



Artificial Intelligence in Assessment and Evaluation

GEA Webinar Series: AI Skill Building for Medical Educators
December 18, 2025 (Session 8)

Share Your Story!

What have you done with AI as a result of your participation in this series?
What new tools have you tried, and for what?
Have you formed your own AI community of practice?

Share your AI growth story with us.

“I used to think... now I think...”

“I didn’t know... but now I...”

“I used to.... but now with AI, I...”



Disclosures

None

The College of Human Medicine Office of CME designates this event for a maximum of 1 *AMA PRA Category 1 Credit(s)*[™]. Physicians should claim only the credit commensurate with the extent of their participation in the activity.



MICHIGAN STATE
UNIVERSITY

College of Human Medicine
Office of Continuing Medical Education

Michigan State University - College of Human Medicine - Office of Continuing Medical Education is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to provide continuing medical education for physicians.

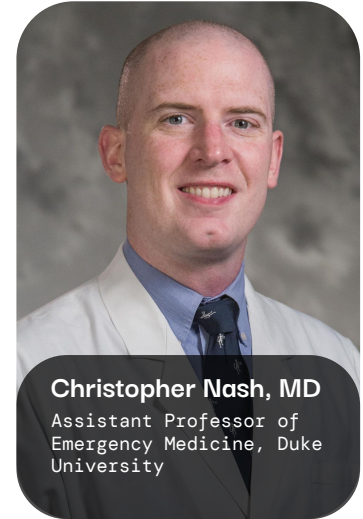
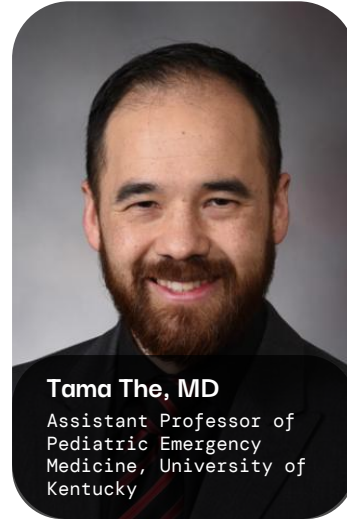
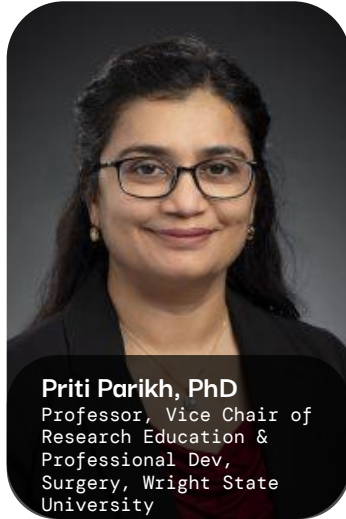
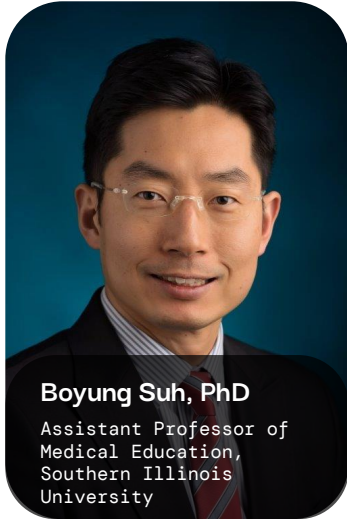
Disclaimer

Although we use certain AI tools in our sessions, we are not promoting some over others.

We are simply choosing certain tools we have access or familiarity with as examples to demonstrate skills.

We do not have time to demonstrate all the available tools.

Introductions



Learning Objectives

01

Introduction

Understand the rationale and opportunities for AI-assessment and evaluation

02

Demonstration

Construct a reliable assessment tool using an application programming interface (API)

03

Reflection

Reflect on considerations for sound and reliable AI-assessment approaches

GEA

Group on Educational Affairs



Why AI Matters *Now* for Assessment and Evaluation? The Opportunities and Limitations of AI in Assessment

Part 1:

The rationale and opportunities for AI-assessment and evaluation

Scalable & Sustainable Competency-Based Assessments

Fast and Scalable

Generate MCQs, OSCE cases, rubrics quickly

CBME and EPA Assessment Volume

More frequent assessments than faculty can manage manually

> [BMC Med Educ. 2025 May 1;25\(1\):641. doi: 10.1186/s12909-025-07241-4.](#)

Is AI the future of evaluation in medical education?? AI vs. human evaluation in objective structured clinical examination

[Murat Tekin](#)¹, [Mustafa Onur Yurdal](#)¹, [Çetin Toraman](#)¹, [Güneş Korkmaz](#)^{2,3}, [İbrahim Uysal](#)⁴

▶ [Med Educ Online. 2025 Jul 17;30\(1\):2532682. doi: 10.1080/10872981.2025.2532682](#) 

Ten tips to harnessing generative AI for high-quality MCQS in medical education assessment

[Comparative Study](#) > [Acad Med. 2024 May 1;99\(5\):508-512.](#)

doi: 10.1097/ACM.0000000000005626. Epub 2023 Dec 28.

[Mohi Eldin Magzoub](#)^{a,✉}, [Imran Zafar](#)^{a,b}, [Fadi Munshi](#)^{b,c}, [Fouzia Shersad](#)^{b,c}

Large Language Models in Medical Education: Comparing ChatGPT- to Human-Generated Exam Questions

[Matthias Carl Laupichler](#), [Johanna Flora Rother](#), [Ilona C Grunwald Kadow](#), [Seifollah Ahmadi](#),
[Tobias Raupach](#)

Personalized Learning



Real-time Feedback and Performance Analysis

Analyze student performance in real-time, providing instant, data-driven feedback



Personalized Learning

Tailored guidance to support competency growth

[Home](#) | [JAMA Network Open](#) | [Vol. 5, No. 2](#)

Original Investigation | Medical Education

Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial

Ali M. Fazlollahi, MSc^{1,2}; Mohamad Bakhaidar, MD, MSc^{1,2,3}; Ahmad Alsayegh, MD^{1,2,3}; et al

Rubrics to Prompts: Assessing Medical Student Post-Encounter Notes with AI

Authors: Andrew R. Jamieson, Ph.D. , Michael J. Holcomb, M.S. , Thomas O. Dalton, M.D. , Krystle K. Campbell, D.H.A. , Sol Vedovato, M.S. , Ameer Hamza Shakur, Ph.D. , Shinyoung Kang, B.S. , David Hein, M.S. , Jack Lawson, B.S. , Gaudenz Danuser, Ph.D. , and Daniel J. Scott, M.D. [Author Info & Affiliations](#)

Published November 25, 2024 | NEJM AI 2024;1(12) | DOI: 10.1056/AIcs2400631 | [VOL. 1 NO. 12](#)

AI-Standardized Clinical Examination Training on OSCE Performance

Authors: Emma Lavigne, M.D. , Antonio Lopez, M.D. , Julien Frandon, M.D., Ph.D. , Geoffrey Blaizot, B.Sc. , Ludovic Gabellier, M.D., Ph.D. , Salma Adham, M.D. , José Ursic Bedoya, M.D., Ph.D. , and Kévin Yaou, M.D., Ph.D. [Author Info & Affiliations](#)

Published July 14, 2025 | NEJM AI 2025;2(8) | DOI: 10.1056/Aloa2500066 | [VOL. 2 NO. 8](#) | [Copyright © 2025](#)

Consistency, Equity & Program-Level Insight



Consistency and Data Insights

Standardize rubrics, Analyze trends from performance and feedback



Optimal Use of Faculty Time

Focus on teaching, mentoring, interpretation, coaching, and research

[nature](#) > [scientific reports](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 03 July 2025

Explainable artificial intelligence for predicting medical students' performance in comprehensive assessments

[Haniye Mastour](#), [Toktam Dehghani](#) , [Ehsan Moradi](#) & [Saeid Eslami](#)

> [BMC Med Educ.](#) 2025 Mar 4;25(1):333. doi: 10.1186/s12909-025-06862-z.

Randomized Controlled Trial > [JAMA Surg.](#) 2025 Sep 1;160(9):993-1003.

doi: 10.1001/jamasurg.2025.2564.

Artificial Intelligence-Augmented Human Instruction and Surgical Simulation Performance: A Randomized Clinical Trial

[Bianca Giglio](#)¹, [Abdulmajeed Albeloushi](#)^{1 2}, [Ahmad Kh Alhaj](#)^{1 2}, [Mohamed Alhantoobi](#)^{1 3}, [Rothaina Saeedi](#)^{1 2}, [Vanja Davidovic](#)¹, [Abicumaran Uthamacumaran](#)¹, [Recai Yilmaz](#)^{1 4}, [Jason Lapointe](#)^{1 5}, [Neevya Balasubramaniam](#)^{1 6}, [Trisha Tee](#)^{1 7}, [Ali M Fazlollahi](#)^{1 6}, [José A Correa](#)⁸, [Rolando F Del Maestro](#)¹



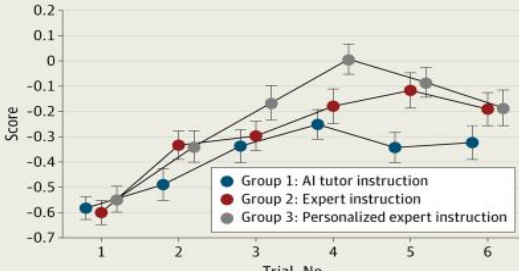

GPT-4 generated answer rationales to multiple choice assessment questions in undergraduate medical education

[Peter Y Ch'en](#)¹, [Wesley Day](#)¹, [Ryan C Pekson](#)¹, [Juan Barrientos](#)¹, [William B Burton](#)¹, [Allison B Ludwig](#)¹, [Sunit P Jariwala](#)¹, [Todd Cassese](#)^{2 3}

Could AI Replace Human Educators?

