SENRS 50th Annual Meeting | Electronic Scientific Exhibit

Head-to-Head: AI Readouts vs. Attending Readouts on a Diagnostic Radiology Subinternship

Musa Malik[1], Ibrahim Kaouk[1], Christian Cordon-Cano[1], Aisha Nur[1], Brandon Garcia Fuentes[1], Michael Fei[2], Avery McCutchen[1], Elinor Grage[1], Kathleen Williams[1], Erica Scherer[1], Kevin Hiatt, MD[1]

[1]Wake Forest University School of Medicine [2]Creighton University School of Medicine

Wake Forest University School of Medicine

The academic core of Atrium Health

Background

- Readouts, where a student dictates a case and reviews it with an attending, are a core part of the diagnostic radiology (DR) subinternship.
- Rising clinical workloads limit the time attendings have for supervision and real-time feedback.[1]
- The cases a student sees depend on the day's worklist, so case volume and pathology differ from student to student.
- Large language models (LLMs) can compare a trainee's report with a reference report and return written feedback.[2–4]
- Published studies of LLM feedback in imaging education have enrolled residents and fellows. Evidence in medical students is limited.[4,5]

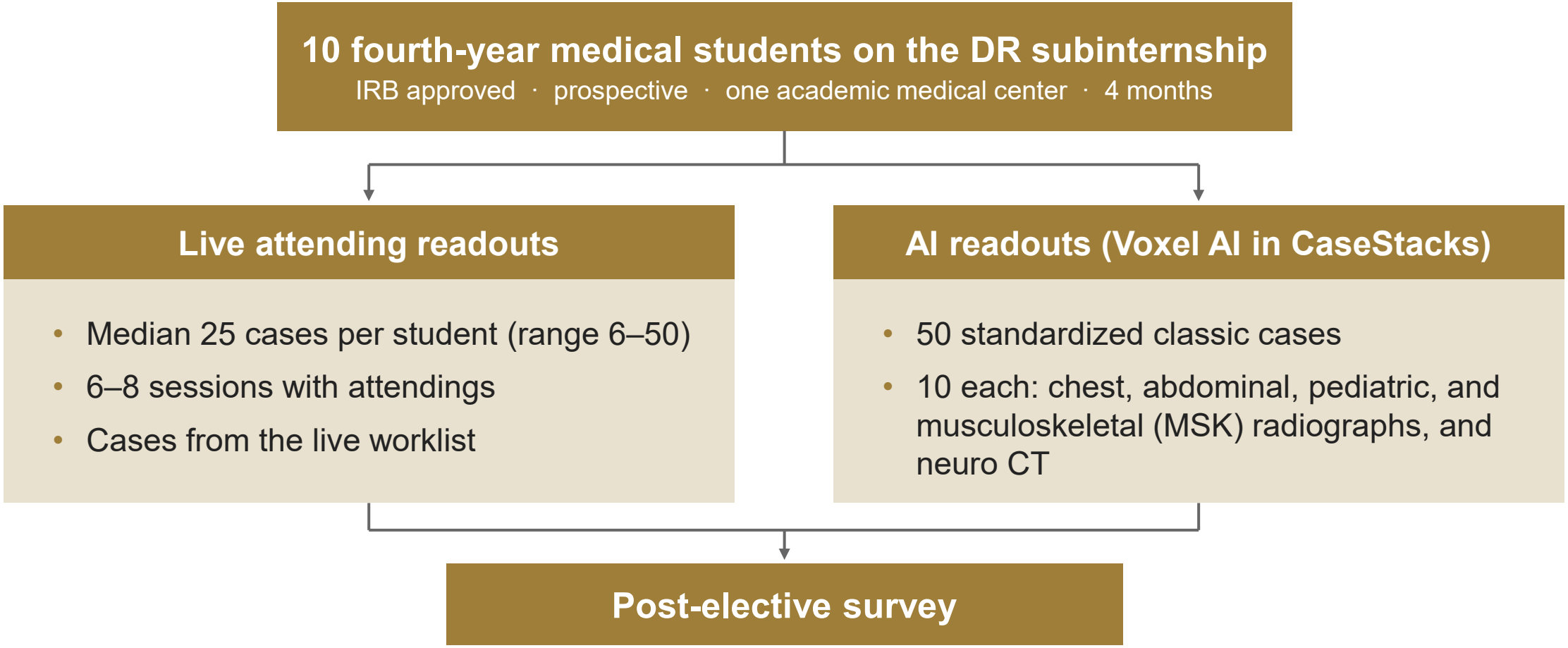


Purpose

To compare AI readouts on a standardized case set with live attending readouts, as rated by the students who completed both.

- How do students rate AI and attending feedback on the same domains?
- Where does AI feedback fall short?
- How much do students trust it, and what role do they want it to have?

