Original Paper

Ambient AI Scribes to Create Educational Feedback Notes for Medical Students: Randomized Trial

Jaideep S Talwalkar[1], MD; David Chartash[2], PhD; Lisa Zhang[2], MD; Michael Makutonin[2], MD; Conrad W Safranek[3], BSc; Anne Elizabeth Sidamon-Eristoff[3], AB; Lee H Schwamm[4], MD; Donald S Wright[2,5], MD, MHS

[1]Departments of Medicine and Pediatrics, Yale School of Medicine, New Haven, CT, United States
[2]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, United States
[3]Yale School of Medicine, New Haven, CT, United States
[4]Departments of Neurology and Biomedical Informatics and Data Sciences, Yale School of Medicine, New Haven, CT, United States
[5]Department of Veterans Affairs, Connecticut VA Healthcare System, West Haven, CT, United States

Corresponding Author:

Jaideep S Talwalkar, MD
Departments of Medicine and Pediatrics
Yale School of Medicine
367 Cedar Street, Bldg D
New Haven, CT 06510
United States
Phone: 1 203-737-4190
Fax: 1 203-737-4199
Email: jaideep.talwalkar@yale.edu

Abstract

Background: High-quality observation and feedback contribute to the development of clinical competence and professional growth in medical education. Faculty often struggle to translate verbal observations into written feedback because of documentation burden and competing demands. Ambient artificial intelligence (AI) scribes, already adopted in clinical practice, may address this challenge by capturing verbal exchanges and generating structured notes.

Objective: The purpose of this study was to examine the use of ambient AI scribes to generate educational feedback notes during a formative medical interviewing workshop for first-year medical students in March and April 2025.

Methods: Thirteen instructors were randomized to control (human-only) or intervention (AI scribe–assisted) workflows to complete narrative feedback forms. The intervention group used an AI scribe to generate transcripts of student-instructor encounters, which were then summarized into feedback notes using a large language model and edited by the instructors before submission. All narratives were scored using the Evaluation of Feedback Captured Tool (EFeCT). Factual accuracy of a subsample of unedited AI feedback summaries was reviewed against the source transcripts. Task load and usability were measured using NASA Task Load Index and System Usability Scale, respectively.

Results: Instructors submitted feedback on 92.2% (94/102) of the students. EFeCT scores on the scale from 0 to 5 were higher for human-edited AI narratives (median 3.00, IQR 2.00-4.00) and unedited AI summaries (median 3.00, IQR 3.00-4.00) than for human-only narratives (median 2.00, IQR 1.75-3.00; P<.001). Human-only narratives were shorter than AI-assisted outputs (P<.001). Review of 117 AI-generated feedback elements showed a 6.8% (n=8) mischaracterization and 1.7% (n=2) hallucination rate, with most errors corrected during editing. Task load was high, and usability was marginal in both the control and intervention groups, with no significant differences (P=.31 and P=.40, respectively).

Conclusions: An ambient AI scribe–assisted workflow improved the quality of written narrative feedback with no observed increase in instructor effort compared to human-only documentation. Although occasional inaccuracies required review, this innovation has the potential to transform feedback documentation.

JMIR Med Educ 2026;12:e89996; doi: 10.2196/89996

Keywords: ambient scribe; artificial intelligence; AI; feedback; medical student education; formative assessment; competency-based education

Introduction

Problem: Documentation Burden and Quality of Feedback in the Modern Assessment Paradigm

High-quality observation and feedback contribute to the development of clinical competence and professional growth in medical education. Effective feedback provides learners with specific and actionable information on their performance, which can promote deliberate practice, motivation, and engagement with ongoing feedback [1-4]. While verbal feedback is critical for collaborative discussion [5] and occurs frequently in clinical [1] and simulation settings [6], it may be limited as a tool for longitudinal reflection as learners recall few feedback points when given verbal feedback alone [5,7]. Written feedback supports self-regulated learning by allowing trainees and their supervisors to review past narratives, set goals, and refine strategies for improvement [1,8]. By providing detailed, context-rich guidance, narrative feedback can help learners focus efforts on achieving professional outcomes expected of them during training [1,9].

In recognition of the value of both verbal and written feedback, accreditation bodies such as the Liaison Committee on Medical Education emphasize the importance of defined competencies within a robust assessment system that includes narrative feedback [10]. Unfortunately, medical educators faced with competing responsibilities and increased educational documentation burden [11] struggle to convert real-time verbal observations into meaningful written feedback [1]. As a result, rather than serving as a tool to promote learning, formative assessment is often reduced to a checklist activity disconnected from best practices for feedback. Brief, nonspecific narratives are common and diminish the educational value of written feedback [11,12].

Solution: Ambient Artificial Intelligence Scribes in Educational Practice

Ambient artificial intelligence (AI) refers to AI embedded into environments, working continuously in the background to support human tasks [13]. Ambient AI scribes have seen rapid adoption to address the growing documentation burden in clinical practice. Ambient AI scribes passively capture and document physician-patient conversations into structured clinical notes [14]. There is early evidence suggesting that these tools reduce administrative burden while maintaining or improving documentation quality [15,16]. The potential for this technology to extend beyond patient care into medical education is compelling. By capturing and structuring verbal formative feedback provided by faculty in real time, ambient AI scribes could generate written records of learner performance. These "educational feedback notes" could transform narrative workflows similar to the way in which AI-generated clinical notes are transforming clinical practice [17].