JAMA Surgery

RCT: Artificial Intelligence–Augmented Human Instruction and Surgical Simulation Performance

<p>POPULATION 40 Men, 46 Women, 1 Unspecified</p>  <p>Medical students enrolled in their preparatory, first, or second year at a Quebec institution Mean (SD) age, 22.7 (4.0) y</p>	<p>INTERVENTION 88 Participants randomized 87 Participants analyzed</p>  <p>30 Artificial intelligence (AI) tutor instruction group Feedback delivered by an AI tutoring system</p> <p>29 Expert instruction group Feedback delivered by an expert instructor using the exact wording provided by AI tutor</p> <p>28 Personalized expert instruction group Personalized feedback delivered by an expert instructor without restriction to AI tutor wording</p>	<p>FINDINGS</p> <p>Personalized expert instruction group achieved significantly higher scores across several practice tasks and the realistic task than AI tutor instruction and expert instruction groups</p>  <table border="1"> <caption>Mean (95% CI) composite expertise score during realistic task</caption> <thead> <tr> <th>Group</th> <th>Score</th> <th>95% CI</th> </tr> </thead> <tbody> <tr> <td>AI tutor instruction</td> <td>-0.35</td> <td>-0.45 to -0.24</td> </tr> <tr> <td>Expert instruction</td> <td>-0.32</td> <td>-0.45 to -0.24</td> </tr> <tr> <td>Personalized expert instruction</td> <td>-0.14</td> <td>-0.25 to -0.04</td> </tr> </tbody> </table>	Group	Score	95% CI	AI tutor instruction	-0.35	-0.45 to -0.24	Expert instruction	-0.32	-0.45 to -0.24	Personalized expert instruction	-0.14	-0.25 to -0.04
Group	Score	95% CI												
AI tutor instruction	-0.35	-0.45 to -0.24												
Expert instruction	-0.32	-0.45 to -0.24												
Personalized expert instruction	-0.14	-0.25 to -0.04												
<p>SETTINGS / LOCATIONS</p>  <p>1 Simulation center in Quebec</p>	<p>PRIMARY OUTCOME</p> <p>Trainee technical skill acquisition and transfer during simulated subpial resection tasks on the NeuroVR. Surgical performance assessed and calculated by AI tutoring system on a scale of -1.00 (novice) to 1.00 (expert)</p>													

AI for Individualized Assessment and Feedback

Opportunities

- MCQ generation
- Academic performance prediction
- Tailored assessment
- 1:1 Coaching
- OSCE scoring
- Faculty teaching evaluation

Platforms

- Clinical reasoning & exam prep (AI Patient Actor, DDx, AMBOSS, Osmosis, Medical Student AI)
- Simulations (Lumeto, MedSimAI, Neural Consult, Labster)
- Assessment management (Gradescope)

Limitations

- Requires faculty oversight
- Lack of rigorous psychometric validation of tools
- Lack of standardization in feedback provision and assessment
- AI's black box nature and biases

Research on AI-Assessment in Medical Education

01	Clinical Knowledge & Competence	<ul style="list-style-type: none">• Clinical diagnostic reasoning• OSCE, OSPE, simulations (low- and high-fidelity, virtual)• Patient notes, medical counseling, physical examination
02	Communication Skills & Professionalism	<ul style="list-style-type: none">• Oral case presentation• Team and inter-professional communication performance• Student and faculty professionalism skills and lapses
03	Feedback Provision & Feedback Quality	<ul style="list-style-type: none">• Gamified anatomy learning for individualized feedback• Automated, real-time feedback on radiology education• Quality of faculty feedback and evaluation methods
04	Academic Performance Prediction & Remediation	<ul style="list-style-type: none">• Prediction of preclerkship performance, exam performance, student success, and course performance• Remediation risk modeling based on narrative comments
05	Curricular Effectiveness	<ul style="list-style-type: none">• Basic science curriculum assessment and development based on student perceptions• International medical e-learning course evaluation

GEA

Group on Educational Affairs



Generative AI to Assess OSCE Notes

The API: How to use Generative AI Reliably

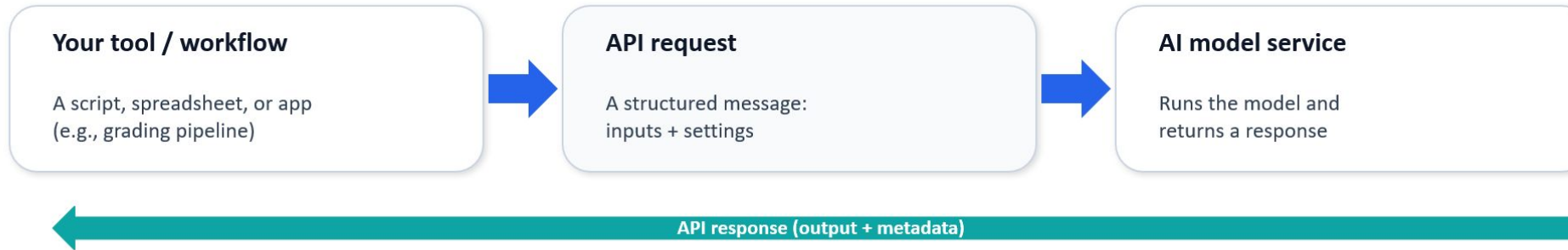
Live Demonstration

Part 2:

Reliable AI-assessment tool construction using an API

What is an API?

An Application Programming Interface lets one piece of software request a service from another — in a consistent, structured way.





NBME

REASONS THE FUTURE

TO MEET THIS MEDICINE

PRODUCE NUTRITION



How much does it cost to hand-grade the OSCEs at UK?

UK has 160 medical students per year	160
Each student performs 9 OSCEs in M3	X 9
Each OSCE is graded twice	X 2
Assume it takes 9 minutes to grade an OSCE	X 9 min
<hr/>	
	25,920 min (432 hrs)

Assuming each hour of faculty time is paid at \$125/hr,
it costs UK **~\$54,000/yr** to grade M3 OSCEs

Comparative Analysis of Large Language Model and Human Rater Scoring of Student Documentation in Objective Structured Clinical Examinations

2024-2026 SEEF Fellows:

Christopher Nash, MD, EdM

Candace Pau, MD

Nayef Chahin, MD

Tama Thé, MD



```
41
42 def grade_section_with_key(rubric_content, answer_key_content, section_content, section_name, temperature, top_p, log_file):
43     # Grades a section using both the rubric and answer key for that section, and logs the conversation.
44     rubric_text = rubric_content.get(section_name.lower(), "No rubric available for this section.")
45     answer_key_text = answer_key_content.get(section_name.lower(), "")
46
47     # Prepare the initial prompt to grade the section
48     messages = [
49         {"role": "system", "content": f"I am a medical educator and I would like your help grading an assignment. "
50                                         f"My students recently completed an activity where they interviewed a patient about their symptoms. "
51                                         f"I have made a scoring rubric that includes the information that should be reported in their post-interview "
52                                         f"The rubric is broken into individual steps. To help you out, I just want you to score each individual section "
53                                         f"Please score each individual section that I provide you based off the rubric and the answer key. Please provide a score and "
54                                         f"Please place their final score after your explanation all by itself as an integer with no markup"},
55         {"role": "user", "content": f"Refer to the rubric: {rubric_text}. "
56                                     f"Here is the answer key for {section_name}: {answer_key_text}. "
57                                     f"Please evaluate the following {section_name} and provide a score: "
58                                     f"{section_content}"}
59     ]
60
61     # Send the initial request to ChatGPT
62     response = client.chat.completions.create(
63         model="gpt-4o-mini",
64         messages=messages,
65         temperature=temperature,
66         top_p=top_p
67     )
68
69     # Get the response content (the full explanation)
70     score_text = response.choices[0].message.content.strip()
71
72     # Log the conversation for tracking
73     log_interaction(log_file, messages, score_text)
```

christopherjnash/OSCE-Grader x +

← ↻ 🏠 🔒 https://github.com/christopherjnash/OSCE-Grader

OSCE-Grader Public

📌 Pin 📄 Unwatch 1 🍴 Fork 0 ⭐ Star 0

👤 main 1 Branch 0 Tags 🔍 Go to file 📄 Add file <> Code

christopherjnash Update README.md 9a86f75 · 6 hours ago 41 Commits

docs	Create modifying_prompt.md	6 hours ago
examples	Rename flankpainnotes.xlsx to sample_student_notes.xlsx	6 hours ago
scripts	Create convert_rubric.py	6 hours ago
.gitignore	Create .gitignore	6 hours ago
LICENSE	Update LICENSE	yesterday
README.md	Update README.md	6 hours ago
requirements.txt	Create requirements.txt	6 hours ago

📖 README 📄 MIT license

OSCE Grader Setup Guide 🚀

An AI-powered grading system for medical student OSCE post-encounter notes, using GPT models to automate grading and provide structured feedback.

