How Icarus Learned Altitude Control: A Cautious Path Leveraging AI to Support Assessment & Evaluation

AEO
OFFICE OF ASSESSMENT, EVALUATION, AND OUTCOMES

Matt Zuckerman, MD
Sean Marshall, MA
Bonnie Kaplan, MD
Tai Lockspeiser, MD, MHPE

AI hub

When I say

Artificial Intelligence & **evaluation**

what words come to mind?

Objectives

- Describe ways AI is currently being used or piloted in health professions assessment and evaluation.
- Apply an “altitude control” framework to decide when AI should augment, rather than automate, assessment decisions.
- Identify specific guardrails (technical, ethical, and educational) for deploying AI in assessment.

Why Assessment & Evaluation Are a Dangerous Place to "Fly"

- Assessment in med ed is high stakes: progression decisions, residency selection, remediation, professionalism
- We are already data-rich but time-poor: narrative comments, OSCE checklists, EPA forms, MSPEs, multisource feedback

The New York Times Magazine

I'm a Professor. A.I. Has Changed My Classroom, but Not for the Worse.

My students' easy access to chatbots forced me to make humanities instruction even more human.

November 25, 2025

Faculty want to fly

Teaching Efficiency

- Major time savings in preparing content, summarizing literature, and creating questions

Enriched Learning Materials

- Allows visualizations, animations, simulations, personalized content, and rapid redesign of teaching media

Individualized Learning Support

- Instant feedback
- Tailored instructional pace

Kastamonu Education Journal. 2025;33(3):448-468.

Faculty describe shifting from knowledge transmitter to facilitator/coach as AI manages routine cognitive load

Kastamonu Education Journal. 2025;33(3):448-468.

Too high

- Bias and inequity
- Illusion of false objectivity
- Tone normalization erases nuance
- Deskilling faculty
- Misalignment with CBME undermines assessment
- Depersonalization and loss of professional autonomy

Templin T, et al. Carolina Digital Repository doi:10.17615/7G2G-BK38; Gordon M, et al. Medical Teacher. 2024;46(4):446-470, Roveta A, et al. AI. 2025;6(9):227. Janumpally R, et al. Front Med. 2025;11. doi:10.3389/fmed.2024.1525604; Yetik SS. Kastamonu Education Journal. 2025;33(3):448-468.

Best AI Grading Tool?

I have more students than I've ever had this year and need to get some more digital/self grading done.

The assessments I use will be a combination of multiple choice and short answer questions. I need to have students write short responses to science questions but haven't found anything that helps grade this part accurately. I have tried Google Forms and although it works great for multiple choice, it has not helped much with the written response questions.

Any suggestions? I would be willing to pay if it's really worth it. Thank you.

0 16

Strangextown • 3mo ago

Maybe sit down and do your job, I don't know. Why would you make them write tests that you won't even check in the first place?

4

ADHTeacher • 3mo ago

Are you going to get your students' explicit, informed consent before feeding their data to AI?

8

Strangextown • 3mo ago

This. I'd be absolutely livid in the students' place.

3

Titsnium • 3mo ago

For short answers, Gradescope has been the most accurate time-saver for me.

Set up a short-answer assignment in Gradescope and turn on AI-assisted grouping; it clusters similar responses so you apply a rubric once to a batch. Build a tight rubric with 3–5 items that mirror your science goals: correct concept, evidence/reference to the prompt, use of vocab, and clear reasoning. Add common deductions (misused term, missing mechanism, unsupported claim) and use keyboard shortcuts to fly through stacks. For quick auto-points before you review, Formative (GoFormative) or Edulastic can auto-score short responses based on required keywords/synonyms with partial credit, then you do a fast manual pass for reasoning quality. If you're on paper for MC, ZipGrade or Akindi handle bubbles; otherwise keep Google Forms for the MC piece.

I lean on Gradescope and Formative/Edulastic for scoring; if a batch looks pasted, I'll spot-check with Smodin's AI/plagiarism detection to keep things clean.

Bottom line: Gradescope + a concise rubric will cut your short-answer grading time the most.

0

Found in FEATURE

AI in the Sky: How Artificial Intelligence and Aviation Are Working Together

Artificial intelligence (AI – also related to Machine Learning, or "ML" as it's called) has reached new heights: a cruising altitude of 10,000 - 70,000 feet to be precise.

VANCE HILDERMAN

AI is my CoPilot: DO-178C

Key international standard for certifying software in commercial aircraft systems

Safety measures include:

- Installing an **external monitor** to assess the decisions of the AI engine from a safety perspective
- Building **redundancy** into the process as a safeguard
- Reverting to a **default safe mode** when unknown or dangerous conditions occur
- Keeping a **human in command** or in the loop
- Monitoring AI through an **independent AI agent**
- Examining the AI output through a traditional backup system, or **safety net**

Use AI to augment human judgment where:

- The task is labor-intensive, repetitive, or pattern-based.
- The stakes are moderated by robust human oversight.
- You are gathering additional signals rather than replacing existing ones.

EASA, EASA. "Easy Access Rules for Acceptable Means of Compliance for Airworthiness of Products, Parts and Appliances (AMC-20)." (2018).