Gap: No Use in Medical Education Settings

Despite the promise of ambient AI scribes in clinical settings, their application in medical education remains unexplored. No studies have described the use of these tools in directly observed encounters or small-group teaching settings, examined whether they can reliably capture narrative feedback in teaching contexts, compared how their outputs align with best practices for written feedback, or determined whether faculty perceive them as usable while actively teaching. No extensions to summarize data into educational feedback notes currently exist. The potential for ambient AI scribes to transform feedback documentation in education remains untested.

This study is the first to our knowledge to examine the use of ambient AI scribes to generate educational feedback notes during directly observed standardized patient (SP) encounters. Specifically, we aimed to (1) evaluate the quality and accuracy of narrative feedback captured by an ambient AI scribe workflow compared to that of feedback provided by humans without AI assistance during a formative medical interviewing workshop for first-year medical students and (2) assess the usability of this technology by medical educators.

Methods

Educational Setting

This study took place in spring 2025 as part of a formative medical interviewing workshop for all first-year medical students at Yale School of Medicine. During the workshop, each medical student spent 20 minutes conducting a complete history with an SP while being observed by an instructor and 3 peers. Students and instructors were permitted to call "time-out" for interspersed guidance and feedback [6]. Upon completion of the interview, 10 minutes were allotted for self-, peer, and instructor feedback. SPs provided no feedback. The exercise was repeated until all 4 students in the room had interviewed separate SPs portraying different scenarios and received individual verbal feedback. After the workshop, instructors completed a postsession feedback narrative (Multimedia Appendix 1) for each student within a learning management system (Medtrics Lab LLC). The workshop was cycled over 4 afternoons, with each student participating once. Instructors devoted 3 and a half hours per workshop, including presession faculty development. This was the seventh iteration of the workshop in the first-year clinical skills curriculum, with each workshop structured similarly, offering students opportunities for repeat practice as their foundational knowledge grows during medical school.

Participants and Platforms

Study participants were 13 clinician educators who had signed up to teach medical interviewing workshops based on their availability before the study was announced. All had completed standard training on teaching medical interviewing, facilitating small groups, working with SPs, and giving feedback. All had facilitated other medical interviewing

workshops earlier in the academic year and completed feedback forms associated with those sessions. Instructors were informed of the study via email and in-person announcements.

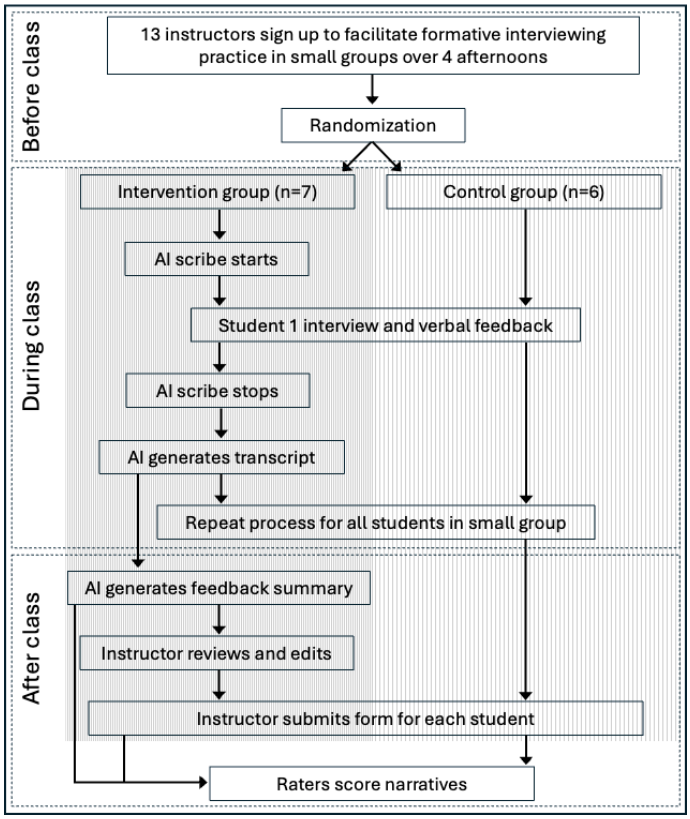
All instructors received the same study-specific training on form completion and use of the verbatim transcription feature within the health system's AI scribe (Abridge Inc), a generative AI tool based on a large language model (LLM) built using medical notes. The AI scribe transcripts were extracted separately from the final medical note that Abridge produces and then fed manually into a private instance of GPT-4o (OpenAI) within the university's secure infrastructure (Clarity Platform; Multimedia Appendix 2) to produce a feedback summary that ignored the direct medical content and instead focused on student feedback. Other tools included a data collection instrument (Qualtrics International Inc) and the learning management system. Those wishing to opt out were instructed to inform a course administrator not involved in the study.

While the entire class of 104 medical students involved in the workshop was given the option to opt out, no students or instructors did.

Feedback Workflows and Data Collection

Instructors were randomized to the control or intervention group through random number generation to reduce bias of fixed effects from prior experience with the educational activity and technology, thereby isolating effects on the quality of the narratives (Figure 1). Instructors were informed of their group assignments immediately prior to the workshop. Students were unaware of their instructor's group assignment, and all instructors placed their smartphones on a counter centrally located in the room at the start of the session. Control group instructors did not activate the AI scribe, and after completing the workshop, composed the "human narratives" on the postsession feedback forms without AI assistance. Intervention group instructors activated the AI scribe, which ran unobtrusively on a smartphone in the background, during each interview and feedback session to create a session transcript. After the workshop, they used a zero-shot prompt (ie, a prompting strategy in which the LLM is asked to perform a task without being given examples, training, or fine-tuning) [18,19] within the Clarity Platform (Multimedia Appendix 3) to extract the educational feedback comments and use them to create a feedback summary from each transcript. The prompt was written to summarize feedback verbalized during the educational session, not to create new feedback. Instructors reviewed and edited these outputs as they deemed necessary before entering them as final narratives, thus turning unedited AI feedback summaries into human-edited AI narratives.