Features

- ✔ Supports Excel, CSV, and text-based student notes
- ✔ Works with structured rubrics (Excel, CSV)
- ✔ Uses ChatGPT to generate detailed, section-by-section grading
- ✔ Customizable grading prompts via config.py
- ✔ Automatic logging for debugging and analysis
- ✔ Easy setup with Python & OpenAI API
- ✔ Includes convert_rubric.py to assist in converting rubric files (PDF/DOCX to structured formats)

About

No description, website, or topics provided.

- 📖 Readme
- 📄 MIT license
- 📈 Activity
- ⭐ 0 stars
- 👁 1 watching
- 🍴 0 forks

Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Languages

- Python 100.0%

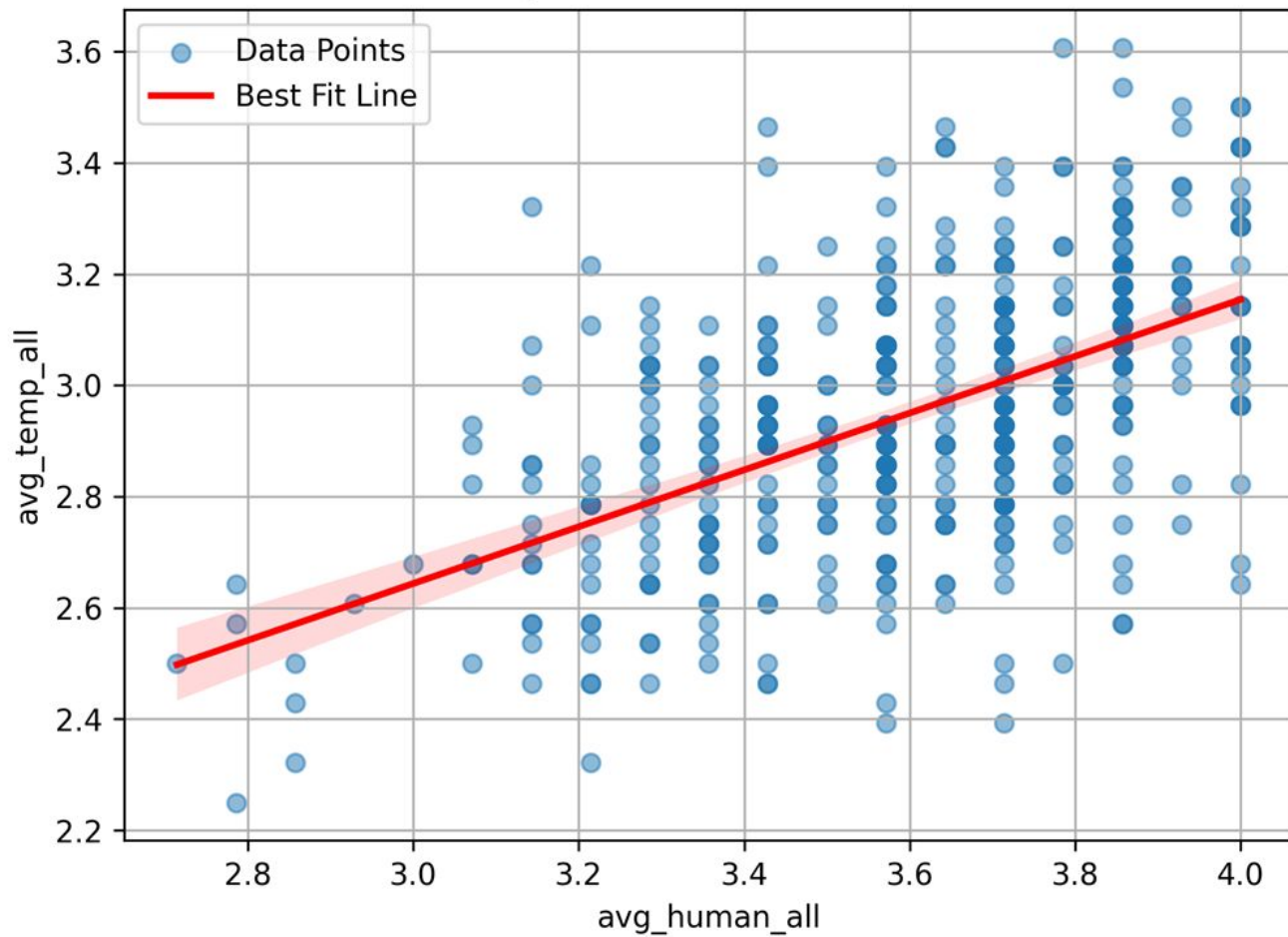
Suggested workflows

Based on your tech stack

- SLSA Generic generator** Configure
Generate SLSA3 provenance for your existing release workflows
- Python package** Configure
Create and test a Python package on

Human vs GPT Expl

Spearman Coeff: 0.526



OVERALL RESULTS



	Average Rating	ICC
Human graders	3.60	0.88
ChatGPT with rationale	2.95	0.76
ChatGPT w/o rationale	2.48	0.94

	ICC	Mean score difference
Human grader x ChatGPT with rationale	0.21	-0.65 <i>(95% CI: 0.61-0.69, p<0.0001)</i>
Human grader x ChatGPT w/o rationale	0.05	-1.12 <i>(95% CI: 1.08-1.17, p<0.0001)</i>
ChatGPT with rationale x ChatGPT w/o rationale	0.29	-0.47 <i>(95% CI: 0.43-0.52, p<0.0001)</i>

Human grader x ChatGPT with rationale: Spearman's $\rho = 0.53$

What's on your mind today?

Ask anything

+ 🕒 Thinking ▾

🎤 🔊

Gemini

PRO



🌟 Hi Tama

Fast is now powered by 3 Flash. Try it [→](#)

Ask Gemini

+ 🛠️ Tools

Fast ▾



Create image

Create video

Write anything

Help me learn

Boost my day

The “Chat Assistant”

Easy, conversational, intuitive.

The path of rapid adoption.

Best Use

Low-stakes ideation, brainstorming, drafting.



Gemini

The “Assessment Engine”

Programmatic, structured, auditable.

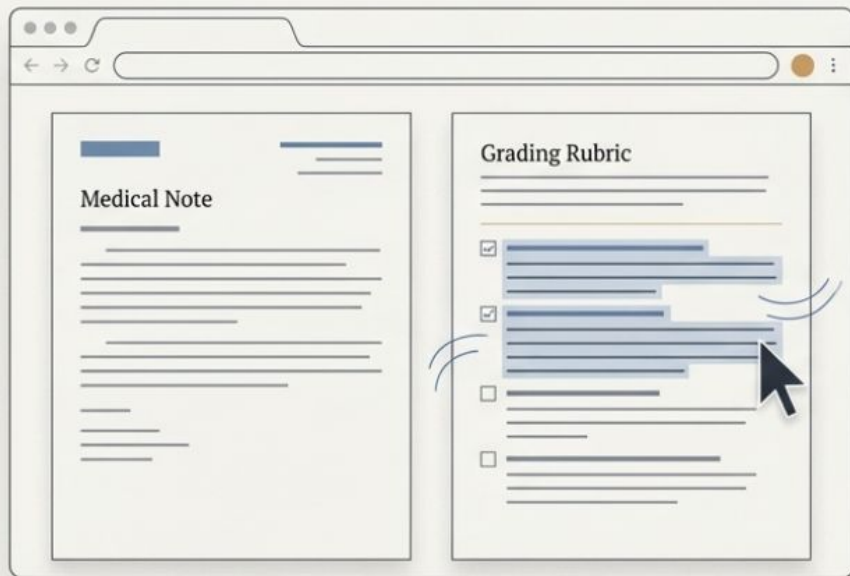
The path of deliberate engineering.

Required Use

High-stakes decisions, grading, validation.