Methods: Study Design



Every student completed both arms. The order of AI and live readouts varied by student.

Methods: How the AI Feedback Worked

1 → 2 → 3 → 4

Student reports
Typed or dictated

Gemini compares
Report vs. radiologist reference, text only

Score and feedback
Shown after each 10-case module

Voxel AI tutor
Follow-up questions in chat

Grading: Google Gemini 3 Flash

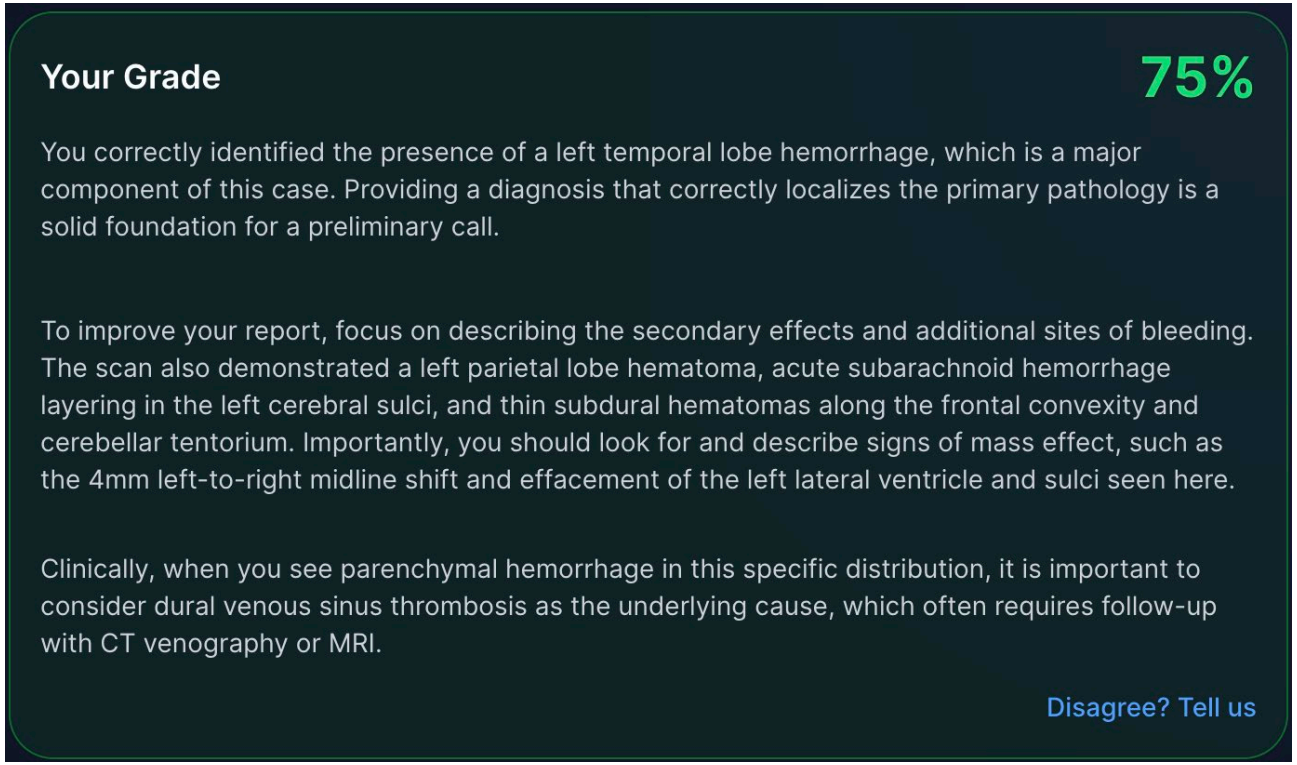
- Text only: the model never saw the images
- Compared the report with the radiologist's reference report, key findings, and diagnosis
- Scored 0–10: diagnosis first, then key findings, then report quality
- Feedback opened with what the student did well
- Lenient, encouraging mode for medical students

