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res*. 2005;15(9):1277-88.
20. Bingham AJ. From data management to actionable findings: a five-phase process of qualitative data analysis. *Int J Qual Methods*. 2023;22.

*continued on next page*

*continued from previous page*

The authors are in the Department of Anesthesiology, University of North Carolina School of Medicine in Chapel Hill, NC: **Daniel J. Rosenkrans** is an assistant professor of anesthesiology; **Katherine B. Owensby** is a research assistant; **Justin C. Magin** is a medical student; **Morgan N. McCain** is a medical student; **Emily G. Teeter** is a professor of anesthesiology; **Samuel N. Blacker** is an associate professor of anesthesiology; **Fei Chen** is an associate professor of anesthesiology.

**Corresponding author:** Daniel J. Rosenkrans, MD, Department of Anesthesiology, University of North Carolina School of Medicine, 101 Manning Drive, Chapel Hill, NC 27599. Telephone: (919) 966-5136, Fax: (984) 974-4873

**Email Address:** Daniel J. Rosenkrans: [Daniel\\_rosenkrans@med.unc.edu](mailto:Daniel_rosenkrans@med.unc.edu)

**Financial disclosure:** None.

#### Abstract

**Background:** High-quality feedback is essential for resident development, yet barriers to its provision persist. Artificial intelligence (AI) offers a promising tool to augment feedback, potentially addressing these challenges. This study explored whether exposing attending anesthesiologists to AI-generated feedback samples before crafting their own feedback improved its quality as rated by residents.

**Methods:** Thirty attending anesthesiologists provided feedback on 2 vignettes addressing issues of preparedness and professionalism. Before delivering feedback

on the second vignette, attendings reviewed an AI-generated feedback sample. Utilizing a validated rubric, 6 blinded residents randomly evaluated the quality of feedback across 2 conditions: attendings alone (human-only) and after attendings reviewed the AI sample (AI-augmented). Feedback ratings and vignette types were compared. Residents were assessed if they could correctly identify the origin of feedback. A qualitative analysis explored attendings' perceptions of using AI for feedback.

**Results:** AI augmentation did not significantly improve feedback quality ratings compared with human-only feedback ( $p = .7$ ). Preparedness feedback was rated higher than professionalism feedback ( $p = .02$ ). Residents could not reliably distinguish the use of AI for feedback ( $\chi^2 = .92, p = .63$ ). Attendings reported that AI provided helpful structure and phrasing, particularly for professionalism issues, and would use it if readily available.

**Conclusions:** Although AI augmentation did not significantly improve feedback quality, it showed promise as a tool for supporting feedback provision, particularly with crafting feedback on nontechnical, more subjective issues. Additional studies are needed to better understand AI as a tool for feedback enhancement.

**Keywords:** Artificial intelligence, feedback, anesthesiologists, internship and residency, education

*continued on next page*

continued from previous page

## Figure

**Figure 1.** Attending assignments. Attendings were randomly assigned to 2 different series of vignettes to prompt feedback samples. Attendings assigned to series A provided feedback on the preparedness vignette first without reading any AI sample before their feedback provision. This feedback sample served as a control and was labeled human-only feedback. Series A attendings then provided feedback on the professionalism vignette after reading the corresponding AI feedback sample for the professionalism vignette. This feedback provision was labeled AI-augmented feedback. Attendings assigned to series B provided feedback the same sequence, human-only followed by AI-augmented with the only difference being the order of the vignettes. Series B attendings provided feedback on the professionalism vignette first, followed by the preparedness vignette.