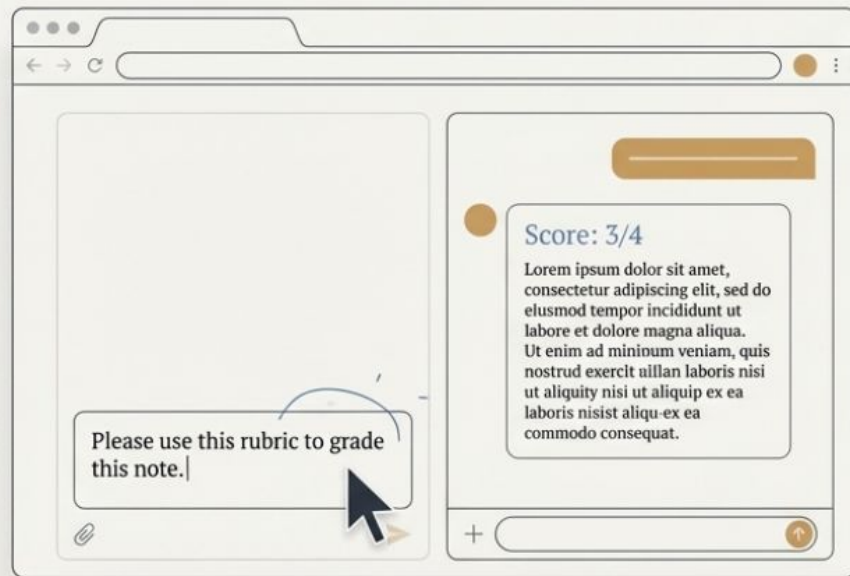


The Seductive Promise of Instant Grading



1

2



3

4

How should I use GenAI in Medical student assessment?

Show thinking ^

Considering GenAI's Role

I'm now focusing on how GenAI can provide formative feedback in medical student assessments. I'm exploring methods to reduce bias and ensure fairness, especially in simulated patient interactions. The ethical implications are top of mind, and I'm investigating how we can align GenAI's output with pedagogical goals to maximize its impact on learning.

Beginning Research Phase

I've transitioned from conceptual exploration to a rigorous literature review. My search strategy is now honed, focusing on practical applications of GenAI within medical student assessment. This includes automated scoring, AI-driven OSCE simulations, and generating effective distractors for MCQs. The NBME's role necessitates a focus on ethical guidelines and high-quality assessments. My next step involves synthesizing findings to identify gaps and prioritize areas for further investigation.

Reviewing Current Landscape

GenAI in Medical Student Assessment ▾

Reviewing Current Landscape

I'm currently focused on the automated scoring of clinical notes, a key area in GenAI for medical student assessment. Research from sources like MDPI and Stanford highlights the promising capabilities of LLMs, such as GPT-4o, in this domain. This research suggests that advanced LLMs can provide a valuable framework to improve the assessment process.

Examining Assessment Methods

I'm now diving into the application of GenAI within the OSCE framework. The potential is substantial; virtual patient actors offer continuous practice and immediate feedback, showing real improvement in diagnostic skills and empathy. Interestingly, I'm also finding that AI can help create OSCEs, generating scripts and checklists. My focus is also on validity and reliability, the crucial role of human oversight in avoiding hallucinations in complex clinical scenarios. I'm seeing that the "LLM-as-a-judge" approach is promising, but will absolutely require human review.

Defining Scope and Framework

I'm now concretizing my thoughts into several key areas: AI-enhanced OSCEs for simulation and scoring, content generation of MCQs with distractor analysis, the creation of personalized learning pathways with targeted feedback, and essential implementation safeguards. My immediate goal is to establish the ideal workflow and pinpoint how models can aid medical education, while addressing issues of validity, bias, and reliability, including things like Intraclass Correlation Coefficient and BERTScore for scoring. Next, I will develop a specific research framework.

Generative AI (GenAI) can be integrated into medical student assessment across formative and

The Crisis of Reliability: Science vs. Stochasticism

CHAT INTERFACE



The Problem: Stochastic Crisis. High default ``temperature`` introduces random noise.

The Result: The same note can receive different grades on subsequent attempts. This is measurement error.

The Killer Flaw: Model Drift. The model you use today is not the same as last month. A landmark study showed GPT-4 accuracy on a task dropping from **97.6% to 2.4%** in three months.

API



The Solution: Deterministic Control. We force ``temperature=0`` in the code.

The Result: The AI transforms from a creative partner into a predictable classification engine.

The Professional Standard: Model Pinning. We lock a specific version (e.g., ``gpt-4-0613``) for the entire academic year, ensuring every student is assessed against the exact same algorithm.

```
41
42 def grade_section_with_key(rubric_content, answer_key_content, section_content, section_name, temperature, top_p, log_file):
43     # Grades a section using both the rubric and answer key for that section, and logs the conversation.
44     rubric_text = rubric_content.get(section_name.lower(), "No rubric available for this section.")
45     answer_key_text = answer_key_content.get(section_name.lower(), "")
46
47     # Prepare the initial prompt to grade the section
48     messages = [
49         {"role": "system", "content": f"I am a medical educator and I would like your help grading an assignment. "
50                                         f"My students recently completed an activity where they interviewed a patient about their symptoms. "
51                                         f"I have made a scoring rubric that includes the information that should be reported in their post-interview "
52                                         f"The rubric is broken into individual steps. To help you out, I just want you to score each individual section "
53                                         f"Please score each individual section that I provide you based off the rubric and the answer key. Please provide a score and "
54                                         f"Please place their final score after your explanation all by itself as an integer with no markup"},
55         {"role": "user", "content": f"Refer to the rubric: {rubric_text}. "
56                                     f"Here is the answer key for {section_name}: {answer_key_text}. "
57                                     f"Please evaluate the following {section_name} and provide a score: "
58                                     f"{section_content}"}
59     ]
60
61     # Send the initial request to ChatGPT
62     response = client.chat.completions.create(
63         model="gpt-4o-mini",
64         messages=messages,
65         temperature=temperature,
66         top_p=top_p
67     )
68
69     # Get the response content (the full explanation)
70     score_text = response.choices[0].message.content.strip()
71
72     # Log the conversation for tracking
73     log_interaction(log_file, messages, score_text)
```

```
41
42 def grade_section_with_key(rubric_content, answer_key_content, section_content, section_name, temperature, top_p, log_file):
43     # Grades a section using both the rubric and answer key for that section, and logs the conversation.
44     rubric_text = rubric_content.get(section_name.lower(), "No rubric available for this section.")
45     answer_key_text = answer_key_content.get(section_name.lower(), "")
46
47     # Prepare the initial prompt to grade the section
48     messages = [
49         {"role": "system", "content": f"I am a medical educator and I would like your help grading an assignment. "
50                                         f"My students recently completed an activity where they interviewed a patient about their symptoms. "
51                                         f"I have made a scoring rubric that includes the information that should be reported in their post-interview "
52                                         f"The rubric is broken into individual steps. To help you out, I just want you to score each individual secti "
53                                         f>Please score each individual section that I provide you based off the rubric and the answer key. Please thi "
54                                         f>Please place their final score after your explanation all by itself as an integer with no markup"},
55         {"role": "user", "content": f"Refer to the rubric: {rubric_text}. "
56                                     f"Here is the answer key for {section_name}: {answer_key_text}. "
57                                     f>Please evaluate the following {section_name} and provide a score: "
58                                     f"{section_content}"}
59     ]
60
61     # Send the initial request to ChatGPT
62     response = client.chat.completions.create(
63         model="gpt-4o-mini",
64         messages=messages,
65         temperature=temperature,
66         top_p=top_p
67     )
68
69     # Get the response content (the full explanation)
70     score_text = response.choices[0].message.content.strip()
71
72     # Log the conversation for tracking
73     log_interaction(log_file, messages, score_text)
```

The Crisis of Security: Data Sovereignty vs. Data Risk

CHAT INTERFACE



The Problem: **Data Sovereignty Risk.** Consumer terms of service allow vendors to train on your inputs.

The Result: Student IP and de-identified patient data enter a public feedback loop.

The Killer Flaw: **No BAA.** Using standard chat tools for any PHI is a **HIPAA violation.**

API



The Solution: **Zero Data Retention (ZDR).** Enterprise API terms legally bind the vendor to process and discard data, not train on it.

The Result: The institution maintains full control over its data.

The Professional Standard: **HIPAA Compliance.** Enterprise APIs from providers like Azure and OpenAI offer comprehensive **Business Associate Agreements (BAAs).**

Search CTRL K

GET STARTED

- Overview
- Quickstart
- Models
- Pricing
- Changelog
- Terms and policies

CAPABILITIES

- Text generation
- Vision
- Image generation
- Audio generation

Cookbook

Forum

Help

Key concepts

Copy page

Key concepts to understand when working with the OpenAI API.

At OpenAI, protecting user data is fundamental to our mission. We do not train our models on inputs and outputs through our API. Learn more on our [API data privacy page](#).

Text generation models

OpenAI's text generation models (often referred to as generative pre-trained transformers or "GPT" models for short), like GPT-4 and GPT-3.5, have been trained to understand natural and formal language. Models like GPT-4 allows text outputs in response to their inputs. The inputs to these models are also referred to as "prompts". Designing a prompt is essentially how you "program" a model like GPT-4, usually by providing instructions or some examples of how to successfully complete a task. Models like GPT-4 can be used across a great variety of tasks including content or code generation, summarization, conversation, creative writing, and more. Read more in our introductory [text generation guide](#) and in our [prompt engineering guide](#).

Assistants

Assistants refer to entities, which in the case of the OpenAI API are powered by large language models like GPT-4, that are capable of performing tasks for users. These assistants operate based on the instructions embedded within the context window of the model. They also usually have access to tools which allows the assistants to perform more complex tasks like running code or retrieving information from a file. Read more about assistants in our [Assistants API Overview](#).