Feature	Automatic Assessment Tools (MCQs, Checklists)	Traditional Machine Learning (Essays)	LLMs (Narrative Notes)
Definition	Fully rule-based automated scoring	Model-based scoring using structured features and labeled datasets	AI-assisted grading using unstructured text and prompt-based outputs
Input	Structured (MCQ answers, OSCE checklist ticks)	Structured (rubric scores, extracted features)	Unstructured free text (student notes)
Training Data	Predefined answers and scoring keys	Requires labeled data with scores and rubric alignment	Fine-tuned prompts using generalized LLMs
Interpretability	High — rules are explicit and explainable	Moderate — interpretable models (e.g., SHAP, weights)	Often opaque — black-box outputs requiring prompt engineering
Flexibility	Low — best for structured, binary input	Low — limited to engineered features	High — can process diverse formats and nuanced expression
Rubric Integration	Direct scoring of checklist or key	Rubrics converted into numeric features	Rubrics converted into prompt structure
Scalability	High — efficient for large-scale objective testing	Limited — separate models needed for each task/context	High — few-shot/zero-shot generalization possible
Performance	High validity/reliability for structured domains	Validated against human scoring	Depends on prompt quality and alignment
Human Oversight	Minimal	Moderate	Essential
Risks	Oversimplification, misalignment with learning goals	Rigid scoring, poor generalization outside trained domain	Hallucination, bias, over-reliance on flawed prompt design

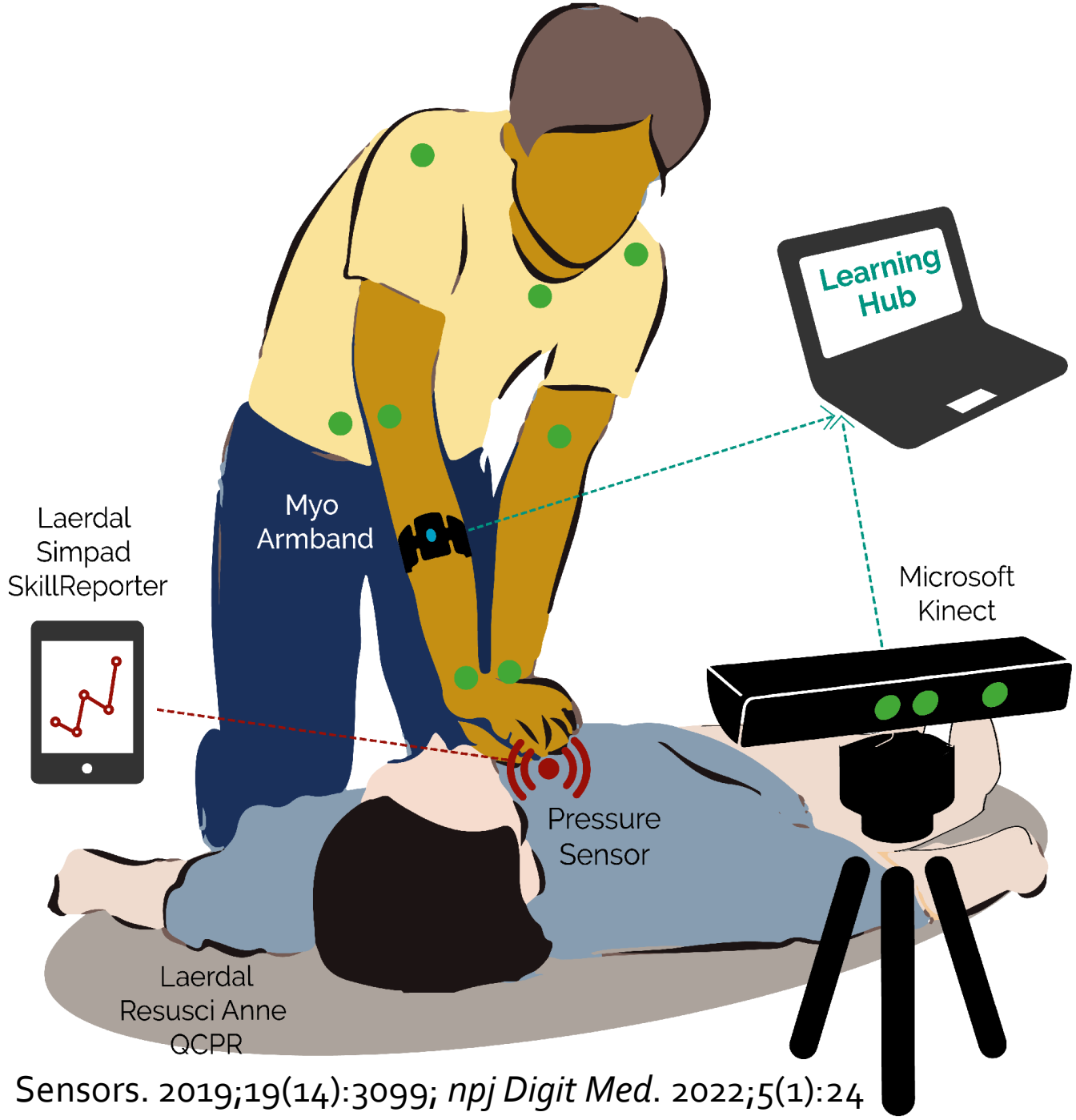
ASC Office of Distance Education; JGIM. 2022;37(9):2230-2238; *NEJM AI*. 2024;1(12):AIcs2400631.

Examples

Technical Skills

•AI-based combines sensors: body-motion tracking + electromyogram + standard manikin data
•Trained neural-network models to detect mistakes in those indicators looking at: rate, depth, release, arm posture, body-weight/posture
•Their system matched the commercial manikin's accuracy on standard metrics (rate, depth, release), and **also** detected additional mistakes: poor arm posture, incorrect body weight usage.