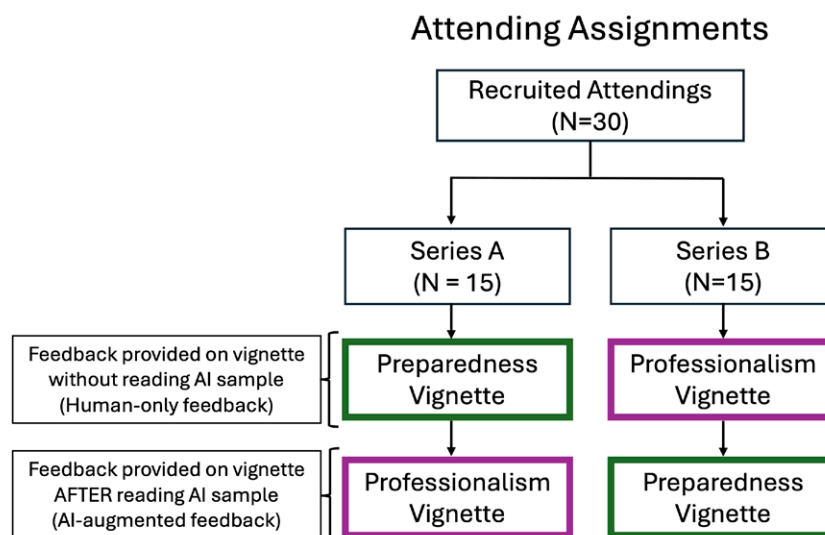


Figure 1: Attendings were randomly assigned to two different series of vignettes to prompt feedback samples. Attendings assigned to Series A provided feedback on the Preparedness vignette first without reading any AI sample before their feedback provision. This feedback sample served as a control and was labeled Human-only feedback. Series A attendings then provided feedback on the Professionalism vignette after reading the corresponding AI feedback sample for the Professionalism vignette. This feedback provision was labeled AI-augmented feedback. Attendings assigned to Series B provided feedback the same sequence, Human-only followed by AI-augmented, with the only difference being the order of the vignettes. Series B attendings provided feedback on the Professionalism vignette first, followed by the Preparedness vignette.

continued on next page

continued from previous page

## Tables

**Table 1.** Demographic Characteristics of Attending Anesthesiologists That Provided Feedback

Attending Anesthesiologist Demographic Information (N = 30)	
Age	
Mean (SD)	41.1 (6.27)
Median [min, max]	42.0 [31.0, 55.0]
Gender	
Female	11 (36.7%)
Male	19 (63.3%)
Years as attending	
Mean (SD)	8.37 (5.67)
Median [min, max]	7.50 [1.00, 20.0]
Practice setting	
Academic	26 (86.7%)
Community	1 (3.3%)
Hybrid (combination of academic and community)	3 (10.0%)
Subspecialization training	
No	6 (20.0%)
Yes	24 (80.0%)
Prior training in giving feedback	
No	18 (60.0%)
Yes	12 (40.0%)
Comfort providing feedback	
Comfortable	7 (23.3%)
Somewhat comfortable	14 (46.7%)
Neither comfortable nor uncomfortable	3 (10.0%)
Somewhat uncomfortable	4 (13.3%)
Uncomfortable	2 (6.7%)
Familiarity with AI	
Not familiar <sup>a</sup>	6 (20.0%)
Basic awareness <sup>b</sup>	18 (60.0%)
Moderate familiarity <sup>c</sup>	3 (10.0%)
Experienced user <sup>d</sup>	3 (10.0%)
Using AI for feedback	
No	28 (93.3%)
Yes	2 (6.7%)

<sup>a</sup> Not familiar: No exposure or minimal exposure to AI technologies, limited understanding of AI applications.

<sup>b</sup> Basic awareness: Some general knowledge about AI and its potential applications, aware of AI models but not necessarily familiar with specific applications.

<sup>c</sup> Moderate familiarity: Familiar with the concept and applications of AI, may have attended workshops or training sessions on AI.

<sup>d</sup> Experienced user: Actively uses AI tools or technologies in daily life or practice.

continued on next page

continued from previous page

## Tables continued

**Table 2.** Demographic Characteristics of Resident Anesthesiologists That Provided Quality Ratings Over Attending Anesthesiologists' Feedback

Resident Anesthesiologist Demographic Information (N = 6)	
Age	
Mean (SD)	31.3 (3.88)
Median [min, max]	30.5 [28, 38]
Gender	
Female	2 (33.3%)
Male	4 (66.7%)
Level of training	
PGY2	2 (33.3%)
PGY3	3 (50.0%)
PGY4	1 (16.7%)
Prior formal training in receiving or providing feedback?	
Yes	0 (0%)
No	6 (100%)
Prior exposure or experience with AI technologies in your medical education or practice?	
Yes	0 (0%)
No	6 (100%)
Familiarity with AI	
Not familiar <sup>a</sup>	2 (33.3%)
Basic awareness <sup>b</sup>	4 (66.7%)
Preferred ways to receive feedback	
Verbal	4 (67.7%)
Written	5 (83.3%)
Electronic	4 (67.7%)
In-person	4 (67.7%)

<sup>a</sup> Not familiar: No exposure or minimal exposure to AI technologies, limited understanding of AI applications.