Figure 1. Depiction of the study workflow. Instructors were randomized to the control or intervention group; the latter used an artificial intelligence (AI) scribe to create a feedback summary. Three sources of narratives were analyzed: unedited AI, human-edited AI, and human only (control). Raters were blinded to the source of the narratives.



All text created through this process was submitted via a data collection instrument designed to preserve student anonymity. The instrument also prompted instructors to complete the NASA Task Load Index [20] and System Usability Scale [21]

to measure instructor cognitive load and general usability, respectively.

Review of Narratives

Feedback narratives were assessed using the Evaluation of Feedback Captured Tool (EFeCT), a feedback quality scoring tool consisting of 5 elements, each of which represents an aspect of good written feedback. One point is given for each element present, for a total score of 0 to 5 [22]. Two blinded, independent raters (DC and LZ) assigned scores asynchronously for each narrative. Reconciliation of mental models for each element was performed after 10 and 25 narratives. After all narratives were scored, the 2 raters synchronously reconciled their scores into a final score for each element for each narrative using the constant comparative method. This final reconciled score was passed to a third rater (JST) to be confirmed. In addition to the use of EFeCT, the raters flagged whether they thought the narrative was generated through AI assistance (binary variable).

Evaluation of Factuality of AI Feedback Summaries

Eight unedited AI feedback summaries were randomly sampled from the intervention cohort to benchmark any potential factual inaccuracies as an exploratory secondary analysis. These summaries were manually delimited into concepts by identifying each independent feedback statement that relied on an observed behavior in the simulation. Each individual feedback concept in the unedited AI feedback summary was cross-referenced with the full transcript of the session by an author (DSW) and classified for factuality (ie, whether the information in the written text corresponded to a real-world fact) [23]. Feedback was classified as correct if it matched the content of the original transcript, mischaracterized if it misrepresented the original context or meaning of the interaction in the transcript, or hallucinated if it referenced events not present in the transcript. Each feedback element classified as mischaracterized or hallucinated was then compared to the human-edited AI narrative to see whether the error was resolved.

Statistical Analysis

Analyses were conducted in R (version 4.2.2; R Foundation for Statistical Computing). For the EFeCT dataset, total EFeCT feedback scores and word counts were compared across sources (human only, unedited AI, and human-edited AI) using Kruskal-Wallis tests, with pairwise Wilcoxon rank-sum tests with Holm correction for multiple comparisons. Individual EFeCT subscores (elements B-F) were treated as binary outcomes and compared across groups using the Fisher exact test and then pairwise when the omnibus test was significant with Holm correction. Factuality analysis was conducted using descriptive statistics. The Fisher exact test was used to compare completion rates. We tested assumptions of normality (Shapiro-Wilk) and homogeneity of variance (Levene) prior to the cognitive load and usability analysis, in which we used independent-sample 2-tailed *t* tests. All tests were 2 tailed with significance set at an α value of .05.

Ethical Considerations

This study received an exemption from review by the Yale University Institutional Review Board due to its educational nature on January 23, 2025 (protocol ID 2000039478).

Results

Each of the 13 instructors taught 1 to 4 afternoons of the 4-workshop cycle. Of the 104 students in the class, 2 (1.9%) were absent due to illness. Instructors submitted feedback forms via the data collection instrument on 92.2% (94/102) of the students. The 46.2% (6/13) of instructors randomized to the control group taught a mean of 1.7 (SD 0.5) sessions and submitted human-only narratives for 84.2% (32/38) of the students in their groups, all of which were submitted on the day of the session. The 53.8% (7/13) of instructors in the intervention group taught a mean of 2.6 (SD 1.1) sessions and submitted unedited AI feedback summaries for 95.3% (61/64) of the students and human-edited AI narratives for 96.9% (62/64) of the students, a completion rate significantly higher than that of the control group (*P*=.049). All were submitted within 1 week of the session, with 95.2% (59/62) submitted on the day of the session.

Median EFeCT scores were 2.00 (IQR 1.75-3.00) for human-only narratives (control group), 3.00 (IQR 2.00-4.00) for human-edited AI narratives, and 3.00 (IQR 3.00-4.00) for unedited AI feedback summaries (Table 1). When narratives were grouped by feedback generation method, a significant difference in total EFeCT scores was observed (*P*<.001; Figure 2). Subsequent pairwise comparisons demonstrated higher EFeCT scores in both the human-edited and unedited AI feedback methods relative to the human-only narratives (*P*=.005 and *P*<.001, respectively), whereas no difference was observed between the 2 AI methods (*P*=.11). Similar significant overall and pairwise effects were observed on 4 of the 5 EFeCT component elements (Table 2 and Figure 3).

Table 1. Review of narratives, including factuality analysis.