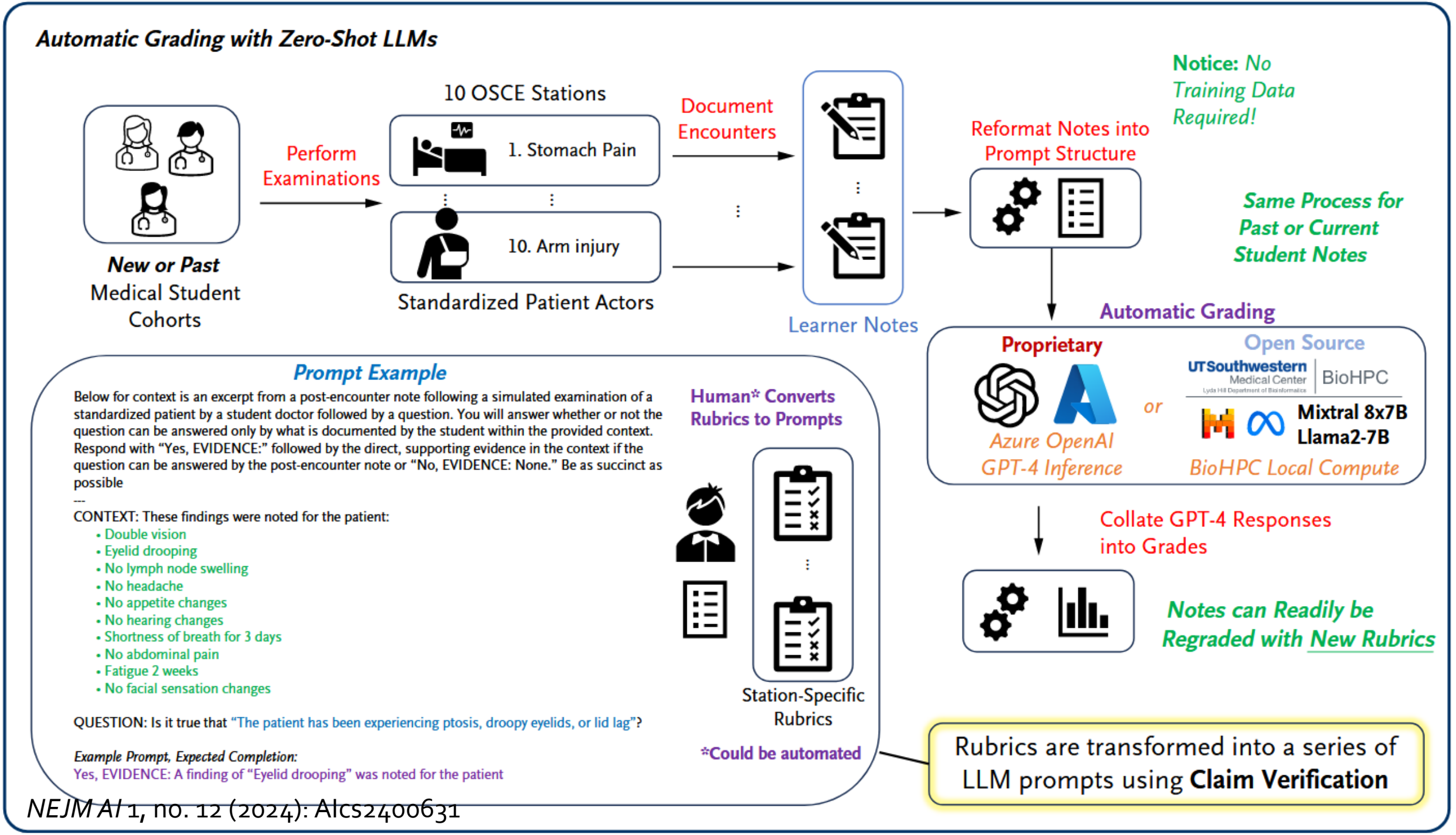
•Additional studies have looked at videos of knot tying skills, endoscopic surgical skills,



Sensors. 2019;19(14):3099; *npj Digit Med*. 2022;5(1):24

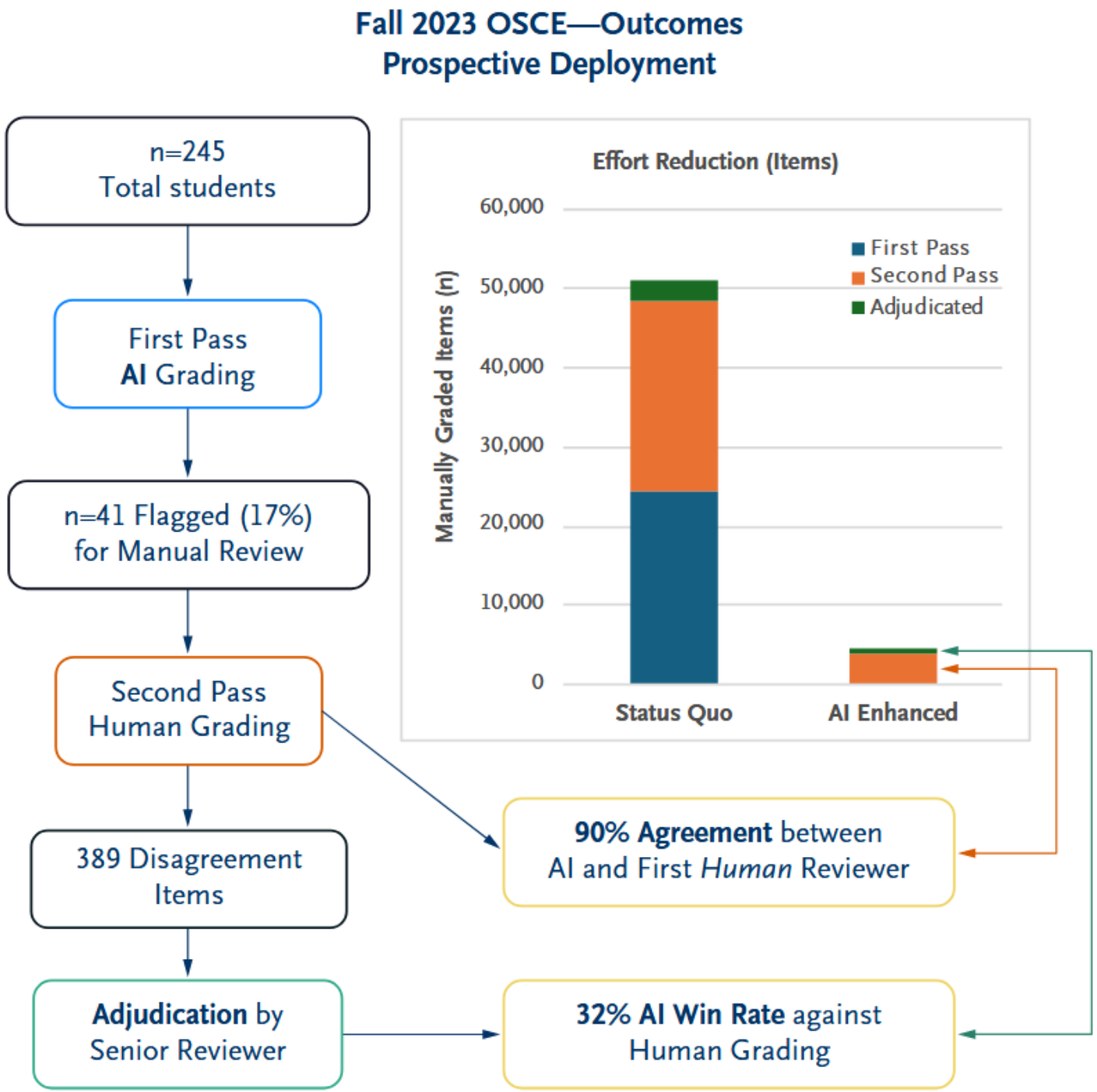
Good OSCE?