<sup>b</sup> Basic awareness: Some general knowledge about AI and its potential applications, aware of AI models but not necessarily familiar with specific applications.

continued on next page

continued from previous page

## Tables continued

**Table 3. Human-Only Versus AI-Augmented Feedback**

	Human-only N = 30 <sup>a</sup>	AI-augmented N = 30 <sup>a</sup>	p-value <sup>b</sup>	Effect Size (90% CI)	Interpretation
Detail	1.88 (0.58)	1.90 (0.77)	> .9	-0.02 (-0.32, 0.28)	very small
Specificity	1.97 (0.67)	2.07 (0.72)	.5	-0.10 (-0.40, 0.20)	very small
Behavior focused	1.90 (0.61)	2.15 (0.65)	.08 <sup>c</sup>	-0.32 (-0.62, -0.01)	small
Constructiveness	1.82 (0.84)	2.02 (0.68)	.3	-0.20 (-0.50, 0.10)	small
Actionability	1.82 (0.79)	1.95 (0.69)	.6	-0.12 (-0.42, 0.18)	very small
Overall effectiveness	1.73 (0.77)	1.82 (0.83)	.7	-0.08 (-0.38, 0.22)	very small

<sup>a</sup> Mean (SD).

<sup>b</sup> Wilcoxon signed rank test with continuity correction.

<sup>c</sup> Statistically significant ( $P < .05$ ).

**Table 4. Comparison Between Human-Only and AI-Augmented Feedback Performance Based on Vignette Types**

	Preparedness N = 30 <sup>a</sup>	Professionalism N = 30 <sup>a</sup>	P-value <sup>b</sup>	Effect Size (90% CI)	Interpretation
Detail	2.07 (0.64)	1.72 (0.68)	.01 <sup>c</sup>	0.53 (0.10, 0.96)	medium
Specificity	2.18 (0.68)	1.85 (0.67)	.01 <sup>c</sup>	0.49 (0.06, 0.92)	small
Behavior focused	2.10 (0.59)	1.95 (0.67)	.3 <sup>c</sup>	0.24 (-0.19, 0.66)	small
Constructiveness	2.10 (0.67)	1.73 (0.81)	.02 <sup>c</sup>	0.49 (0.06, 0.92)	small
Actionability	2.12 (0.68)	1.65 (0.73)	.01 <sup>c</sup>	0.66 (0.22, 1.1)	medium
Overall effectiveness	1.97 (0.72)	1.58 (0.83)	.02 <sup>c</sup>	0.49 (0.06, 0.92)	small

<sup>a</sup> Mean (SD).

<sup>b</sup> Wilcoxon signed rank test with continuity correction.

<sup>c</sup> Statistically significant ( $P < 0.05$ ).

continued on next page

*continued from previous page*

## **Supplemental Online Material**

---

### **Supplemental Digital Content A:**

#### **Professionalism**

**Vignette:** You are an attending anesthesiologist and a member of your residency's clinical competency committee. Multiple attendings have expressed concerns about a resident's professionalism, stating that the resident resists feedback throughout the clinical day. Evaluations describe scenarios of the resident frequently dismissing suggestions from the attending, often justifying their behavior by quoting research studies. Other evaluations highlight the resident failing to complete requests from attendings, such as checking labs throughout a case and failing to administer medications that the resident did not feel were necessary. The resident has average performance on patient care and medical knowledge milestones relative to their peers. However, their actions have led to below-average scores on the professionalism milestones.

**Task:** How would you provide feedback to this resident? Use examples where necessary. Be concise and conversational.

#### **AI-Generated Feedback to this vignette:**

I wanted to have a chat with you about some concerns that have come up regarding your professionalism during your residency. It's important for your overall development as a physician, so let's discuss a few instances that were raised.

Firstly, it's great that you're well-versed in research studies, but it's equally important to maintain a respectful and collaborative attitude towards attendings and colleagues. When you find yourself disagreeing with a suggestion, consider framing it as a discussion rather than dismissing it outright. For example, you could say something like, 'I've come across a recent study that suggests an alternative approach. What are your thoughts on that?' This way, you're still sharing your knowledge but in a more constructive way.

In terms of completing requests, it's vital that you follow through with attending instructions promptly. If there's an issue or a concern, discuss it with them rather than deciding not to administer medications or skipping tasks. Communication is key here. For instance, you could say, 'I noticed this particular medication was suggested. Can we talk about the rationale behind it?'