Outcome	Narrative feedback workflow			Groupwise comparison	
	Human-only narratives (n=32)	Human-edited AI narratives[a] (n=62)	Unedited AI narratives (n=61)	*P* value	Effect size estimate (ε^2)
EFeCT[b] score, median (IQR)	2.00 (1.75-3.00)	3.00 (2.00-4.00)	3.00 (3.00-4.00)	<.001[c]	0.11
Word count, median (IQR)	45.50 (33.00-62.00)	312.50 (207.25-355.00)	344.00 (310.00-396.00)	<.001[d]	0.40

Outcome	Narrative feedback workflow			Groupwise comparison	
	Human-only narratives (n=32)	Human-edited AI narratives[a] (n=62)	Unedited AI narratives (n=61)	*P* value	Effect size estimate (ε^2)
Mischaracterizations among AI feedback elements, n (%)	—[e]	1/117 (0.9)	8/117 (6.8)	—	—
Hallucinations among AI feedback elements, n (%)	—	2/117 (1.7)	2/117 (1.7)	—	—

[a]AI: artificial intelligence.
[b]EFeCT: Evaluation of Feedback Captured Tool.
[c]Via the Kruskal-Wallis test. Additional pairwise comparisons demonstrated higher scores for edited (*P*=.005) and unedited (*P*<.001) AI narratives relative to human-only narratives.
[d]Additional pairwise comparisons demonstrated that all differences were statistically significant.
[e]Not applicable.

Figure 2. Evaluation of Feedback Captured Tool (EFeCT) scores by narrative source: human only, unedited artificial intelligence (AI), and edited AI. The EFeCT consists of 5 elements, each of which represents an aspect of good written feedback. One point is given for each element present for a total score of 0 to 5.

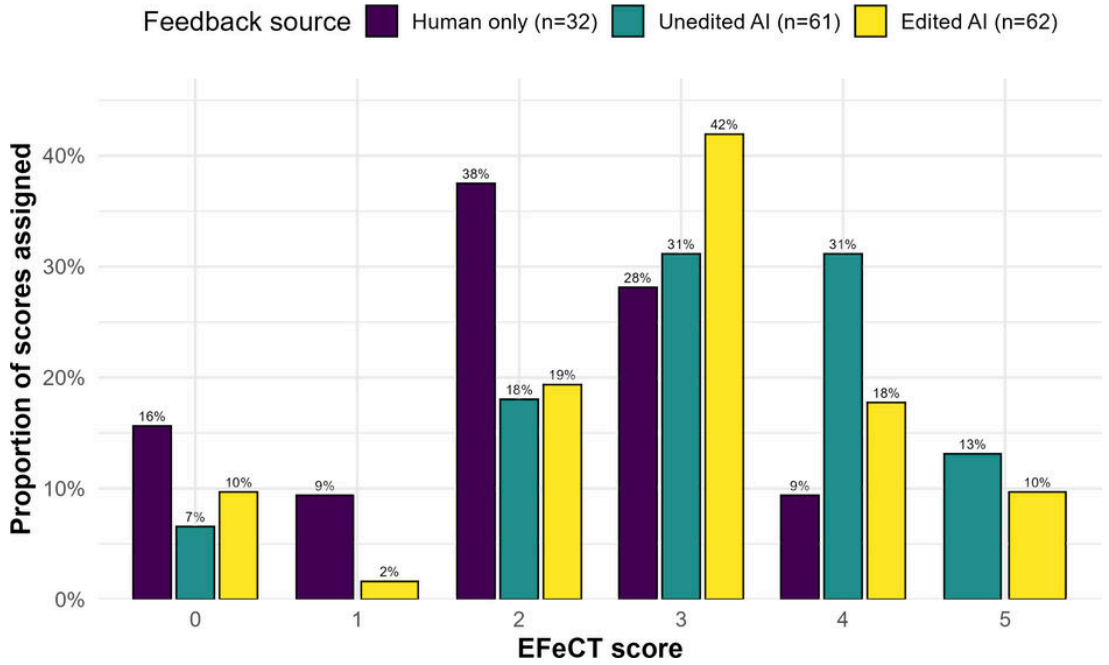


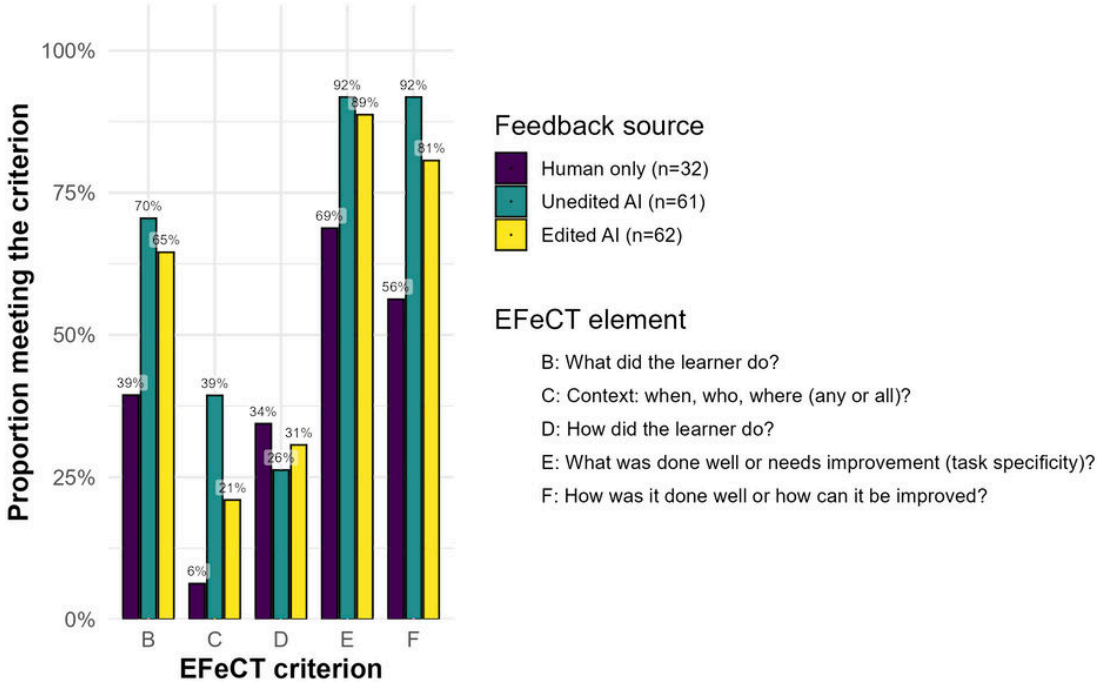
Table 2. Evaluation of Feedback Captured Tool (EFeCT) element comparisons.