- Rubric-based assessments of student notes (from a high-stakes clinical skills exam) with feedback and scores generated using GPT-4 via Microsoft Azure
- Prompts were designed to simulate the rubric criteria and ask the AI to generate feedback and scores.



NEJM AI 1, no. 12 (2024): AIcs2400631

Score Agreement:
•AI scores correlated moderately with human scores in most domains.
•Agreement was highest in the “appropriate language and structure” domain and lowest in clinical reasoning documentation.
Feedback:
•AI-generated feedback was broadly accurate but often lacked the depth or specificity of faculty comments.
Efficiency:
•Once prompts were refined, AI assessments required **seconds** per note compared to **minutes** for faculty reviewers.
Prompt Engineering Matters:
•Sequential prompting yielded slightly better alignment with human scoring than simultaneous prompting.
•The design and calibration of prompts significantly affected output quality.



NEJM AI 1, no. 12 (2024): AIcs2400631

Our First Flight

Medical Student Performance Evaluation (Dean's Letter)

- Estimated 30,000 – 35,000 MSPEs annually in the U.S.
- AAMC now recommends standardized formatting and language
- About 4 in 5 program directors prioritize academic history, professionalism, academic progress, and a **summary paragraph**
  - May grow in importance as Step1 is pass/fail

Med Educ Online 2021;26(1):1876315; Clin Orthop Relat Res 2021;479(8):1679-1687

School of Medicine
UNIVERSITY OF COLORADO
ANSCHUTZ MEDICAL CAMPUS

Office of Student Life
Fitzsimons Bldg N1219
13001 East 17th Pl, Mail Stop C292
Aurora, CO 80045
P 303-724-6407
F 303-724-6409

MEDICAL STUDENT PERFORMANCE EVALUATION
For
XXXXXXX
September 20XX

XXXXXXX XXXXXXX is a fourth-year student at the University of Colorado, School of Medicine in Aurora, Colorado.

NOTEWORTHY CHARACTERISTICS

ACADEMIC HISTORY

Date of Initial Matriculation in Medical School	XXXXXXX
Date of Expected Graduation from Medical School	XXXXXXX
Please explain any extensions, leave(s), gap(s) or break(s) in the student's educational program.	XXXXXXX
Not Applicable or definition	
Information about the student's prior, current, or expected enrollment in, and the month and year of the student's expected graduation from dual, joint, or combined degree programs.	XXXXXXX
Was the student required to repeat or otherwise remediate any course work during her medical education? If yes please explain.	No
Add explanation or delete if no.	
Was the student the recipient of any adverse action(s) by the medical school or its parent institution?	No

ACADEMIC PROGRESS

Professional Performance*

Preclinical Coursework

Clerkships
Official transcript grades in the Clinical years are Honors, High Pass, Pass, Pass with Remediation, or Fail. The NBME Subject Exams are used in several disciplines and are

Pass/Fail. A student who receives a failing score the first time they take a Subject Exam is no longer eligible for Honors; students who fail a Subject Exam twice or more are eligible only for Pass with Remediation once they pass the exam. Our grading is criterion based and grades are assigned by discipline specific grading committees.

DHLIC-MSPE Summary PASS (Pass/Fail Only)

DHLIC-Emergency Medicine PASS (Pass/Fail Only)