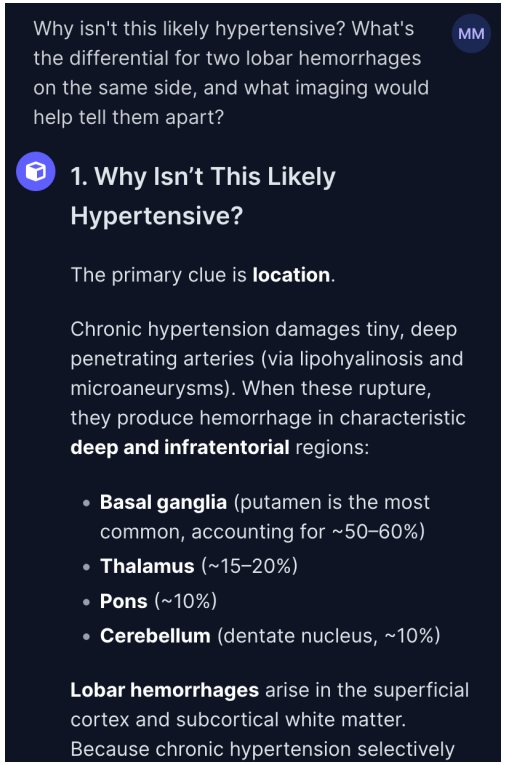
Voxel AI tutor: Gemini Flash

- AI tutor chat panel, available on every case
- Used the case's teaching write-up as context
- Adjusted explanations to the learner's training level
- Could search and cite an open-access radiology textbook

Grading model: gemini-3-flash-preview. Dictated reports were transcribed with Google speech-to-text.

Methods: Example AI Readout





Example on a practice case. The submitted report called only a left temporal hemorrhage and attributed it to hypertension. The feedback (left) named the missed parietal hematoma, the subarachnoid and subdural blood, and the midline shift, and raised venous sinus thrombosis. The student then asked the tutor why hypertension was unlikely (right).

Methods: Survey and Analysis

One post-elective survey, with the same wording for AI and attending feedback

Rated 1–5 for each source

- Error identification
- Terminology
- Report structure
- Urgent findings
- Specificity

Also asked

- Forced choice: which source did better in each domain
- Usefulness by case type
- Trust and attitudes
- Concern about AI, before and after
- Free-text comments

Analysis

- Paired AI vs. attending items: Wilcoxon signed-rank test
- Free text: coded by theme
- n = 10; every student rated both sources

1 = strongly disagree, 5 = strongly agree.

Methods: Matched Survey Items

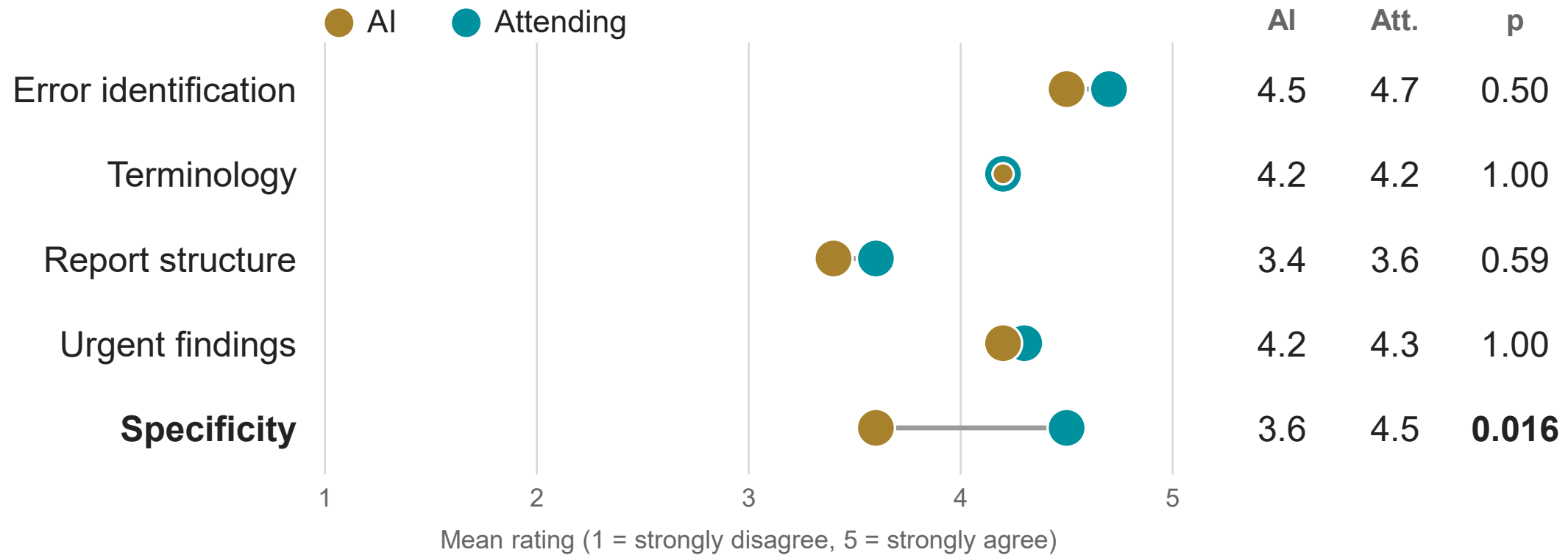
Domain	Item, asked once for attending feedback and once for AI feedback
Error identification	"…identified errors I had made in describing imaging findings (e.g., missed findings, mischaracterized findings)."
Terminology	"…corrected my use of radiology-specific terminology and descriptors."
Report structure	"…addressed the structure and organization of my report (e.g., findings vs. impression, ordering of findings by clinical significance)."
Urgent findings	"…addressed my communication of clinically actionable or urgent findings."
Specificity	"…was specific enough that I knew exactly what to change in my next dictation."