Element and pairwise comparison	Pairwise *P* value[a]	Risk difference	Groupwise comparison[b]	
			P value	Effect size, Cramér *V*
What did the learner do? Interpretation: explicit statements that the participant performed a medical interview or completed a medical history.			.01	0.21
Unedited AI[c] vs human-edited AI	.56	0.06		
Human only vs human-edited AI	.06	–0.25		
Human only vs unedited AI	.01	–0.31		
Context: when, who, where? Interpretation: a positive score was assigned if the raters were able to recognize which patient was being interviewed based on the context provided.			.001	0.27
Unedited AI vs human-edited AI	.06	0.18		
Human only vs human-edited AI	.08	–0.15		
Human only vs unedited AI	.002	–0.31		
How did the learner do? Interpretation: there was an explicit statement verbalized that the participant performed a medical interview AND a qualifier was provided stating how well they performed.			.70	0
Unedited AI vs human-edited AI	—[d]	—		
Human only vs human-edited AI	—	—		
Human only vs unedited AI	—	—		

Element and pairwise comparison	Pairwise *P* value[a]	Risk difference	Groupwise comparison[b]	
			P value	Effect size, Cramér *V*
What was done well or needs improvement? Interpretation: any explicit statement of feedback on a specific task.			.01	0.23
Unedited AI vs human-edited AI	.76	0.03		
Human only vs human-edited AI	.05	–0.20		
Human only vs unedited AI	.02	–0.23		
How was it done well or how can it be improved? Interpretation: any explicit statement that highlighted a specific element of a task that was done well or needed improvement.			<.001	0.31
Unedited AI vs human-edited AI	.11	0.11		
Human only vs human-edited AI	.03	–0.24		
Human only vs unedited AI	<.001	–0.36		

[a]Pairwise comparisons via the Fisher exact test with Holm correction for multiple comparisons; effect size reported as the risk difference (proportion meeting the criterion in group 1 minus group 2). A positive risk difference indicates higher EFeCT element attainment in the first-listed group.
[b]Via the Fisher exact test on 3 × 2 tables.
[c]AI: artificial intelligence.
[d]Not applicable.

Figure 3. Evaluation of Feedback Captured Tool (EFeCT) scores by element. The EFeCT consists of 5 elements. One point is given for each element present. Element lettering begins with B because the "A" designation is used when the narrative section is blank, which did not apply to this study. AI: artificial intelligence.



Human-only narratives were significantly shorter (median 45.50, IQR 33.00-62.00 words) than those generated in human-edited (median 312.50, IQR 207.25-355.00 words) and unedited (median 344.00, IQR 310.00-396.00 words) AI feedback workflows (*P*<.001 in both cases). Human-edited AI narratives were shorter than unedited summaries (*P*=.001; Multimedia Appendix 4). A total of 117 AI-generated feedback elements were manually reviewed for factuality, demonstrating 8 (6.8%) mischaracterizations and 2 (1.7%) hallucinations in comparison to the source transcript. Most errors were corrected by instructors prior to submission of human-edited AI narratives (Table 1). Raters correctly identified 93.5% (58/62) of AI-assisted narratives and 100% (32/32) of human-only narratives.

Task load related to completion of the narratives as measured using the NASA Task Load Index did not differ between the control (mean 47.08, SD 12.39) and intervention (mean 54.32, SD 9.45) groups (*P*=.31; mean difference 95% CI –8.28 to 22.75), with scores indicating high cognitive load for both workflows [24]. Similarly, general usability as measured using the System Usability Scale did not differ between the control (mean 58.75, SD 19.55) and intervention (mean 49.70, SD 13.65) groups (*P*=.40; mean difference 95% CI –33.25 to 15.15), with scores indicating marginal acceptability [21].

Discussion

Quality of Narratives

The integration of an ambient AI scribe into educational workflows improved the quality of written feedback in a medical student workshop with no observed increase in instructor effort. Feedback narratives generated with AI assistance received significantly higher scores on the EFeCT instrument than those written by clinician educators without AI assistance. By passively capturing verbal feedback and restructuring it into written narratives, AI scribes may address a long-standing challenge in competency-based medical education: translating verbal observations into high-quality written narratives. This is the first study to our knowledge to explore the use of ambient AI scribes to create educational feedback notes, with improvement in documentation quality [15,16] and timeliness [25] similar to that observed with ambient scribes in clinical care.