DHLIC-Family Medicine HIGH PASS

DHLIC-Internal Medicine HONORS

DHLIC-Obstetrics and Gynecology PASS

DHLIC-Pediatrics PASS

DHLIC-Psychiatry HONORS

DHLIC-Surgery HIGH PASS

Process

Assessment Comments

Example comments & MSPE specific to grade and course

Draft

Verification

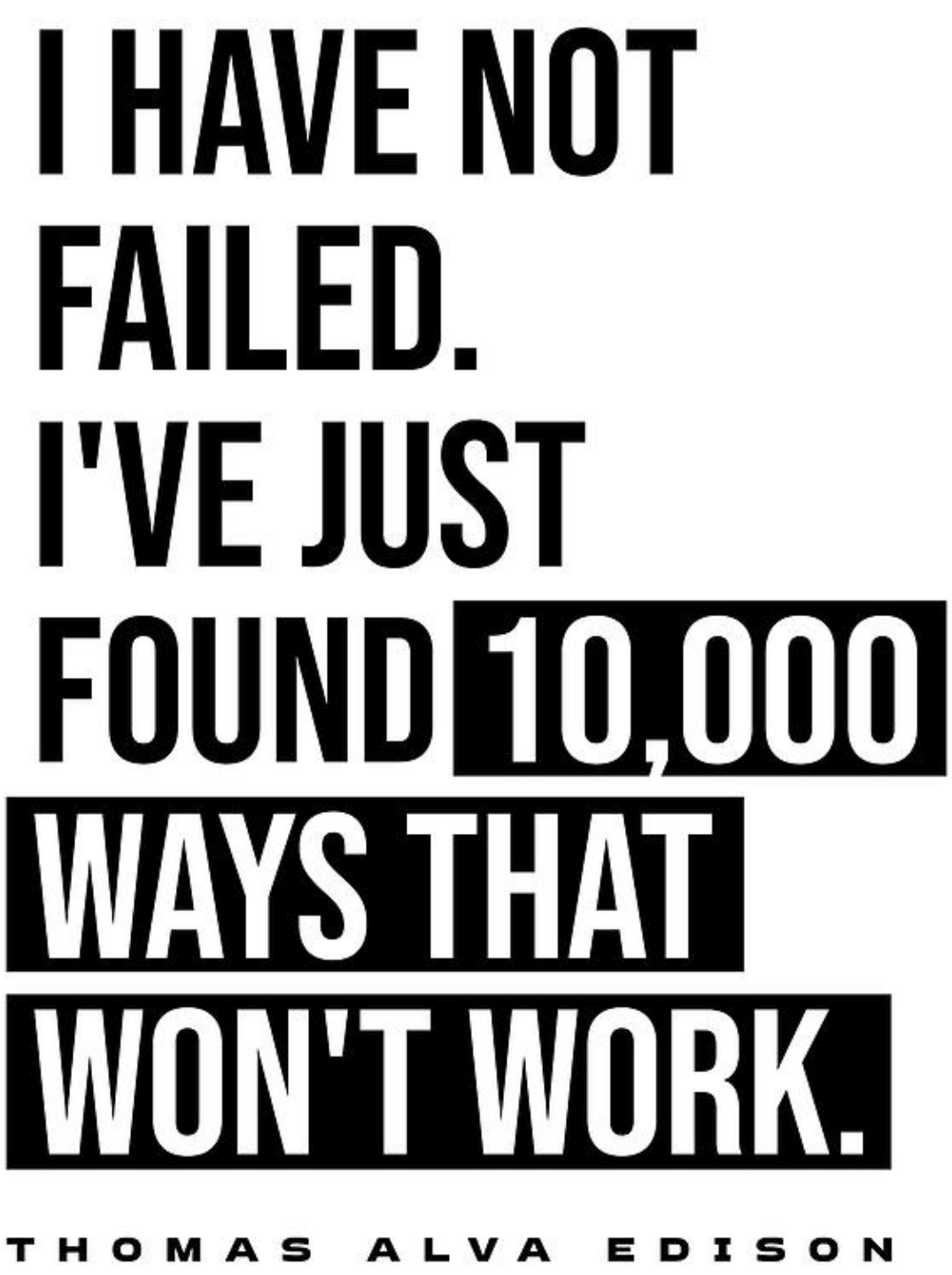
Approval

1. Automated check against comments
2. Team review
3. Dean review
4. Student review

Beyond CoPilot: Programming and Prompt Generation

```
# Fetch Current Student Data

# Find preferred names from the comments

# Load few-shot examples for specialty/grade

# Build the system prompt

# Send to Azure OpenAI Chat Completions

# Save summaries to SQL
```

```
# Build the system prompt
system_message = f"""
You are an expert medical education evaluator. Your task is to generate narrative MSPE (Medical Student Performance Evalu

Your output must reflect the voice and judgment of a faculty member who observed the student in that specialty. Your tone

Follow these rules carefully:

TONE AND STRUCTURE:
- Each summary must be a single paragraph **concise (120-145 words)** and strengths-focused, with tone modulated to matcl
- Use professional, academic tone that is evaluative and narrative, not reflective, not conversational, not informal.
- Use **natural human rhythm and narrative variation** — as if each summary were written by a different evaluator.
- Use a mix of sentence lengths and formats. Avoid starting every sentence or summary with the student's name or verbs li
- Use varied verbs across summaries: “applied,” “led,” “developed,” “refined,” “responded,” etc.
- Do NOT use phrases like "commendable," "excellent clinical skills," "strong work ethic," or “ready for residency.”
- DO NOT end with a generic conclusion. You may end mid-action or on a specific clinical insight — no closing line is rec

CLOSING SENTENCE RULES:
- Do NOT end with generic phrases like:
- “The student is prepared for residency.”
- “They left a lasting impression.”
- “He consistently impressed throughout the rotation.”
- “Their performance demonstrates readiness for clinical training.”
These are generic and unhelpful.

Instead, allow the summary to end on:
- A specific skill the student applied
- A patient case they contributed to
- A reflection or insight from the evaluator comments

Let the summary stop mid-action or at the resolution of a case. Do NOT add any concluding or summarizing line unless it i
```


I HAVE NOT
FAILED.
I'VE JUST
FOUND 10,000
WAYS THAT
WON'T WORK.
THOMAS ALVA EDISON

Learning Lessons

- Keeping each specialty data set separate to avoid mis-categorization
- Asking prompt to emphasize later evaluations to reflect growth
- Few shot examples from past cohort to maintain tone and language
  - Match examples based on specialty and grade
- Prompt engineering to avoid repetitive language
- Coding to check for consistent name and gender
- Multi-agent process to allow for generation, validation, and improvement