Results: Participants

Characteristic	Students (n = 10)
Intended specialty	Diagnostic radiology 6, interventional radiology 1, other 3
Prior dictation experience	None 4, a few times 5, extensive 1
Live cases dictated with attendings	Median 25 (range 6–50)
AI case-feedback encounters	500 in total (50 per student)
Comfort with AI tools before the elective	Mean 3.5 of 5
How closely they read the AI feedback	Every comment 1, most comments 7, skimmed 2
Time per case (self-reported)	AI 13 min, attending 18 min

Other specialties: internal medicine, pathology, and physical medicine and rehabilitation.

Results: Head-to-Head Ratings



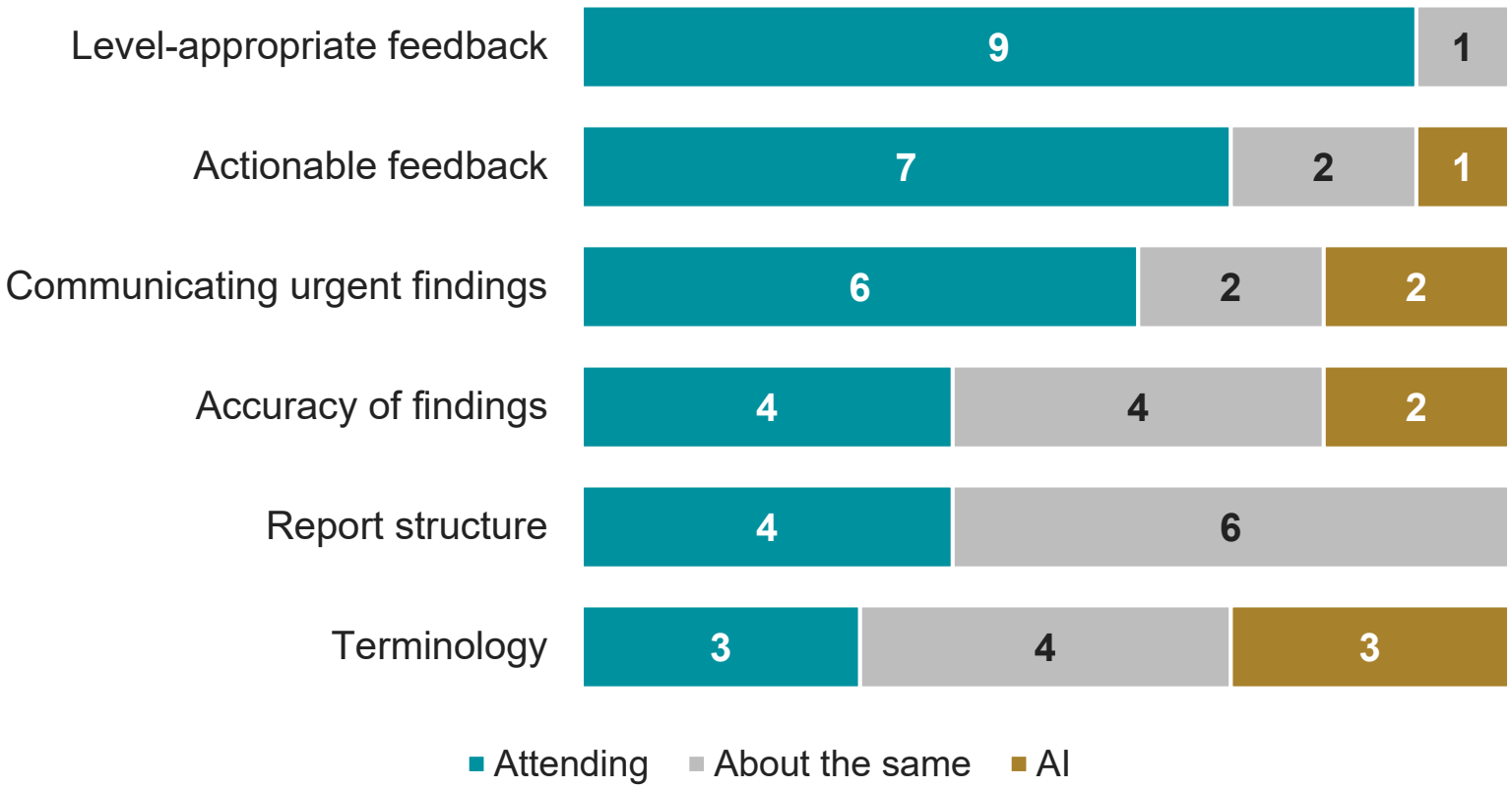
	AI	Att.	p
Error identification	4.5	4.7	0.50
Terminology	4.2	4.2	1.00
Report structure	3.4	3.6	0.59
Urgent findings	4.2	4.3	1.00
Specificity	3.6	4.5	**0.016**

- No significant difference in 4 of 5 domains
- Attending feedback rated higher for specificity: knowing exactly what to change next

Wilcoxon signed-rank test, n = 10. In a sample this small, a non-significant difference does not establish equivalence.