Your patient care and medical knowledge are on track, which is great to see. However, your actions seem to be affecting your professionalism scores negatively. Remember, professionalism encompasses teamwork and effective communication.

*continued on next page*

*continued from previous page*

## **Supplemental Online Material *continued***

---

I suggest we work on improving your collaboration skills, active listening, and open communication with the team. It will not only benefit your interactions but also your overall growth as a resident. Please let me know if you'd like some resources or additional support in this area.

We all have room to grow, and I'm confident that with some adjustments, you'll continue to excel in your residency. Let's work together to address these concerns.

*continued on next page*

*continued from previous page*

## **Supplemental Online Material *continued***

---

### **Preparedness**

**Vignette:** You are an attending anesthesiologist working in the operating room with a resident for the first time. The cases consist of a mix of thoracic surgeries with varying patient acuity. Throughout the day, the resident appears consistently unprepared for cases. You note that the norepinephrine infusion was in the room but not ready to be administered, there was no transducer setup for the pre-planned arterial line, and the bronchoscope was not prepared before placing a double-lumen endotracheal tube. You discussed needing all of these things the evening before with the resident. The clinical day ends and the resident asks you for feedback on their performance.

**Task:** How would you provide feedback to this resident? Use examples where necessary. Be concise and conversational.

#### **AI-Generated Feedback to this vignette:**

I'm glad you're asking for feedback, as it's a key part of your development. Today, I noticed a few areas where your preparation could be improved, which is essential for efficient and safe patient care, especially in complex cases like thoracic surgeries.

For example, the norepinephrine infusion wasn't ready for immediate use. In thoracic surgeries, having vasoactive medications prepared is critical due to potential hemodynamic instability. It's important to anticipate and be prepared for such situations.

Also, the arterial line transducer wasn't set up. Continuous arterial pressure monitoring is vital in thoracic surgeries for real-time hemodynamic assessment, so having this equipment ready beforehand is crucial.

Lastly, the bronchoscope needed for the double-lumen endotracheal tube wasn't prepared. Ensuring tools for airway management are ready is key, especially for lung isolation procedures.

We discussed these preparations yesterday, so it's important to follow through to ensure patient safety. I suggest creating a checklist for each case, including equipment setup and specific surgical considerations. This will help you stay organized and ensure nothing is overlooked.

Remember, it's okay to ask questions or seek help when you're unsure. We're here to support each other and provide the best care for our patients. Let's focus on these areas for improvement and work together to enhance your skills and our patient care. Keep up the good work, and let's aim for continual progress.

*continued on next page*

*continued from previous page*

## **Supplemental Online Material *continued***

---

### **Supplemental Digital Content B:**

#### **Interview Procedure**

##### **Preparation:**

- Ensure the video conferencing (zoom) is working properly.
- Each attending has been assigned to receive **Series A** or **Series B** of the vignettes.
- Prepare to email each document for the assigned series
- Chat GPT Feedback Series A.1.pdf, Chat GPT Feedback Series A.2.pdf
- Chat GPT Feedback Series B.1.pdf, Chat GPT Feedback Series B.2.pdf

##### **Introduction (Phone/Video Conference Start):**

- Greet the attending anesthesiologist and introduce yourself as the research assistant.
- Inform them about the need for access to email during the interview.
- Remind them the conversation will be recorded.

##### **Protocol Explanation:**

- Explain that they will be presented with two scenarios. They should provide feedback as if speaking to a resident anesthesiologist.
- Inform them you will not respond to their feedback and to avoid asking questions.

##### **First Scenario:**

- Email the first scenario to the attending anesthesiologist.
- Chat GPT Feedback Series A.1, or B.1
- Instruct them to open the email and review the scenario.
- Ask them to begin providing feedback when ready.
- They may use any notes if necessary.
- Listen to the feedback silently without interruptions.
- Avoid providing any verbal or nonverbal cues.

##### **Second Scenario:**

- Email the second scenario to the attending anesthesiologist.
- Chat GPT Feedback Series A.2, or B.2
- Inform them that the second scenario will be accompanied by AI-generated feedback.
- Instruct them to review the scenario and the AI-generated feedback.
- Ask them to provide their feedback when ready.
- Explain they may use the AI-generated feedback or any notes if necessary.
- Listen to the feedback silently without interruptions.

*continued on next page*

*continued from previous page*

## **Supplemental Online Material *continued***

---

- Avoid providing any verbal or nonverbal cues.