The Crisis of Accountability: Proof vs. A Black Box

CHAT INTERFACE



The Problem: No Audit Trail. The only record is a user's personal chat history.

The Result: In a grade appeal, the defense is, "I typed it in and this is what it said." There is no verifiable proof.

API



The Solution: The ``system_fingerprint``. Every API call returns a unique ID representing the exact backend configuration state.

The Result: We can produce an immutable log: *Who* made the call, *What* was the prompt, *When* it happened, *Which* model was used, and *The Result*. This is **forensic-level evidence**.

Case ID: Nausea and Vomiting (M. Neal)

History & Physical Exam (Please list required documentation for each section)

History Documentation	Physical Exam Documentation
Acute on chronic onset nausea/vomiting <ul style="list-style-type: none"> Current episode started two days ago Worsening or intractable Abdominal bloating Avoiding large meals Sick exposure via grandchild who goes to daycare History of diabetes OR insulin-dependent diabetes No fever	Tachycardia Minimal pain with palpation diffusely No rebound tenderness

4	3	2	1
History Documentation			
Almost all key information is present, concise and well organized with little irrelevant information	Most key positive historical findings are documented and well organized, missing pertinent negatives	Many key historical findings are omitted. History is disorganized or poorly documented (obscuring what is key)	Key history findings are inaccurately documented or omitted
Physical Exam Documentation			
Almost all key physical exam findings are present, concise and well organized with little irrelevant information	Most key positive physical exam findings are documented and well organized, but missing pertinent negatives	Many key physical exam findings are omitted. Exam is disorganized or extensively detailed (obscuring what is key)	Key physical exam findings are inaccurately documented or omitted
Organization			
All features are organized, with information presented in appropriate sections.	Most features are organized with information presented in appropriate sections.	Some features are disorganized with information in the incorrect place or repeated and/or occasional sections missing.	Most features are disorganized, with appropriate sections missing.

Medical Decision Making

Assessment/Summary Statement

Essential Elements	
1.	Acute on chronic OR recurrent
2.	Vomiting OR nonbilious emesis OR non-bloody emesis
3.	History of diabetes
4.	Sick exposure
Optional	Bloating OR afebrile OR no diarrhea

Assessment/Summary Statement			
4	3	2	1
All essential elements present, and statement has 5 or fewer total elements	Missing 1 essential element and/or statement has more than 5 total elements	Missing multiple essential elements	No summary statement included

Differential Diagnosis & Support

Rank	Diagnosis	Supporting Evidence
#1	Gastroparesis	Prior episodes of nausea vomiting and bloating Prior episodes of both after meals History of diabetes on insulin Limiting patient from eating big meals
#2	Viral Gastroenteritis OR Viral Gastritis	Acute onset vomiting Worsening abdominal bloating Exposure to grandchild who goes to daycare No fever No diarrhea
Optional #3	Gastric outlet obstruction	Intractable nausea and vomiting Progressive worsening symptoms Abdominal bloating

4	3	2	1
Differential Diagnosis			
Pertinent, prioritized differential diagnosis reflecting correct interpretation of findings	Differential diagnosis generally consistent with findings, with minor flaws in prioritization that would not negatively impact patient care	Differential diagnosis somewhat consistent with findings, but overly broad and/or inaccurately prioritized	Differential diagnosis is inconsistent with findings
Differential Diagnosis Support			
Excellent differential diagnosis well supported, links to diagnoses are correct and complete	Appropriately supported differential diagnosis, with missing or incorrect links that would not negatively impact patient care	Weakly supported differential diagnosis, with several incorrect links between findings and diagnosis	Differential diagnosis lacks accurate links to patient findings

Initial Diagnostic Plan

None

Acceptable: Abdominal x-ray/KUB, CT abdomen/pelvis, BMP, A1c

Unacceptable: CBC, endoscopy, stool studies

4	3	2	1
Initial Diagnostic Plan			
Plan for diagnostic workup is effective and efficient, includes all essential tests, and few or no unnecessary tests	Reasonable plan for diagnostic workup, includes most essential but also some unnecessary tests	Ineffective plan for diagnostic workup, essential tests missed and/or irrelevant tests included ("shotgunning")	Diagnostic workup is inconsistent with differential diagnoses and/or places patient at unnecessary risk

grader_int_only.py

grader_with_explanations.py

script.bat

Rubric with Common Features Replicated in Excel

standardrubric.xlsx

each section)

Exam Documentation

ia

ain with palpation diffusely

nd tenderness

Avoiding target meals

Sick exposure via grandchild who goes to daycare

History of diabetes OR insulin-dependent diabetes

No fever

4	3	2	1
History Documentation			
Almost all key information is present, concise and well organized with little irrelevant information	Most key positive historical findings documented and well organized with pertinent negatives	Many key historical findings are omitted. History is disorganized or poorly documented (obscuring what is key)	Key history findings are inaccurately documented or omitted
Physical Exam Documentation			
Almost all key physical exam findings are present, concise and well organized with little irrelevant information	Most key positive physical exam findings documented and well organized with missing pertinent negatives	Many key physical exam findings are omitted. Exam is disorganized or extremely detailed (obscuring what is key)	Key physical exam findings are inaccurately documented or omitted
Organization			
All features are organized, with information presented in appropriate sections.	Most features are organized with information presented in appropriate sections.	Some features are disorganized with information in the incorrect place or repeated and/or occasional sections missing.	Most features are disorganized, with appropriate sections missing.

Medical Decision Making

Assessment/Summary Statement

Essential Elements	
1.	Acute on chronic OR recurrent
2.	Vomiting OR nonbilious emesis OR non-bloody emesis
3.	History of diabetes
4.	Sick exposure
Optional	Bloating OR afebrile OR no diarrhea

Assessment/Summary Statement			
4	3	2	1
All essential elements present, and statement has 5 or fewer total elements	Missing 1 essential element and/or statement has more than 5 total elements	Missing multiple essential elements	No summary statement included

Differential Diagnosis & Support

Rank	Diagnosis	Supporting Evidence
#1	Gastroparesis	Prior episodes of nausea vomiting and bloating Prior episodes of both after meals History of diabetes on insulin Limiting patient from eating big meals
#2	Viral Gastroenteritis OR Viral Gastritis	Acute onset vomiting Worsening abdominal bloating Exposure to grandchild who goes to daycare No fever No diarrhea
Optional #3	Gastric outlet obstruction	Intractable nausea and vomiting Progressive worsening symptoms Abdominal bloating

4	3	2	1
Differential Diagnosis			
Pertinent, prioritized differential diagnosis reflecting correct interpretation of findings	Differential diagnosis generally consistent with findings, with minor flaws in prioritization that would not negatively impact patient care	Differential diagnosis somewhat consistent with findings, but overly broad or inaccurately prioritized	Differential diagnosis is inconsistent with findings
Differential Diagnosis Support			
Excellent differential diagnosis well supported, links to diagnoses are correct and complete	Appropriately supported differential diagnosis with missing or incorrect links that would not negatively impact patient care	Weakly supported differential diagnosis, with several incorrect links between findings and diagnosis	Differential diagnosis lacks accurate links to patient findings

Initial Diagnostic Plan

None

Acceptable: Abdominal x-ray/KUB, CT abdomen/pelvis, BMP, A1c

Unacceptable: CBC, endoscopy, stool studies

4	3	2	1
Initial Diagnostic Plan			
Plan for diagnostic workup is effective and efficient, includes all essential tests, and few or no unnecessary tests	Reasonable plan for diagnostic workup, includes most essential but also some unnecessary tests	Ineffective plan for diagnostic workup, essential tests missed and/or irrelevant tests included ("shotgunning")	Diagnostic workup is inconsistent with differential diagnoses and/or places patient at unnecessary risk

Rubric with Case-Specific Features Created in Excel

Case ID: Nausea and Vomiting (M. Neal)

History & Physical Exam (Please list required documentation for each section)

History Documentation	Physical Exam Documentation
Acute on chronic onset nausea/vomiting <ul style="list-style-type: none"> Current episode started two days ago Worsening or intractable Abdominal bloating Avoiding large meals	Tachycardia Minimal pain with palpation diffusely No rebound tenderness

- kneepainnotes.xlsx
- nasalcongestionkey.xlsx
- nasalcongestionnotes.xlsx
- nauseavomitingkey.xlsx
- nauseavomitingnotes.xlsx
- syncopekey.xlsx

2	1
Many key historical findings are omitted. History is disorganized or poorly documented (obscuring what is key)	Key history findings are inaccurately documented or omitted
Many key physical exam findings are omitted. Exam is disorganized or extensively detailed (obscuring what is key)	Key physical exam findings are inaccurately documented or omitted
Some features are disorganized with information in the incorrect place or repeated and/or occasional sections missing.	Most features are disorganized, with appropriate sections missing.