Results: Forced-Choice Comparisons

Which source did better? (number of students)

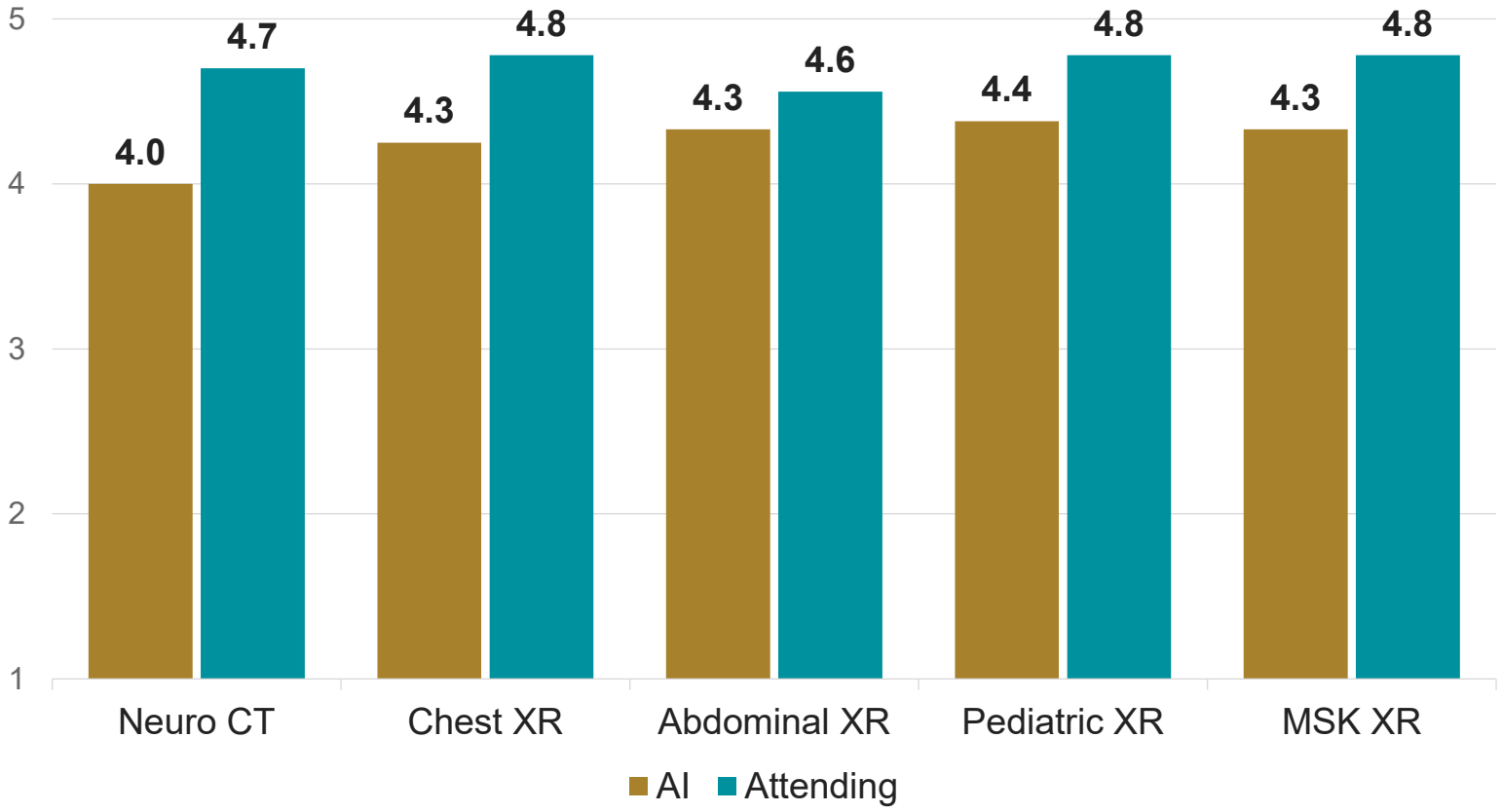


- Attendings were preferred for level-appropriate (9 of 10) and actionable (7 of 10) feedback
- Terminology was an even split
- AI was never the majority choice

n = 10. "Slightly more" and "substantially more" responses are combined for each source.