The improvement observed with AI-assisted narratives is educationally meaningful: a 1-point increase on the EFeCT instrument exceeded the gains reported in the original validation study by Ross et al [22], where multiyear faculty development and feedback were required to achieve a smaller degree of improvement. Several factors may explain why AI-assisted narratives were scored higher on nearly all elements of the EFeCT instrument. Because the EFeCT instrument rewards inclusion of specific feedback elements, the longer AI-assisted narratives may have scored higher because they provided the detail needed to explicitly include those elements. The zero-shot LLM, guided only by a simple prompt, drew on a broad knowledge base of feedback principles, generating more structured, context-rich narratives. In contrast, despite prior extensive training in delivering verbal and written feedback, instructor-written narratives omitted best practice suggestions, as has been described previously. In an integrative review on feedback practices, Bing-You et al [26] identified inadequate feedback as a frequent theme in the literature even among experienced educators and those who had received feedback training. Gingerich et al [12] highlight a tendency for faculty to either omit written feedback or provide only vague, nonspecific comments, particularly when they feel that essential feedback has already been delivered verbally or is considered "unwritable" due to social or relational dynamics. Additionally, instructors may have relied on the assumption that students will sufficiently retain verbal feedback given during the encounter despite evidence to the contrary [7]. Finally, despite having volunteered, instructors may have faced competing professional demands that accumulated during the 3.5-hour workshop, potentially limiting time spent on feedback forms [1,11].

The impact of an AI-assisted workflow in narrative feedback may be even greater outside of research settings. In this study, all instructors were clinician educators who were aware that their notes would be reviewed as part of the research protocol. This may have resulted in more detailed narratives from instructors in the control group than they would have submitted otherwise. Additionally, nearly all narratives were submitted immediately. In real-world educational environments, where delays in form completion are common and specificity is often lost due to poor recall [27], the potential benefit of ambient AI scribes may be even greater. Alternatively, observed differences in completion rates may be attributable to novelty of the intervention workflow, warranting iterative evaluation of sustained impact.

Accuracy and Reliability of AI Outputs

As a secondary, exploratory outcome, we also attempted to measure the rate of hallucinations and mischaracterizations. Within the sample of unedited feedback summaries that were examined, factual inconsistencies with the underlying transcript were rare; however, the certainty of this estimate is limited by the size of the sample reviewed. Instructors corrected most of the errors in the unedited AI feedback summaries, resulting in shorter human-edited AI narratives that retained higher EFeCT scores. This finding underscores the importance of human oversight over AI outputs. Accurate documentation in medical education remains paramount given its importance for learner development, institutional accountability, and public trust in training outcomes [9].

Task Load and Usability

We did not observe reductions in task load or improvements in usability when using AI scribes compared to the traditional manual human workflow. Both approaches were rated as cognitively demanding and only marginally usable. Notably, the manual workflow itself was associated with high task load, echoing prior work showing that educational documentation can add extraneous demands that detract from core teaching activities [1,11,28]. Cognitive load theory dictates that, when administrative processes create high extraneous load, educators have fewer cognitive resources available for essential instructional and clinical tasks [29]. This dynamic likely contributes to the tendency of faculty to abbreviate written feedback despite their awareness of best practices.

Our findings contrast with reports from clinical medicine, where ambient AI scribes have consistently reduced cognitive burden and improved usability for clinicians [14,16,30,31]. The discrepancy likely reflects the research workflow we used, which required manual copying and pasting among 3 separate platforms (Abridge, Clarity, and Medtrics), all of which were used at no added cost as they were already available within the university and health system. By design, the copy-paste workflow introduced inefficiencies to capture interim outputs for study purposes that would be unnecessary in practice. In a production-grade system, this would be replaced by automated data flow, which may lead to better task load and usability relative to our intervention, so our findings related to cognitive load and usability should be interpreted with caution. Additionally, with well-designed systems, cognitive load [29] and usability [21] typically improve with experience, suggesting that adoption over time may mitigate some of the present challenges as workflows become smoother and more familiar.

Limitations

This study has several limitations. It was conducted at a single institution with a small number of highly trained educators, limiting generalizability. Participants were diligent in reviewing AI outputs, which may not reflect the behaviors of a broader faculty population in usual teaching environments. However, in real-world practice, we can reasonably expect human-only narratives to be less diligently submitted, potentially further increasing the observed differences in quality between AI-assisted and human-only narratives.

Additionally, the AI workflow relied on robust verbal feedback having occurred as the scribe was not enabled to generate its own feedback. We used this safeguard to ensure that feedback was consistent with session objectives.

Second, some contamination may have occurred if instructors in the intervention arm occasionally defaulted to human-only narratives, potentially due to usability challenges. Raters were excellent at identifying AI-assisted narratives, so the rare occasions when raters misattributed an AI-generated narrative to the human-only group raise the possibility that technical difficulties or workflow barriers led some intervention instructors to manually compose narratives. If this occurred, the observed differences in quality between AI-assisted and human-only narratives may, in fact, underestimate the true effect of AI assistance.

Finally, while the EFeCT instrument provides a framework for evaluating written feedback, this was its first application to our knowledge to AI-generated narratives. The EFeCT instrument focuses on structure and content but does not capture tone or learner-perceived utility.

Future Directions

Next steps should expand on these findings in several directions. It is essential to engage learners and clinical coaches to determine whether AI-assisted narratives are useful or actionable for reflection and growth. Through prompt engineering, AI outputs can be iteratively refined to more closely align with the needs of learners and coaches and adjusted to improve factuality [32]. Additionally, integrating AI scribe content into streamlined workflows and data pipelines could reduce task load and facilitate adoption. Finally, applications in other educational settings—such as clinical precepting and bedside teaching—represent promising areas where ambient AI could help close the gap between frequent verbal feedback and inadequate written documentation.