Medical Decision Making

Assessment/Summary Statement

Essential Elements	
1.	Acute on chronic OR recurrent
2.	Vomiting OR nonbilious emesis OR non-bloodly emesis
3.	History of diabetes
4.	Sick exposure
Optional	Bloating OR afebrile OR no diarrhea

4	3	2	1
All essential elements present, and statement has 2 or fewer total elements	Missing 1 essential element and/or statement has more than 5 total elements	Missing multiple essential elements	No summary statement included

Differential Diagnosis & Support







Rank	Diagnosis	Supporting Evidence
#1	Gastroparesis	Prior episodes of nausea vomiting and bloating Prior episodes of both after meals History of diabetes on insulin Limiting patient from eating big meals
#2	Viral Gastroenteritis OR Viral Gastritis	Acute onset vomiting Worsening abdominal bloating Exposure to grandchild who goes to daycare No fever No diarrhea
Optional #3	Gastric outlet obstruction	Intractable nausea and vomiting Progressive worsening symptoms Abdominal bloating

4	3	2	1
Pertinent, prioritized differential diagnosis reflecting correct interpretation of findings	generally consistent with findings, with minor flaws in prioritization that would not negatively impact patient care	Differential diagnosis somewhat consistent with findings, but overly broad and/or inaccurately prioritized	Differential diagnosis is inconsistent with findings
Excellent differential diagnosis well supported, links to diagnoses are correct and complete	Appropriately supported differential diagnosis, with missing or incorrect links that would not negatively impact patient care	Weakly supported differential diagnosis, with several incorrect links between findings and diagnosis	Differential diagnosis lacks accurate links to patient findings

Initial Diagnostic Plan

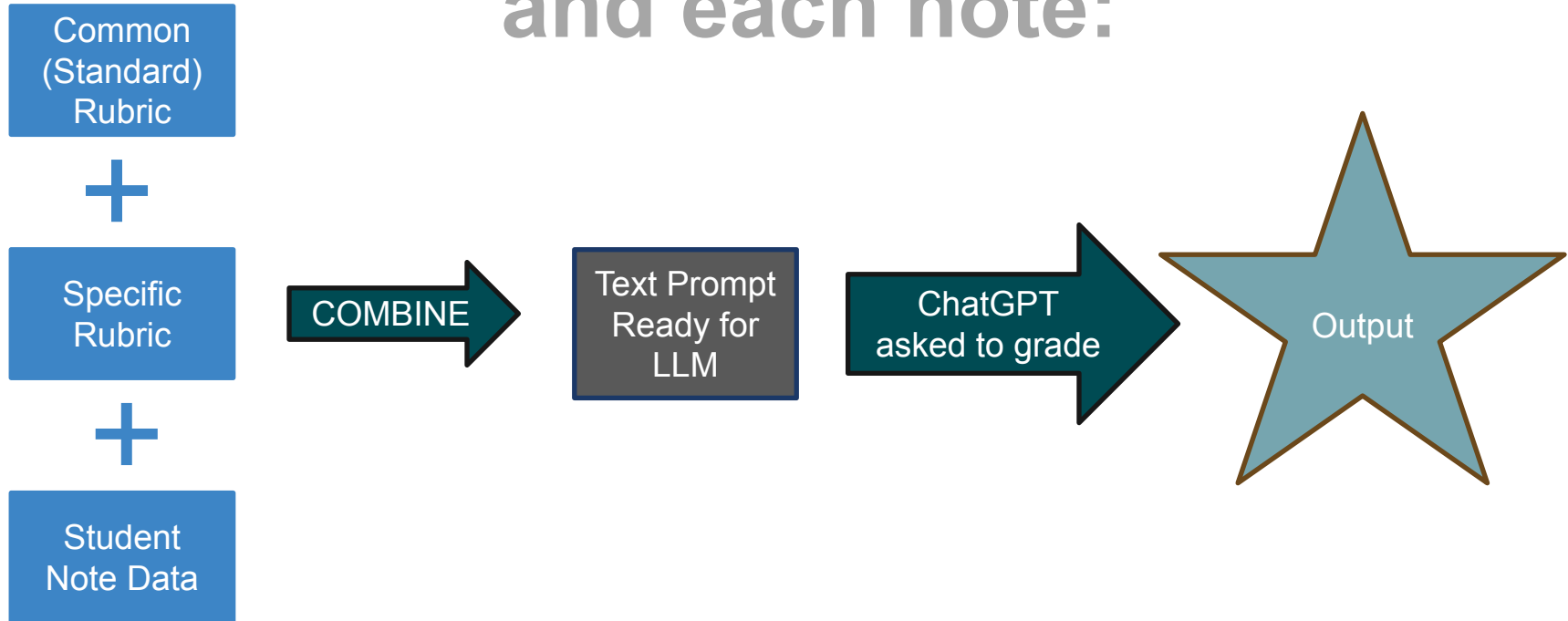
None
Acceptable: Abdominal x-ray/KUB, CT abdomen/pelvis, BMP, A1c
Unacceptable: CBC, endoscopy, stool studies

4	3	2	1
Plan for diagnostic workup is effective and efficient, includes all essential tests, and few or no unnecessary tests	Reasonable plan for diagnostic workup, includes most essential but also some unnecessary tests	Ineffective plan for diagnostic workup, essential tests missed and/or irrelevant tests included ("shotgunning")	Diagnostic workup is inconsistent with differential diagnoses and/or places patient at unnecessary risk

-  kneepainnotes.xlsx
-  nasalcongestionkey.xlsx
-  nasalcongestionnotes.xlsx
-  **nauseavomitingkey.xlsx**
-  nauseavomitingnotes.xlsx
-  syncopekey.xlsx

	A	B	C	D	E	F	G
1	hpi	pex	sum	ddx	support	plan	
	chronic onset nausea/v omiting - Current episode started two days ago - Worsenin g or intractabl e Abdomina l bloating Avoiding large meals Sick exposure via grandchil d who goes to daycare History of	Tachycar dia Minimal pain with palpation diffusely No rebound tenderne ss	1) Acute on chronic OR recurrent 2) Vomiting OR nonbiliou s emesis OR non- bloody emesis 3) History of diabetes 4) Sick exposure Optional) Bloating OR afebrile OR no	1) gastropar esis 2) viral gastroent eritis OR vital gastritis optional #3) gastric outlet obstructi on	esis - Prior episodes of nausea vomiting and bloating - Prior episodes of both after meals - History of diabetes on insulin - Limiting patient from eating big meals Viral Gastroent eritis OR Viral Gastritis - Acute onset	Acceptab le: Abdomina l x- ray/KUB, CT abdomen /pelvis, BMP, A1c Unaccept able: CBC, endoscop y, stool studies	
2	diabetes		diarrhea				

For each rubric domain and each note:



HPI Example:

HPI Shared
Component



Nausea
Vomiting
HPI Items



Student's
Written HPI

----- Interaction -----

system: I am a medical educator and I would like your help grading an assignment. My students recently completed an activity where they interviewed a patient about their symptoms. I have made a scoring rubric that includes the information that should be reported in their post-interview note. The rubric is broken into individual steps. To help you out, I just want you to score each individual section that I will provide you one by one. Please score each individual section that I provide you based off the rubric and the answer key. Please think through this step-by-step. Please place their final score after your explanation all by itself as an integer with no markup

user: Refer to the rubric: 4) Almost all key information is present, concise and well organized with little irrelevant information

3) Most key positive historical findings are documented and well organized, missing pertinent negatives

2) Many key historical findings are omitted. History is disorganized or poorly documented (obscuring what is key)

1) Key history findings are inaccurately documented or omitted. Here is the answer key for hpi:

Acute (6 hours) hematemesis

- Recurrent OR multiple times OR every 30-60min
- Worsening

Abdominal pain, for one month

- Worsening
- Burning OR sharp OR stabbing
- Upper abdomen OR towards the left
- Recurrent OR Episodic

Significant alcohol use (4-5 beers nightly)

Smoking

NSAID use. Please evaluate the following hpi and provide a score: Ms. Maxine Kidd is a middle aged woman presenting to the ER with a 1 month history of abdominal pain and bloating as well as a 6 hour history of vomiting with recent hematemesis. Patient reports 1 month ago she began experiencing minor stomach pain and bloating localized to the epigastric region and LUQ associated with meals especially spicy foods and alcohol. This pain progressed to being fairly constant. Tonight she woke up with sharp pain in the same regions associated with vomiting initially present with just food that progressed to red streaks and eventually gross blood of approximately 0.5 cup according to the patient. She has attempted taking Malox with little to no relief of symptoms. PMH: chronic back pain for approximately 6 years FH: cirrhosis in her uncle, DM in father, Alzheimers dementia in mother SOCIAL: patient drinks about 4-5 beers daily, uses cigarettes and smokes 1ppd for about 20 years, denies IVDU and is not currently sexually active ROS: Patient endorses sore throat associated with the onset of the vomiting tonight as well as abdominal pain for 1 month. Denies symptoms relates to vision, swallowing, chest pain, SOB, diarrhea, urinary difficulties, extremity swelling, weakness, or any fevers or chills.