Results: Usefulness by Case Type

Mean usefulness rating (1–5)



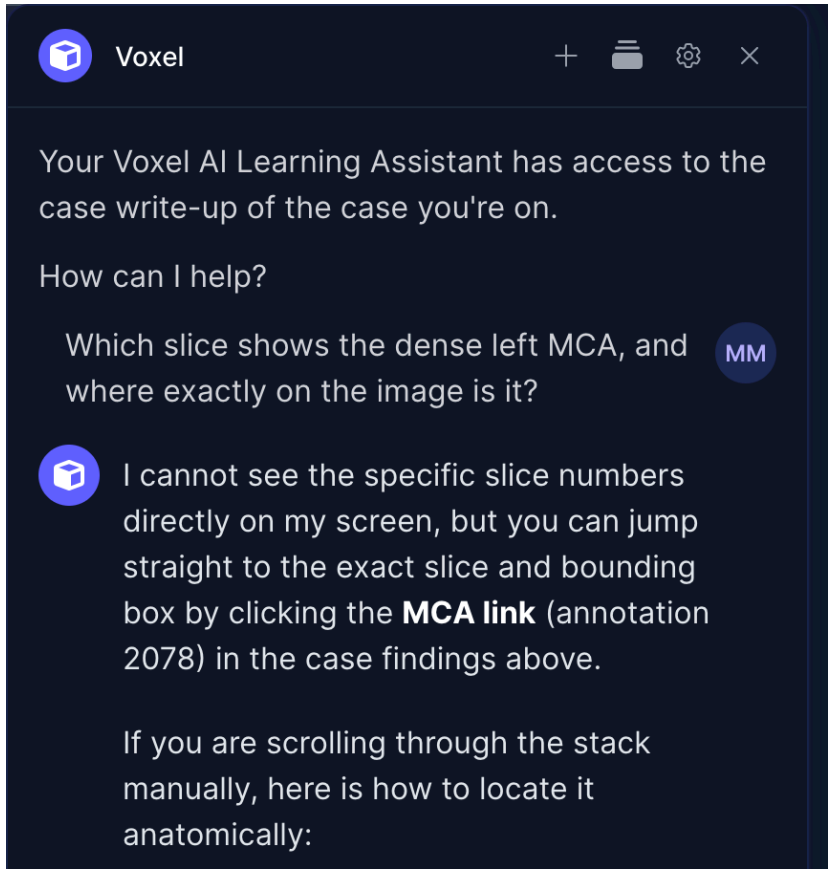
- Neuro CT, the only cross-sectional set, had the lowest AI rating and the widest gap
- Attending feedback rated 4.6 to 4.8 for every case type
- No paired difference was significant (p = 0.13 to 0.63)

Students could mark N/A, so n = 8–10 per bar. Wilcoxon signed-rank test on paired responses.

Results: Where AI Fell Short

- 8 of 10 said the AI could not reliably show where a finding was on the image
  - They asked for annotation, a way to point at a structure and ask about it, and multiplanar views
- The grading model compared text only and never saw the images
- No student reported a hallucination in 500 case-feedback encounters
  - 1 student described the model agreeing with an incorrect finding the student defended

"The most significant limitation was the inability to put … an arrow [to] see exactly where the finding was on the image."



Example: asked where the dense MCA is, the Voxel AI tutor pointed to a pre-made annotation instead of locating it on the image.

Results: What Students Wrote

Attending feedback: most useful	
6	Showed how they approach a study
3	Pointed to findings on the image
3	Real-time back-and-forth
2	Feedback matched to their level

AI feedback: most useful	
5	Fast, and available on their own time
2	Explained findings they had missed
2	Helped with report wording
2	Follow-up questions or self-testing