Taste Testers

- Longitudinal Integrated Curriculum directors
- Leadership from Office of Student Life and Office of Assessment, Evaluations, and Outcomes
- Residency Program Directors
- Specialty Advisors
- Clinical Content Directors

Behind the curtain

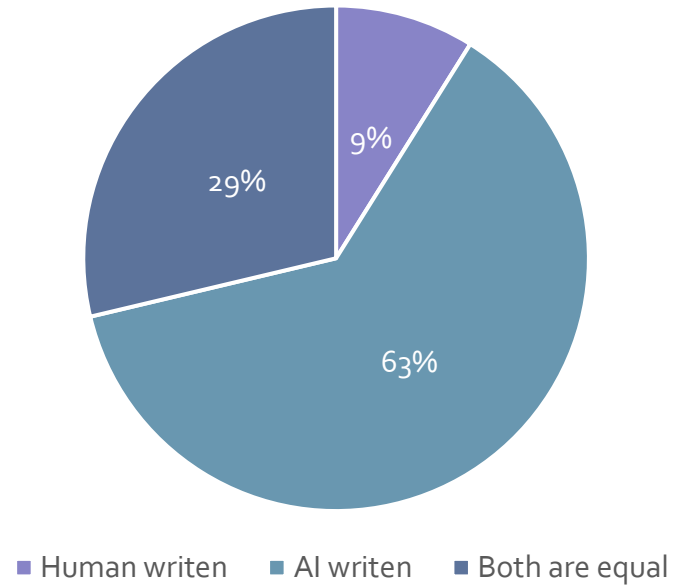
Is this summary paragraph written by an LLM or by a human?

65% accuracy

Preference

All of the AI generated paragraphs were preferred

Which of these two paragraphs do you prefer?



Sometimes we root for the home team

I think version B is the AI generated one, and if I am right I find those often seem **very generic and level out differences between students.** Also, I think "a respectable differential" and "getting tasks done quickly and with high quality" **don't sound like what a person would say** - the second phrase sounds grammatically incorrect, or at least awkward.

Version B is human generated

Efficiency

- Generated 189 letters x 7 paragraphs = 1,323 paragraphs
- 5 minutes per paragraph vs 20 minutes per paragraph

Time Saved
15 min x 1,323 paragraphs
= 330 hours

OSL Survey

- Administrative stakeholders who review letters with students reported needing to edit five or fewer, usually for tone or minor language adjustments.
- They rated the overall quality of letters as excellent or good compared with prior years, and no one preferred returning to manual drafting.
- Noted that the quality of the letters improved over time (prompt iteration, process improvement)

Guardrails and Flight plans

Abstinence is not an Option

Educators will have to place themselves on the continuum between the two poles of:

(a) Ignoring AI tools and (b) Educating about AI tools.

And on the other perpendicularly intersecting axis

(a) Banning AI tools and (b) Whole-heartedly adopting them.

The aspiration to ban such technologies is a losing proposition; embracing them and using them judiciously is a better and more productive choice

International Journal for Educational Integrity 19, no. 1 (2023): 9

NBME

Data Privacy

Incorporate clear rules and assurances of how data will be collected, how personal data will be protected, how long data will be retained, and whether it's acceptable to share or repurpose data.

Accountability

Consider regulatory requirements regarding the development and deployment of AI technologies to ensure compliance. Be prepared to explain the aims, motivations and reasons for using AI.

Fairness

Begin with the assumption that the use of AI is never neutral or impartial, and acknowledge that the use of AI, especially for tasks like automated decision-making, could lead to discriminatory outcomes.

Transparency

Prepare to share information and an explanation about how AI is used through methods appropriate to your audience.

Human Control and Oversight

Maintain human oversight to ensure that an organization's use of AI systems is transparent, explainable and trustworthy.

Promotion of Human Values

Support the principle of non-maleficence, or "do not harm," applying to both foreseeable and negative outcomes"

Ai In Assessment: Ethics, Innovation And Research. NBME

An "Altitude Control" Framework for AI in Assessment

Stakes

- Low: formative quizzes, draft notes, coaching feedback
- Medium: course grades, OSCE scores
- High: progression decisions, graduation, licensure

Role

- Assistive: drafting, summarizing, tagging, suggesting scores.
- Hybrid: AI proposes, human confirms/edits
- Automated: AI decision with spot-checking only

Role

Stakes

	Assistive	Hybrid	Automated
High **(Progression, graduation, licensure)**	**RECOMMENDED** AI for summaries, visualization, pattern highlighting **Humans make final decisions**	**CAUTION** May be appropriate with extensive oversight, transparency, and appeal mechanisms	**NOT RECOMMENDED** Too high risk for automated decisions
Medium **(Course grades, OSCE scores)**	**APPROPRIATE** AI drafts, suggests, highlights areas needing attention **Human review essential**	**USE WITH CARE** AI scoring + human verification. Especially for borderline cases	**USE WITH CARE** Only with robust validation and spot-checking
Low **(Formative quizzes, coaching)**	**EXCELLENT USE** Drafting, feedback, practice questions, coaching suggestions	**EFFICIENT** Light human oversight for quality assurance	**AI-FIRST** Great for practice, formative feedback

Responsible AI:
How to Fly Without Melting the Wax

- Use **secure, enterprise** LLM instances with strong data governance; avoid public models for identifiable data
- Maintain **clear documentation**
  - What data go into the model?
  - What tasks the model performs?
  - How outputs are checked, stored, and audited?
- Adopt or adapt existing AI governance frameworks (e.g., FUTURE-AI, reporting checklists for gen-AI in health care)
  - Development of policies for use of AI