OVERALL RESULTS



	Average Rating	ICC
Human graders	3.60	0.88
ChatGPT with rationale	2.95	0.76
ChatGPT w/o rationale	2.48	0.94

	ICC	Mean score difference
Human grader x ChatGPT with rationale	0.21	-0.65 <i>(95% CI: 0.61-0.69, p<0.0001)</i>
Human grader x ChatGPT w/o rationale	0.05	-1.12 <i>(95% CI: 1.08-1.17, p<0.0001)</i>
ChatGPT with rationale x ChatGPT w/o rationale	0.29	-0.47 <i>(95% CI: 0.43-0.52, p<0.0001)</i>

Human grader x ChatGPT with rationale: Spearman's $\rho = 0.53$

2026 Emerging Technologies for Teaching and Learning: Digital Demonstrations Virtual Conference

[Register Now](#)

[Already registered?](#)

February 4, 2026 – February 5, 2026

12:00 PM-5:00 PM ET

🕒 Viewing in Eastern Time

When registering for the 2026 Emerging Technologies for Teaching and Learning: Digital Demonstrations Virtual Conference, you'll be prompted to sign in to your AAMC account or set one up if you haven't already done so.

If you have any issues accessing or creating your AAMC account, please email memberservices@aaamc.org for assistance or visit our [Account Help](#) page.

About the Meeting

AI in Medical Education – Innovative and Practical Uses

Temp = 1.0:

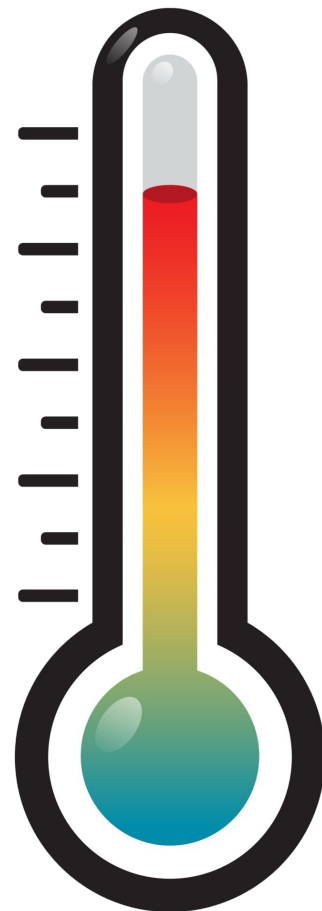
(More expressive, more creative.)

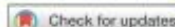
Temp = 0.5:

(Less creative, less likely to hallucinate)


Temp = 0.0:

(Highly deterministic, factual.)





Using large language models (LLMs) to apply analytic rubrics to score post-encounter notes

Christopher Runyon 

Growth and Innovation, NBME, Philadelphia, PA, USA

ABSTRACT

Background: Large language models (LLMs) show promise in medical education. This study examines LLMs' ability to score post-encounter notes (PNs) from Objective Structured Clinical Examinations (OSCEs) using an analytic rubric. The goal was to evaluate and refine methods for accurate, consistent scoring.

Methods: Seven LLMs scored five PNs representing varying levels of performance, including an intentionally incorrect PN. An iterative experimental design tested different prompting strategies and temperature settings, a parameter controlling LLM response creativity. Scores were compared to expected rubric-based results.

Results: Consistently accurate scoring required multiple rounds of prompt refinement. Simple prompting led to high variability, which improved with structured approaches and low-temperature settings. LLMs occasionally made errors calculating total scores, necessitating external calculation. The final approach yielded consistently accurate scores across all models.

Conclusions: LLMs can reliably apply analytic rubrics to PNs with careful prompt engineering and process refinement. This study illustrates their potential as scalable, automated scoring tools in medical education, though further research is needed to explore their use with holistic rubrics. These findings demonstrate the utility of LLMs in assessment practices.

ARTICLE HISTORY

Received 22 November 2024
Accepted 6 May 2025

KEYWORDS

Assessment; learning outcomes; clinical skills; teaching & learning; standardized patients

Fine-tuning

All

Successful

Failed

[Learn more](#)

[+ Create](#)

ft:gpt-4o-2024-08-06:personal::AA118zHw

9/21/2024, 4:20 PM

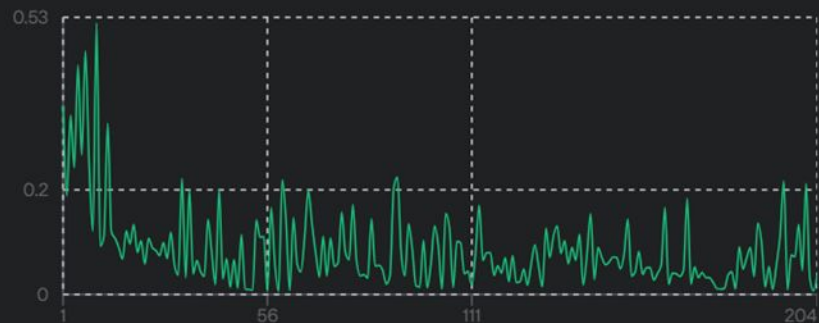
ft:gpt-4o-mini-2024-07-18:personal:grader:9z9jb5xl

8/22/2024, 5:23 PM

[Metrics](#)

Training loss

0.0373



[Messages](#)

[Metrics](#)

16:29:22

The job has successfully completed

[Job](#)

[Playground](#)

christopherjnash/OSCE-Grader x +

← ↻ 🏠 🔒 https://github.com/christopherjnash/OSCE-Grader

OSCE-Grader Public

📌 Pin 📄 Unwatch 1 🍴 Fork 0 ⭐ Star 0

👤 main 1 Branch 0 Tags 🔍 Go to file 📄 Add file <> Code

christopherjnash Update README.md 9a86f75 · 6 hours ago 41 Commits

docs	Create modifying_prompt.md	6 hours ago
examples	Rename flankpainnotes.xlsx to sample_student_notes.xlsx	6 hours ago
scripts	Create convert_rubric.py	6 hours ago
.gitignore	Create .gitignore	6 hours ago
LICENSE	Update LICENSE	yesterday
README.md	Update README.md	6 hours ago
requirements.txt	Create requirements.txt	6 hours ago

📖 README 📄 MIT license

OSCE Grader Setup Guide 🚀

An AI-powered grading system for medical student OSCE post-encounter notes, using GPT models to automate grading and provide structured feedback.

Features

- ✔ Supports Excel, CSV, and text-based student notes
- ✔ Works with structured rubrics (Excel, CSV)
- ✔ Uses ChatGPT to generate detailed, section-by-section grading
- ✔ Customizable grading prompts via config.py
- ✔ Automatic logging for debugging and analysis
- ✔ Easy setup with Python & OpenAI API
- ✔ Includes convert_rubric.py to assist in converting rubric files (PDF/DOCX to structured formats)

About

No description, website, or topics provided.

- 📖 Readme
- 📄 MIT license
- 📈 Activity
- ⭐ 0 stars
- 👁 1 watching
- 🍴 0 forks

Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Languages

- Python 100.0%

Suggested workflows

Based on your tech stack

- SLSA Generic generator** Configure
Generate SLSA3 provenance for your existing release workflows
- Python package** Configure
Create and test a Python package on

AI-Powered Evaluation Analysis

Original Faculty Comments

Comments for MSPE and Student Review

Has a good fund of knowledge. Knew many differentials to consider Good empathy and interpersonal skills. Make sure you work on your differential diagnosis. Many of these patients can have multiple causes for regional pain. Great job with history taking and coming up with plans which is above and beyond what is expected in first month of third year. Also did well with efficiency and understanding which patients looked sick vs not sick in the ED She has an excellent knowledge base and is advanced at providing differentials and recommending workup. Did well on a busy shift. Always present and curious to learn. Notes presented appropriate information Student Dr. X4 was excellent on our EM TeachER shift together. X4 was an engaged learner, easy to get along with, eager to learn, and willing to be a helpful team member. We did her Direct Observation Form, saw necessary RCEs like pts with CC of headaches and seizures, and went to critical call outs. X4 went to a floor code, helped perform bedside POCUS, and assisted in a critically ill trauma patient. We also talked through practice questions and high yield medical topics, and she is clearly intelligent. X4 will do very well this year, in her residency, and in the future after that.

Comments for Student-Only Review

Continue to read and expand on differentials. Develop a broad differential Continue to work on physical exam findings and developing broad differentials based on different body systems. N/a Continue to expand differentials and show interest in all fields, you will do well! Great job! Keep up the good work. Let us know if we can help you at all this year!

AI-Generated Formative (Student-Facing) Evaluation

Feedback Summary for Your Emergency Medicine Clerkship

Strengths

You have demonstrated a solid foundation of medical knowledge and interpersonal skills throughout your clerkship. Your ability to take comprehensive histories and formulate plans is commendable, especially as a first-month third-year student. Your engagement during busy shifts and your curiosity to learn have been noted as significant strengths. For instance, your participation in critical scenarios, such as a floor code and assisting in a critically ill trauma patient, showcases your ability to handle high-pressure situations effectively.

***Quotes from Faculty:** "X4 was an engaged learner, easy to get along with, eager to learn, and willing to be a helpful team member."*

Areas for Development and Prescriptions for Improvement

Area for Development: Differential Diagnosis

Observed Behavior (The 'What'): While you have a good understanding of differential diagnoses, there is room for improvement in developing broader differentials, particularly for patients with regional pain and complex presentations.