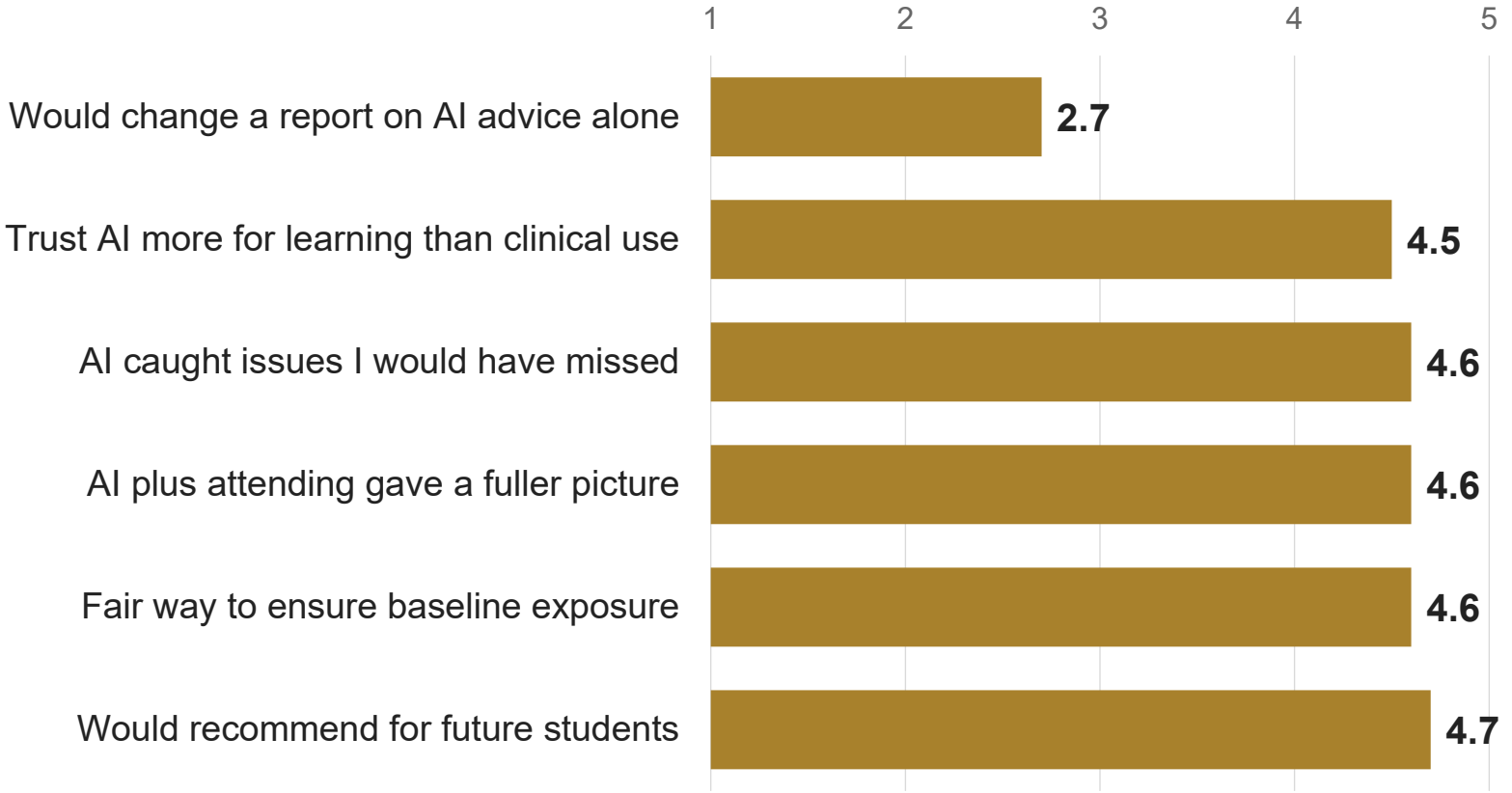
Attending feedback: main limitation	
8	Limited attending time
1	Case mix on the worklist
1	Variable teaching between attendings
1	No way to review feedback later

AI feedback: main limitation	
7	Could not show where a finding was
2	Not matched to their level
1	Unclear how reports were graded
1	Retained less from AI feedback

Number of students whose free-text answer raised each theme (n = 10; an answer could raise more than one). Not every theme raised by a single student is listed. An eighth student raised the spatial limitation under suggestions.

Results: Trust and Attitudes

Mean agreement with each statement (1–5)



7 of 7

students who got conflicting feedback trusted the attending

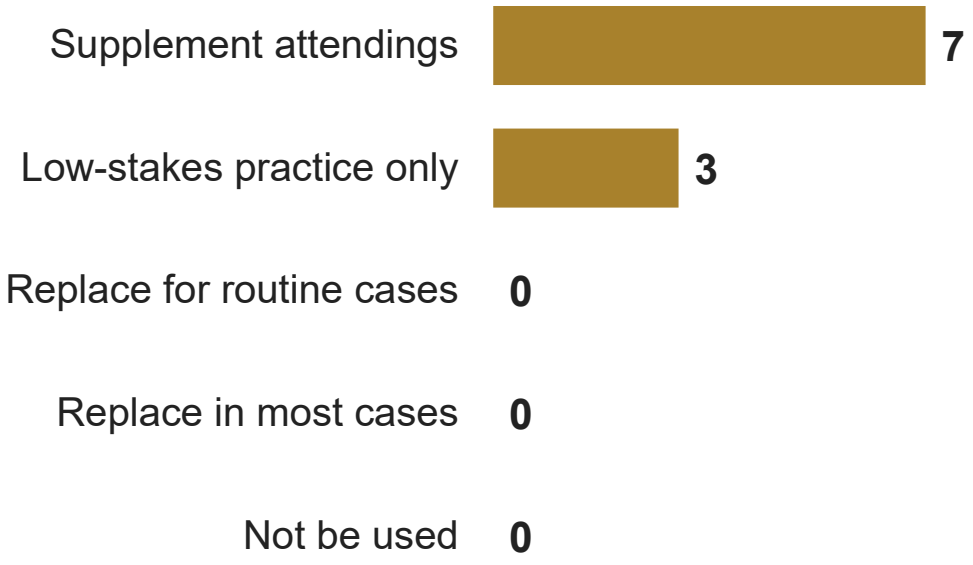
8 of 10

spent less time per case with AI (13 vs. 18 minutes)