BMJ. 2025;388:e081554

Responsible AI:
How to Fly Without Melting the Wax

- Bias & fairness checks:
  - Compare outputs across demographic groups
  - Look for patterns of inflated or deflated evaluations
  - Include diverse stakeholders in reviewing and calibrating prompts and outputs
- Transparency with learners:
  - Tell students when AI is used in assessment processes
  - Provide a pathway for students to review and challenge AI-assisted narratives
- Faculty development:
  - Train faculty not just in "how to use AI," but in how to critique AI outputs as artifacts of assessment—similar to how we teach critical appraisal of literature (more to come)
- Statistical analysis

Where We Might Be Flying Next

Narrative analytics:

- LLMs tagging themes (clinical reasoning, teamwork, professionalism) in evaluation comments to help CCCs see patterns earlier

Real-time formative feedback:

- AI assistants giving structured feedback on student notes, presentations, and simulated encounters, aligned to rubrics

Multimodal assessment:

- Combining transcripts, audio, video, and EMR interactions to create more holistic pictures of performance

Precision medical education:

- Adaptive curricula where assessment data feeds individualized learning plans, case selection, and simulation

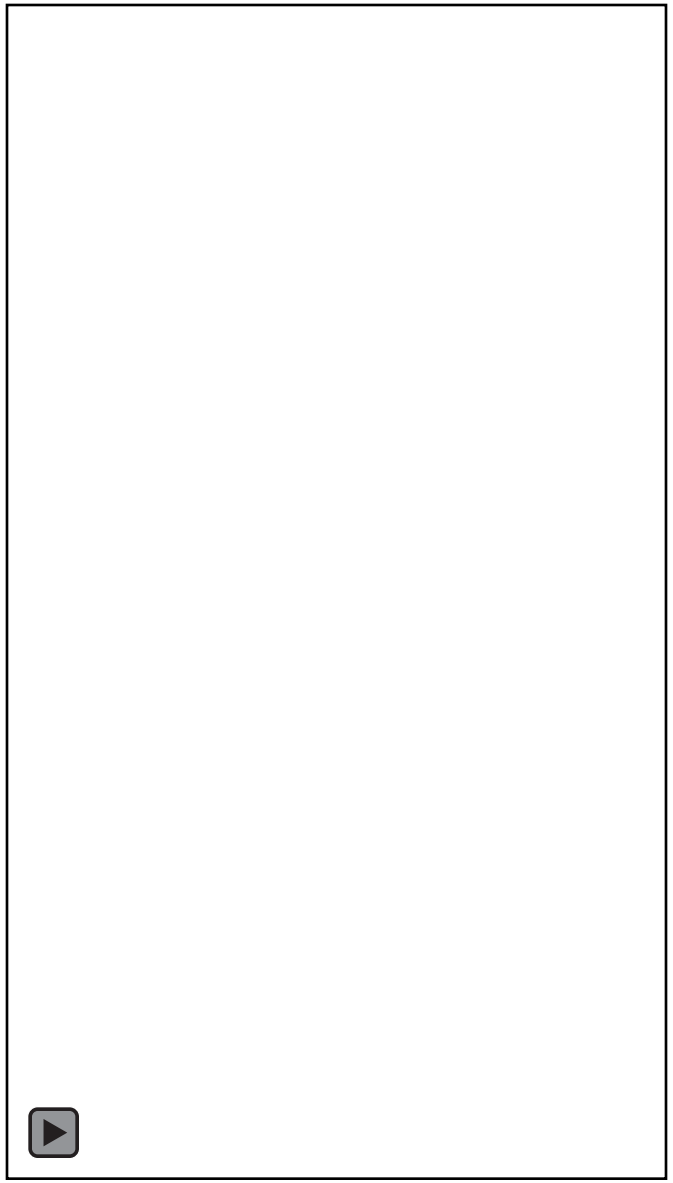
Medical Teacher. 2024;46(4):446-470; *JMIR Medical Education*. 2024;10(1):e56342. *BioMedInformatics*. 2025;5(3):37.

Escape the Labyrinth

- **AI in assessment is inevitable; equity and rigor are not.** We have to build those in on purpose.
- **Use AI where it frees humans to be more human**—to coach, to observe, to listen—not where it replaces our core professional responsibilities.
- **Treat every new AI assessment tool like a new assessment instrument**, requiring validity evidence, fairness checks, and clear use-cases.

- **Our job as educators is altitude control:**
  - Don't hug the water (fear).
  - Don't charge the sun (hype).
  - Fly a thoughtful, transparent, and evaluable flight path.

Medical Teacher. 2025;47(9):1410-1424; *BMC Med Educ*. 2025;25(1):129.



REWRITING ICARUS

Here is what they don't tell you:

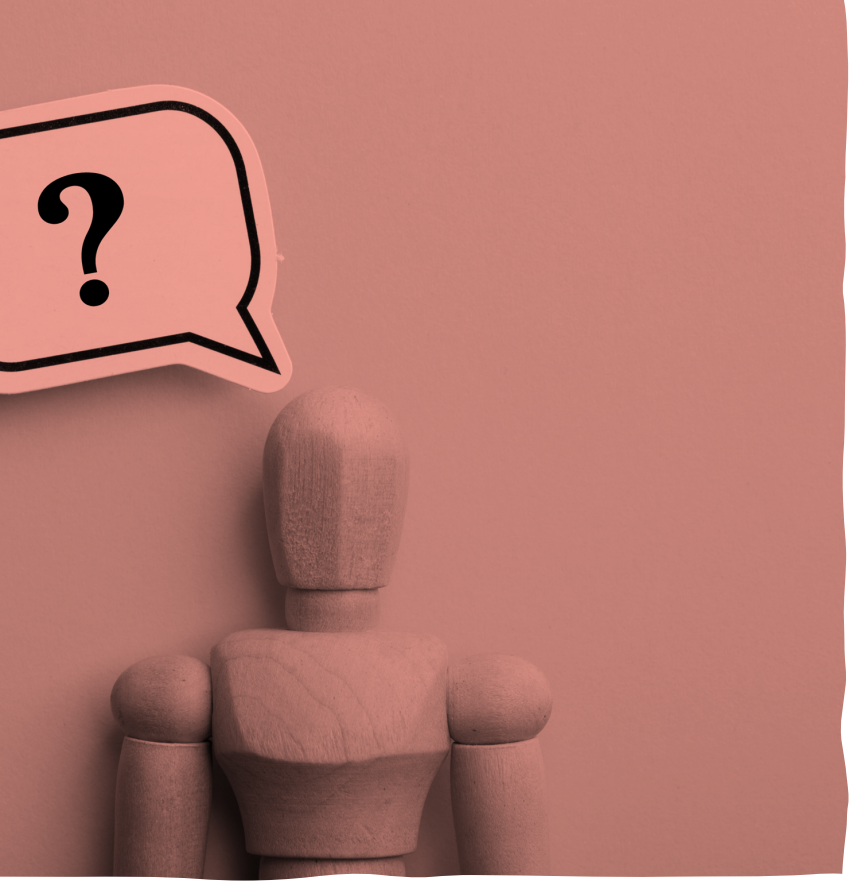
Icarus laughed as he fell.
Threw his head back and
yelled into the winds,
arms spread wide,
teeth bared to the world.

(There is a bitter triumph
in crashing when you should be
soaring.)

The wax scorched his skin,
ran blazing trails down his back,
his thighs, his ankles, his feet.
Feathers floated like prayers
past his fingers,
close enough to snatch back.
Death breathed burning kisses
against his shoulders,
where the wings joined the harness.
The sun painted everything
in shades of gold.

(There is a certain beauty
in setting the world on fire
and watching from the centre
of the flames.)

—Fiona



Q&a

matthew.zuckerman@cuanschutz.edu

Please reach out for further discussion about AI, assessment, faculty development.

AEO
OFFICE OF ASSESSMENT, EVALUATION, AND OUTCOMES

School of Medicine
UNIVERSITY OF COLORADO
ANSCHUTZ MEDICAL CAMPUS

Matt Zuckerman, MD
Sean Marshall, MA
Bonnie Kaplan, MD
Tai Lockspeiser, MD, MHPE

AI hub

Selected References

1. **Schaye V, Guzman B, Burk-Rafel J, et al.** Development and validation of a machine learning model for automated assessment of resident clinical reasoning documentation. *J Gen Intern Med.* 2022;37(9):2230-2238. doi:10.1007/s11606-022-07526-0
2. **AI and Auto-Grading in Higher Education: Capabilities, Ethics, and the Evolving Role of Educators.** ASC Office of Distance Education. Accessed December 9, 2025. https://ascode.osu.edu/news/ai-and-auto-grading-higher-education-capabilities-ethics-and-evolving-role-educators
3. **Lekadir K, Frangi AF, Porras AR, et al.** FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ.* 2025;388:e081554. doi:10.1136/bmj-2024-081554
4. **Di Mitri D, Schneider J, Specht M, Drachsler H.** Detecting mistakes in CPR training with multimodal data and neural networks. *Sensors.* 2019;19(14):3099. doi:10.3390/s19143099
5. **Yetik SS.** The impact of artificial intelligence on university teaching processes: an analysis based on faculty perspectives. *Kastamonu Education Journal.* 2025;33(3):448-468. doi:10.24106/kefdergi.1748350
6. **Rabbani SA, El-Tanani M, Sharma S, et al.** Generative artificial intelligence in healthcare: applications, implementation challenges, and future directions. *BioMedInformatics.* 2025;5(3):37. doi:10.3390/biomedinformatics5030037

7. **Janumpally R, Nanua S, Ngo A, Youens K.** Generative artificial intelligence in graduate medical education. *Front Med.* 2025;11. doi:10.3389/fmed.2024.1525604
8. **Jamieson AR, Holcomb MJ, Dalton TO, et al.** Rubrics to prompts: assessing medical student post-encounter notes with AI. *NEJM AI.* 2024;1(12):AIcs2400631. doi:10.1056/AIcs2400631
9. **Burke HB, Hoang A, Lopreiato JO, et al.** Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. *JMIR Med Educ.* 2024;10(1):e56342. doi:10.2196/56342
10. **Roveta A, Castello LM, Massarino C, Francese A, Ugo F, Maconi A.** Artificial intelligence in medical education: a narrative review on implementation, evaluation, and methodological challenges. *AI.* 2025;6(9):227. doi:10.3390/ai6090227
11. **Feigerlova E, Hani H, Hothersall-Davies E.** A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Med Educ.* 2025;25(1):129. doi:10.1186/s12909-025-06719-5
12. **Templin T, Fort S, Padmanabham P, et al.** Framework for bias evaluation in large language models in healthcare settings. *NPJ Digit Med.* 2025;8:414. doi:10.1038/s41746-025-01786-w
13. **Masters K, MacNeil H, Benjamin J, et al.** Artificial intelligence in health professions education assessment: AMEE Guide No. 178. *Med Teach.* 2025;47(9):1410-1424. doi:10.1080/0142159X.2024.2445037
14. **Gordon M, Daniel M, Ajiboye A, et al.** A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med Teach.* 2024;46(4):446-470. doi:10.1080/0142159X.2024.2314198
15. **Ark T.** Trust but Verify: Using AI in Health Professions Education. Presented at: Medical Education Day 2025; May 29, 2025. Accessed December 10, 2025. https://videos.med.wisc.edu/videos/122883