### **Questions asked after the Second Scenario:**

- (2) Do you feel like the AI-generated feedback helped you, or do you feel like it didn't make an impact?
- (2) Do you feel like this is a tool you might use in the future if it were easily available?

### **Conclusion:**

- Thank the participant for their valuable insights and participation.

*continued on next page*

*continued from previous page*

## **Supplemental Online Material *continued***

---

### **Supplemental Digital Content C:**

#### **Standardized Grading Rubric**

##### **Detail:**

- 0: Not detailed
- 1: Somewhat detailed
- 2: Detailed
- 3: Highly detailed

##### **Specificity:**

- 0: Not specific
- 1: Somewhat specific
- 2: Specific
- 3: Highly specific

##### **Behavior-Focused:**

- 0: Not behavior-focused
- 1: Somewhat behavior-focused
- 2: Behavior-focused
- 3: Highly behavior-focused

##### **Constructiveness:**

- 0: Not constructive
- 1: Somewhat constructive
- 2: Constructive
- 3: Highly constructive

##### **Actionability:**

- 0: Not actionable
- 1: Somewhat actionable
- 2: Actionable
- 3: Highly actionable

##### **Overall Effectiveness Rating:**

- 0: Ineffective
- 1: Somewhat effective
- 2: Effective
- 3: Highly effective

##### **Origin:**

- Human
- AI
- Unsure

*continued on next page*

continued from previous page

## Supplemental Online Material *continued*

### Supplemental Digital Content D:

#### AI-Only Feedback Performance Based on Vignette Types

	Preparedness (N=6) <sup>1</sup>	Professionalism (N=6) <sup>1</sup>	P-value <sup>2</sup>
<b>Detail</b>	2.50 (0.84)	1.83 (0.75)	0.2
<b>Specificity</b>	2.67 (0.82)	2.17 (0.75)	0.3
<b>Behavior Focused</b>	2.33 (1.21)	2.50 (0.55)	>0.9
<b>Constructiveness</b>	2.33 (0.82)	2.67 (0.52)	0.8
<b>Actionability</b>	2.33 (0.82)	2.50 (0.55)	>0.9
<b>Overall Effectiveness</b>	2.33 (0.82)	2.50 (0.55)	0.8

1. Mean (SD)
2. Wilcoxon signed rank test with continuity correction

\* Statistically Significant (P < 0.05)

continued on next page

continued from previous page

## Supplemental Online Material *continued*

### Supplemental Digital Content E:

Interrater reliability assessed using intraclass correlation coefficient (ICC)

Feedback Rating Domain	ICC	95% Confidence Interval		F Test With True Value 0			
		Lower Bound	Upper Bound	Value	<i>f</i>	<i>df</i>	Sig
Detail	0.502	0.171	0.702	2.02	5.9	6.0	p=0.004
Specificity	0.361	-0.078	0.620	1.55	5.9	5.46	p=0.005
Behavior Focused	0.160	-0.388	0.494	1.19	5.9	9.48	p=0.005
Constructiveness	0.384	-0.033	0.632	1.62	5.9	9.33	p=0.005
Actionability	0.443	0.082	0.664	1.86	5.9	5.11	p=0.005
Overall Effectiveness	0.514	0.178	0.712	2.04	5.8	8.04	p=0.004

Estimates and their 95% confident intervals calculated based on a mean-rating ( $k = 2$ ), absolute-agreement, 2-way random-effects model.

continued on next page

continued from previous page

## Supplemental Online Material continued

### Supplemental Digital Content F:

	Gender					
	Human-Only			AI-Augmented		
	Female (N=11)	Male (N=19)	P-value	Female (N=11)	Male (N=19)	P-value
<b>Detail</b>						
Mean (SD)	2.09 (0.437)	1.76 (0.632)	0.14	2.14 (0.745)	1.76 (0.770)	0.21
<b>Specificity</b>						
Mean (SD)	2.05 (0.522)	1.92 (0.750)	0.63	2.36 (0.674)	1.89 (0.699)	0.08
<b>Behavior Focused</b>						
Mean (SD)	1.91 (0.539)	1.89 (0.658)	0.95	2.32 (0.603)	2.05 (0.664)	0.29
<b>Constructiveness</b>						
Mean (SD)	1.86 (0.745)	1.79 (0.902)	0.82	2.27 (0.518)	1.87 (0.723)	0.12
<b>Actionability</b>						
Mean (SD)	1.95 (0.650)	1.74 (0.872)	0.48	2.18 (0.560)	1.82 (0.730)	0.16
<b>Overall Effectiveness</b>						
Mean (SD)	1.68 (0.643)	1.76 (0.856)	0.79	2.14 (0.674)	1.63 (0.863)	0.11