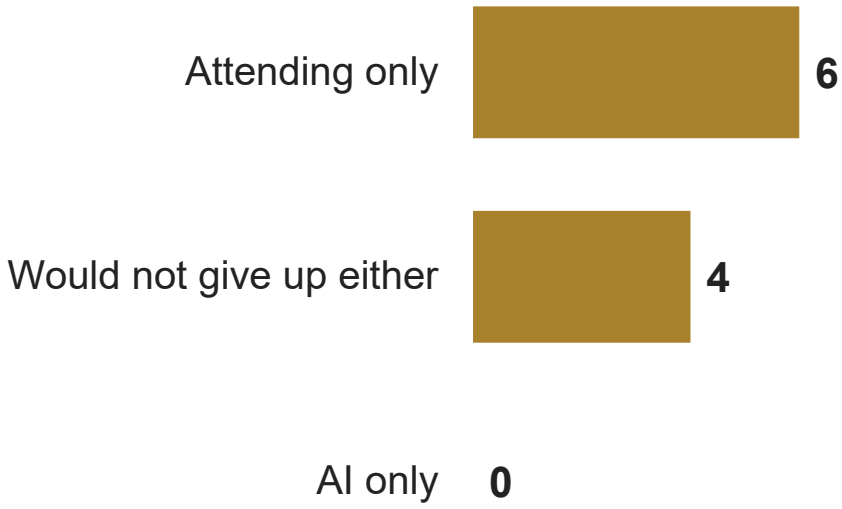
n = 10. Time per case was self-reported. The other 3 students did not encounter conflicting feedback.

Results: Future Role of AI Readouts

What role should AI readouts play?



If you could keep only one type of readout



Concerned about AI's impact on radiology: **5 of 10 before → 2 of 10 after**

Number of students, n = 10. Mean concern fell from 3.4 to 2.5 of 5, which was not statistically significant (p = 0.13).

Limitations and Next Steps

Limitations

- 10 students at one site
- Self-reported ratings, not measured report accuracy
- AI and attending arms used different cases
- Order of AI and live readouts was not controlled
- The AI graded text only, in a lenient mode
- The survey was not anonymous

Next steps (from student suggestions)

- Mark findings on the image, and let students point at a structure and ask
- Give feedback after every case, not after each 10-case module
- Add a structured report template, as in PACS
- Label feedback by training level
- Measure report accuracy over time

Conclusion

- Students rated AI feedback similarly to attending feedback on error identification, terminology, report structure, and urgent findings.
- Attending feedback was more specific and better matched to the student's level.
- The main limitation of AI feedback was spatial: a text-only model could not show where a finding was.
- A standardized AI case set can increase the number of cases and the range of pathology each student dictates.
- Students favored keeping AI readouts as a complement to attending readouts.

References

1. Lyo SK, Cook TS. Integrating large language models into radiology education: an interpretation-centric framework for enhanced learning while supporting workflow. J Am Coll Radiol. 2025;22(11):1271-1280. doi:10.1016/j.jacr.2025.07.003

2. Bala W, Li H, Moon J, Trivedi H, Gichoya J, Balthazar P. Enhancing radiology training with GPT-4: pilot analysis of automated feedback in trainee preliminary reports. Curr Probl Diagn Radiol. 2025;54(2):151-158. doi:10.1067/j.cpradiol.2024.08.003

3. Lyo S, Mohan S, Hassankhani A, Noor A, Dako F, Cook T. From revisions to insights: converting radiology report revisions into actionable educational feedback using generative AI models. J Imaging Inform Med. 2025;38(2):1265-1279. doi:10.1007/s10278-024-01233-4

4. Al-Mashhadani MM, Ajaz F, Guraya SS, Ennab F. Use of large language models for providing automated feedback in medical imaging education: a systematic review. Front Med (Lausanne). 2026;13:1803921. doi:10.3389/fmed.2026.1803921

5. Kıyak YS, İş-Kara T, Emekli E. Applications and outcomes of large-language-model-generated feedback in undergraduate medical education: a scoping review. Med Sci Educ. 2026;36(1):81-99. doi:10.1007/s40670-025-02621-3

Contact: Musa Malik, musa.malik@wfusm.edu