Clinical Significance (The 'Why'): A comprehensive differential diagnosis is crucial in Emergency Medicine, as patients often present with overlapping symptoms. Being able to consider multiple causes will enhance your diagnostic accuracy and improve patient outcomes.

Actionable Suggestions (The 'How'):

1. Regularly review case studies or clinical scenarios that involve patients with regional pain to practice developing broad differentials.
2. Engage in discussions with peers or mentors about differentials for various presentations, ensuring you consider multiple body systems.
3. Utilize online resources or textbooks to expand your knowledge of less common conditions that could present similarly to common complaints.

***Supporting Evidence:** "Make sure you work on your differential diagnosis. Many of these patients can have multiple causes for regional pain."*

Area for Development: Physical Exam Skills

Observed Behavior (The 'What'): There is a noted need for continued development in physical exam findings related to various body systems.

AI-Generated Summative (MSPE-Facing) Evaluation

The student demonstrates a solid foundation in medical knowledge, evidenced by her ability to generate differential diagnoses and propose appropriate workups, particularly in emergency situations. She exhibits strong interpersonal skills, effectively engaging with patients and colleagues, which enhances her learning environment. Faculty noted her proficiency in history taking and her capacity to discern the acuity of patient conditions, further highlighting her clinical reasoning skills. However, there is a need for continued development in her differential diagnosis formulation, particularly in complex cases where multiple etiologies may be present. The student actively seeks opportunities to learn and contribute, as shown by her participation in critical procedures and her engagement during teaching sessions. Overall, she has shown commendable performance in her early clinical experiences, with clear potential for further growth in her diagnostic skills and clinical decision-making as she progresses in her training.

GEA

Group on Educational Affairs



Part 3: Reflections on sound and reliable AI-assessment approaches

Pedagogical and Theoretical Frameworks

Passive, Interactive, Creative, Replacement, Amplification, Transformation (PICRAT)

- P: learning time, assessment attempts
- I: Socratic type assessment
- C: personalized learning
- R: MCQ generation
- A: personalized feedback
- T: large data set analysis

Kimmons *et al.* (2020)

Technology, Pedagogy, Content, Knowledge (TPACK)

- Examine the how (pedagogy), what (content), and why of technological integration

Schmidt *et al.* (2009)

Considerations for Continuous Improvement

01

Theoretical frameworks and pedagogical knowledge

02

Continuous faculty oversight and development

03

Reflections on bias, equity, access, and ethical issues

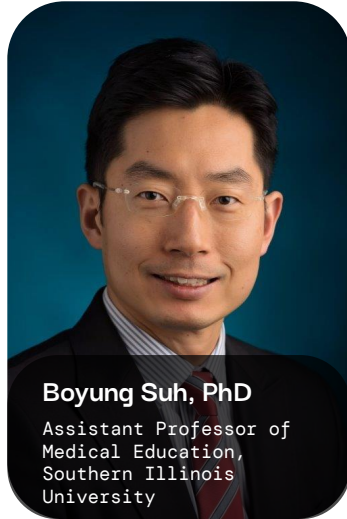
04

Learner psychological safety, privacy, and consent

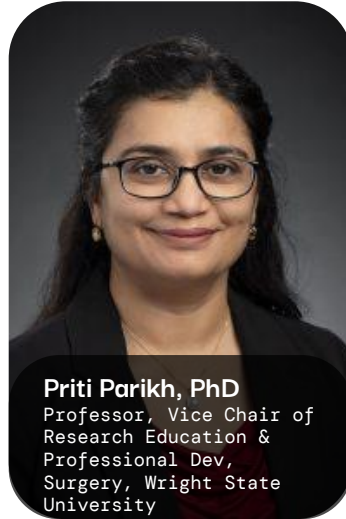
References

- Masters K, MacNeil H, Benjamin J, et al. Artificial Intelligence in Health Professions Education assessment: AMEE Guide No. 178. *Medical Teacher*. 2025/09/02 2025;47(9):1410-1424.
- Costa-Dookhan KA, Adirim Z, Maslej M, et al. Applications of Artificial Intelligence for Nonpsychomotor Skills Training in Health Professions Education: A Scoping Review. *Academic Medicine*. 2025;100(5):635-644.
- Turner, Laurah PhD; Hashimoto, Daniel A. MD; Vasisht, Shubha; Schaye, Verity MD. Demystifying AI: Current State and Future Role in Medical Education Assessment. *Academic Medicine* 99(4S):p S42-S47, April 2024.
- Cole, Jacob MD; Duncan, Joshua MD; Cole, Rebekah PhD, MEd. Using Pretrained Large Language Models for AI-Driven Assessment in Medical Education. *Academic Medicine* 100(12):p 1442-1446, December 2025.
- Kimmons R, Graham CR, West RE. The PICRAT model for technology integration in teacher preparation. *Contemp Issues Technol Teach Educ*. 2020;20(1):176–198.
- Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., & Shin, T. S. (2009). Technological Pedagogical Content Knowledge (TPACK): The Development and Validation of an Assessment Instrument for Preservice Teachers. *Journal of Research on Technology in Education*, 42(2), 123–149.

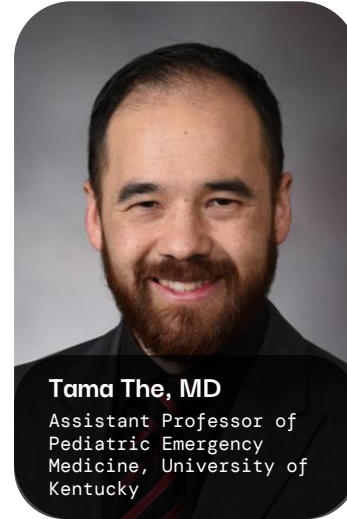
Contact Us



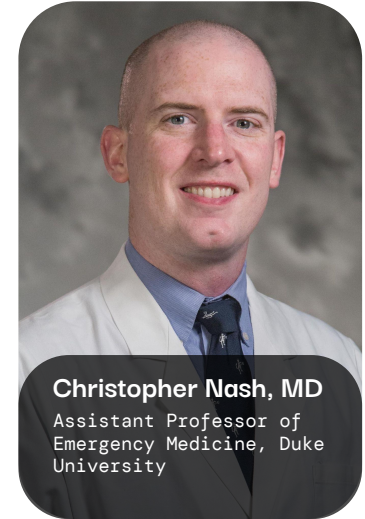
bsuh46@siumed.edu



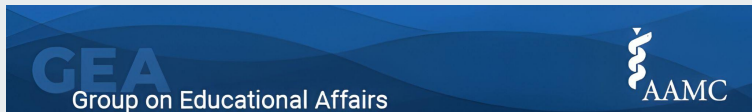
priti.parikh@wright.edu



tama.the@uky.edu



chris.nash@duke.edu



CME & Session Evaluation



The College of Human Medicine Office of CME designates this event for a maximum of 1 *AMA PRA Category 1 Credit(s)*[™]. Physicians should claim only the credit commensurate with the extent of their participation in the activity.



MICHIGAN STATE
UNIVERSITY

College of Human Medicine
Office of Continuing Medical Education

Michigan State University - College of Human Medicine - Office of Continuing Medical Education is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to provide continuing medical education for physicians.



Next Session (Jan. 15, 1-2 pm, EST)

Preparing for AI Integration in Clinical Education

CME ACTIVITY INFORMATION

Activity Name: Enduring: AI for Assessment and Evaluation in Medical Education

Date: December 22, 2025 – May 14, 2028

Activity Code: 95382

Speaker Name: Christopher Nash, MD, EdM; Priti Parikh, PhD; Boyung Suh, PhD; Tama The, MD

Target Audience: Regional physicians, residents, students, faculty

Learning Objectives: *at the conclusion of this educational activity, learners will be able to:*

Objective 1: Understand the rationale and opportunities for AI-assessment and evaluation.

Objective 2: Construct a reliable assessment tool using an application programming interface (API).

Objective 3: Reflect on considerations for sound and reliable AI-assessment approaches.

Accreditation Statement:

The Michigan State University is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to provide continuing medical education for physicians.

Michigan State University designates this activity for a maximum of 1.0 AMA PRA Category 1 Credits™. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Summary of Financial Disclosures:

All planners, reviewers, faculty presenters have nothing to disclose.

Commercial Support Disclosure:

No commercial support was provided for this CME activity

Review Date: DECEMBER 22, 2025

Original Release Date: DECEMBER 18, 2025

Termination Date: MAY 14, 2028

To Earn CME Credit: *Completion of each of these steps is required to earn CME credit.*

1. Complete the attest process to your time in attendance by visiting <https://cmetracker.net/MSU> (case sensitive).
 - Click on the **Sign In** option on the left menu
 - Enter your email and password to log into the system. You will be required to create a profile if you have not used the system before.
 - Enter the activity code provided on this sheet.
 - Complete the evaluation and attest to your time in attendance, then follow the screen instructions to print your certificate. Make sure your computer is set to allow pop-ups from the site or the certificate will not show.

If you have any issues obtaining your certificate, please contact the CME Office at handyrya@msu.edu or by calling 517-884-8873.