Associations between attending gender and resident rating of human-only and AI-augmented feedback. Multiple measures from each attending were combined using the average score. ANOVA and T-test was used to study the associations. SD = Standard deviation.

continued on next page

continued from previous page

## Supplemental Online Material continued

\* Statistically Significant (P < 0.05)

Prior Feedback Training						
	Human-Only			AI-Augmented		
	Female (N=11)	Male (N=19)	P- value	No (N=18)	Yes (N=12)	P- value
<b>Detail</b>						
Mean (SD)	1.86 (0.614)	1.92 (0.557)	0.8	1.89 (0.654)	1.92 (0.949)	0.92
<b>Specificity</b>						
Mean (SD)	2.06 (0.616)	1.83 (0.749)	0.38	2.06 (0.591)	2.08 (0.900)	0.92
<b>Behavior Focused</b>						
Mean (SD)	1.78 (0.624)	2.08 (0.557)	0.18	2.11 (0.631)	2.21 (0.689)	0.69
<b>Constructiveness</b>						
Mean (SD)	1.69 (0.807)	2.00 (0.879)	0.34	1.97 (0.581)	2.08 (0.821)	0.67
<b>Actionability</b>						
Mean (SD)	1.69 (0.843)	2.00 (0.707)	0.31	1.89 (0.608)	2.04 (0.811)	0.56
<b>Overall Effectiveness</b>						
Mean (SD)	1.64 (0.763)	1.88 (0.801)	0.42	1.75 (0.624)	1.92 (1.08)	0.6

Associations between attending prior feedback training and resident rating of human-only and AI-augmented feedback. Multiple measures from each attending were combined using the average score. ANOVA and T-test was used to study the associations. SD = Standard deviation.

\* Statistically Significant (P < 0.05)

continued on next page

continued from previous page

## Supplemental Online Material *continued*

Age				
	Human-Only		AI-Augmented	
Feedback Domain	Correlation	P-value	Correlation	P-value
Detail	0.09662	0.61	-0.3021	0.11
Specificity	-0.057	0.77	-0.2776	0.14
Behavior Focused	0.07425	0.7	-0.0409	0.83
Constructiveness	0.07154	0.71	-0.2974	0.11
Actionability	0.08576	0.65	-0.2956	0.11
Overall Effectiveness	0.14239	0.45	-0.2109	0.26

Associations between attending age and resident rating of human-only and AI-Augmented feedback. Multiple measures from each attending were combined using the average score. Pearson correlation was used to study the associations.

\* Statistically Significant ( $P < 0.05$ )

Years as Attending				
	Human-Only		AI-Augmented	
Feedback Domain	Correlation	P-value	Correlation	P-value
Detail	-0.0127	0.95	-0.2952	0.11

continued on next page

*continued from previous page*

## **Supplemental Online Material *continued***

<b>Specificity</b>	-0.133	0.48	-0.246	0.19
<b>Behavior Focused</b>	-0.1441	0.45	-0.1403	0.46
<b>Constructiveness</b>	-0.1272	0.50	-0.3952	0.03*
<b>Actionability</b>	-0.0535	0.79	-0.3868	0.04*
<b>Overall Effectiveness</b>	-0.0398	0.84	-0.243	0.2

Associations between years as an attending and resident rating of human-only and AI-Augmented feedback. Multiple measures from each attending were combined using the average score. Pearson correlation was used to study the associations.

\* Statistically Significant ( $P < 0.05$ )

*continued on next page*

continued from previous page

## Supplemental Online Material continued

### Supplemental Digital Content G

#### Instances of Qualitative Code Co-Occurrence

	Preparedness	Professionalism
<b>When to use</b>	0	6  <i>“I think it might be even more useful in like professionalism [conversations].” (Attending 0.694)</i>
<b>When not to use</b>	4  <i>“If it’s some[thing] like with the patient or the resident not having the operating room ready like for that, I feel like I might not necessarily need it.” (Attending 0.454)</i>	0

Co-occurrence (and exemplary quotes) of codes “When to use” and “When not to use” vs. “Preparedness” and “Professionalism”. The numbers in the matrix represent the frequency with which each pair of codes occur together.