Conclusions

Ambient AI scribes represent a promising innovation in medical education, where the challenge of creating high-quality narrative feedback persists. In this study, AI-assisted feedback narratives—produced through the combination of automated transcription and zero-shot prompting of an LLM—were of higher quality than human-only narratives with no observed increased human effort. While usability gains were not realized in this research workflow, integration into streamlined educational systems could replicate the successes of ambient AI scribes observed in clinical practice. With continued refinement and oversight, ambient AI scribes have the potential to strengthen feedback culture; support longitudinal assessment; and, ultimately, enhance the learning experience for medical students.

Acknowledgments

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the US Department of Veterans Affairs or the US government. The authors used large language models in the preparation of the manuscript to assist with literature search (OpenEvidence), formatting references (GPT-4o; OpenAI), and mock peer review (GPT-4o). GPT-4o was accessed using a generative artificial intelligence tool housed within Yale University's secure infrastructure in which inputs are not added to the external training dataset. There is no plagiarism of text or images related to our use of artificial intelligence.

Funding

No external funding was received.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the need to protect student privacy, as the data were collected within a single class of students at a single medical school. The data are available from the corresponding author on reasonable request.

Authors' Contributions

JST contributed to conceptualization, formal analysis, investigation, methodology, project administration, resources, software, supervision, visualization, and writing—original draft. DC contributed to conceptualization, data curation, investigation, methodology, validation, and writing—original draft. LZ contributed to data curation, investigation, validation, and writing—review and editing. MM contributed to data curation, formal analysis, validation, and writing—review and editing. CWS contributed to conceptualization, methodology, and writing—review and editing. AES-E contributed to conceptualization, methodology, and writing—review and editing. LHS contributed to conceptualization, methodology, resources, software, supervision, and writing—review and editing. DSW contributed to data curation, formal analysis, investigation, methodology, software, validation, visualization, and writing—review and editing.

Conflicts of Interest

LHS reports consulting with Medtronic, LifeImage, Genentech, and Penumbra unrelated to this work; serving as a content advisor on digital health to the *Stroke* editorial board; holding an artificial intelligence–related patent pending (US2024031761); and representing Yale New Haven Health System on a hospital advisory committee for Abridge, an ambient artificial intelligence clinical documentation company. All other authors declare no other conflicts of interest.

Multimedia Appendix 1

Postsession form completed by the instructor for each student. This standard formative assessment form is used across curricular sessions during which medical students are directly observed performing the patient care skills of history taking or physical examination. Instructors compose a feedback narrative in the final open-response box.
[PNG File (Portable Network Graphics File), 224 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Clarity Platform technical specifications.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Zero-shot prompt in GPT-4o used by the intervention group.
[DOCX File (Microsoft Word File), 14 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Sample feedback notes submitted by instructors to first-year medical students following the formative medical interviewing workshop. Three representative samples from each workflow are presented with the removal of student identifiers. Control group instructors created the narratives manually. Intervention group instructors created the narratives with the use of an ambient artificial intelligence scribe.
[DOCX File (Microsoft Word File), 19 KB-Multimedia Appendix 4]

References

1. Cooper D, Holmboe ES. Competency-based medical education at the front lines of patient care. N Engl J Med. Jul 24, 2025;393(4):376-388. [doi: 10.1056/NEJMra2411880] [Medline: 40700689]
2. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ. Jan 2019;53(1):76-85. [doi: 10.1111/medu.13645] [Medline: 30073692]
3. Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide No. 7. Med Teach. Mar 2006;28(2):117-128. [doi: 10.1080/01421590600622665] [Medline: 16707292]
4. Lai MM, Roberts N, Mohebbi M, Martin J. A randomised controlled trial of feedback to improve patient satisfaction and consultation skills in medical students. BMC Med Educ. Aug 20, 2020;20(1):277. [doi: 10.1186/s12909-020-02171-9] [Medline: 32819352]
5. Natesan S, Jordan J, Sheng A, et al. Feedback in medical education: an evidence-based guide to best practices from the Council of Residency Directors in Emergency Medicine. West J Emerg Med. May 5, 2023;24(3):479-494. [doi: 10.5811/westjem.56544] [Medline: 37278777]
6. Talwalkar JS, Cyrus KD, Fortin AH. Twelve tips for running an effective session with standardized patients. Med Teach. Jun 2020;42(6):622-627. [doi: 10.1080/0142159X.2019.1607969] [Medline: 31033363]
7. Humphrey-Murto S, Mihok M, Pugh D, Touchie C, Halman S, Wood TJ. Feedback in the OSCE: what do residents remember? Teach Learn Med. 2016;28(1):52-60. [doi: 10.1080/10401334.2015.1107487] [Medline: 26787085]
8. Sargeant JM, Mann KV, van der Vleuten CP, Metsemakers JF. Reflection: a link between receiving and using assessment feedback. Adv Health Sci Educ Theory Pract. Aug 2009;14(3):399-410. [doi: 10.1007/s10459-008-9124-4] [Medline: 18528777]
9. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32(8):676-682. [doi: 10.3109/0142159X.2010.500704] [Medline: 20662580]
10. Publications. Liaison Committee on Medical Education. URL: https://lcme.org/publications/ [Accessed 2026-05-10]
11. Szulewski A, Braund H, Dagnone DJ, et al. The assessment burden in competency-based medical education: how programs are adapting. Acad Med. Nov 1, 2023;98(11):1261-1267. [doi: 10.1097/ACM.0000000000005305] [Medline: 37343164]
12. Gingerich AN, Lingard L, Sebok-Syer SS, Watling CJ, Ginsburg S. "Praise in public; criticize in private": unwritable assessment comments and the performance information that resists being written. Acad Med. Nov 1, 2024;99(11):1240-1246. [doi: 10.1097/ACM.0000000000005839] [Medline: 39137257]

13. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catalyst. Apr 16, 2025;6(5). [doi: 10.1056/CAT.25.0040]
14. Gams M, Gu IYH, Härmä A, Muñoz A, Tam V. Artificial intelligence and ambient intelligence. J Ambient Intell Smart Environ. 2019;11(1):71-86. [doi: 10.3233/AIS-180508]
15. Cain CH, Davis AC, Broder B, et al. Quality assurance during the rapid implementation of an AI-assisted clinical documentation support tool. NEJM AI. 2025;2(4). [doi: 10.1056/AIcs2400977]
16. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catalyst. Feb 21, 2024;5(3). [doi: 10.1056/CAT.23.0404]
17. Fu P. Transforming clinical workflows with artificial intelligence (AI)-based technologies. In: Zheng K, Westbrook J, Patel VL, editors. Reengineering Clinical Workflow in the Digital and AI Era. Springer; 2025:35-55. [doi: 10.1007/978-3-031-82971-0_3]
18. Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. J Med Internet Res. 2025;27:e72644-e72644. [doi: 10.2196/72644] [Medline: 40955776]
19. CHART Collaborative, Huo B, Collins GS, et al. Reporting guideline for chatbot health advice studies: the CHART statement. JAMA Netw Open. Aug 1, 2025;8(8):e2530220. [doi: 10.1001/jamanetworkopen.2025.30220] [Medline: 40747871]
20. Hart SG. NASA-Task Load Index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet. Oct 2006;50(9):904-908. [doi: 10.1177/154193120605000909]
21. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum Comput Interact. Jul 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
22. Ross S, Hamza D, Zulla R, Stasiuk S, Nichols D. Development of and preliminary validity evidence for the EFeCT feedback scoring tool. J Grad Med Educ. Feb 2022;14(1):71-79. [doi: 10.4300/JGME-D-21-00602.1] [Medline: 35222824]
23. Saurí R, Pustejovsky J. Are you sure that this happened? Assessing the factuality degree of events in text. Comput Linguist. Jun 2012;38(2):261-299. [doi: 10.1162/COLI_a_00096]
24. Prabaswari AD, Basumerda C, Utomo BW. The mental workload analysis of staff in study program of private educational organization. IOP Conf Ser Mater Sci Eng. 2019;528:012018. [doi: 10.1088/1757-899X/528/1/012018]
25. Moura LM, Mishuris RG, Metlay JP, et al. Hybrid ambient clinical documentation and physician performance: work outside of work, documentation delay, and financial productivity. J Gen Intern Med. Apr 2026;41(5):1294-1303. [doi: 10.1007/s11606-025-09979-5] [Medline: 41249645]
26. Bing-You R, Varaklis K, Hayes V, Trowbridge R, Kemp H, McKelvy D. The feedback tango: an integrative review and analysis of the content of the teacher-learner feedback exchange. Acad Med. Apr 2018;93(4):657-663. [doi: 10.1097/ACM.0000000000001927] [Medline: 28991848]
27. Lee GB, Chiu AM. Assessment and feedback methods in competency-based medical education. Ann Allergy Asthma Immunol. Mar 2022;128(3):256-262. [doi: 10.1016/j.anai.2021.12.010] [Medline: 34929390]
28. Li SX, Li CMF, Jenkins ME, Venance SL, Florendo-Cumbermack A. Insights from the transition to competency-based medical education in neurology programs. Can J Neurol Sci. Nov 23, 2023:1-5. [doi: 10.1017/cjn.2023.318] [Medline: 37994542]
29. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE guide no. 86. Med Teach. May 2014;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
30. Wright DS, Kanaparthy NS, Melnick ER, et al. The effect of ambient artificial intelligence scribes on trainee documentation burden. Appl Clin Inform. Aug 2025;16(4):872-878. [doi: 10.1055/a-2647-1142] [Medline: 40602775]
31. Olson KD, Meeker D, Troup M, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open. Oct 1, 2025;8(10):e2534976. [doi: 10.1001/jamanetworkopen.2025.34976] [Medline: 41037268]
32. Debnath T, Siddiky MN, Rahman ME, et al. A comprehensive survey of prompt engineering techniques in large language models. TechRxiv. Preprint posted online on Oct 28, 2025. [doi: 10.36227/techrxiv.174140719.96375390/v1]

Abbreviations

AI: artificial intelligence
EFeCT: Evaluation of Feedback Captured Tool
LLM : large language model
SP: standardized patient

Edited by Marco Montagna; peer-reviewed by Fumitoshi Fukuzawa, Olivia Ng, Yoshikazu Asada; submitted 19.Dec.2025; final revised version received 09.Apr.2026; accepted 29.Apr.2026; published 28.May.2026

Please cite as:
Talwalkar JS, Chartash D, Zhang L, Makutonin M, Safranek CW, Sidamon-Eristoff AE, Schwamm LH, Wright DS
Ambient AI Scribes to Create Educational Feedback Notes for Medical Students: Randomized Trial
JMIR Med Educ 2026;12:e89996
URL: https://mededu.jmir.org/2026/1/e89996
doi: 10.2196/89996