Using LLMs to Assist with Grading OSCE Notes

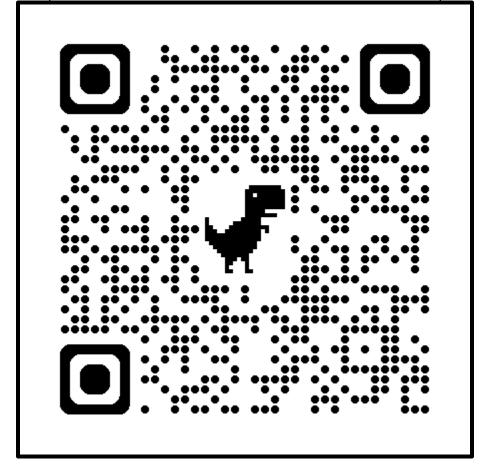
Christopher Runyon

Albert Einstein School of Medicine

Davidoff Day

April 7, 2025

Presentation Slides:



(both presentations available)

Disclaimer

The information presented in this document/presentation is solely for informational purposes and does not necessarily reflect the official stance or views of NBME. The opinions expressed herein are those of the authors/presenters and may not represent the views or policies of NBME. While every effort has been made to ensure the accuracy and reliability of the information provided, NBME makes no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability with respect to the content contained herein. Any reliance you place on such information is therefore strictly at your own risk.

Learning Objectives

• Understand the differences between analytic and holistic rubrics and their implications for GPT-assisted grading.

 Apply effective prompt engineering techniques for assessing medical education tasks.

 Identify key considerations for implementing GPT-based grading systems.

OSCE Scoring Rubrics

Analytic Rubrics

- Identifies specific information to be included
- Each element scored independently
- Time intensive
 - Potentially inaccurate
- Construct underrepresentation
- Easier for feedback?

Clinical Concept	Point Value
32-year-old male (5 if gender omitted)	1
right heel hurts (-1 if side omitted)	2
3-week history of pain	1
pain is worse in morning or after prolonged inactivity	1
pain is better after activity	1
no recent trauma, redness, or swelling	1
avid runner	1
normal vital signs	1
unremarkable family or social history	1
tenderness to deep palpation of the right medial heel	1
Diagnosis is plantar fasciitis	3
Diagnosis is heel fracture	1
If note is about a different clinical condition	0; discard rubric

Scoring Rubrics

Analytic Rubrics

- Identifies specific information to be included
- Each element scored independently
- Time intensive
 - Potentially inaccurate
- Construct underrepresentation
- Easier for feedback?

Holistic Rubrics

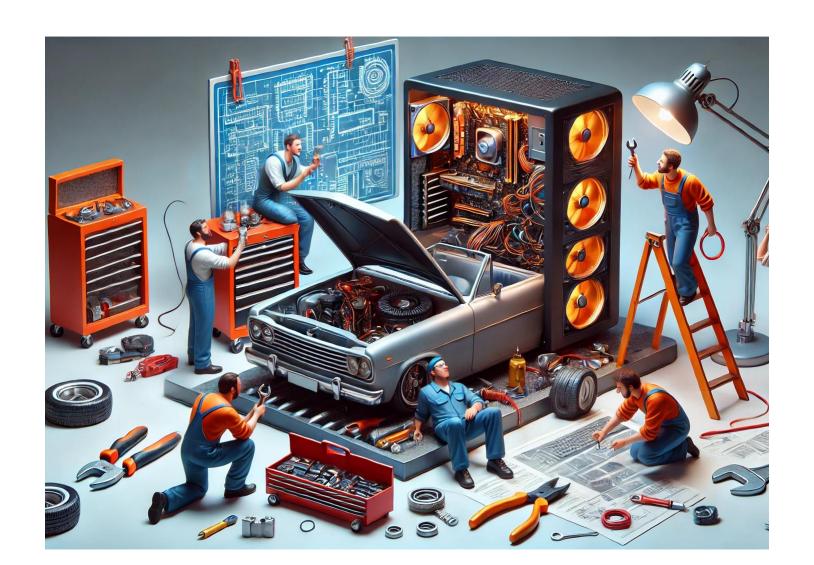
- More global assessment
- Usually has score categories
- Easier to implement
- More subjective
 - Validity threat
- Time-intensive feedback

ASSESSMENT of REASONING TOOL



Learner: Evaluator: MIEDICINE				
Did the Learner	Assessment			
Did the Learner	Minimal	Partial	Complete	
Collect/report history and examination data in a hypothesis-directed manner?	Non-directed in questioning and exam Asked questions without clear focus on potential diagnoses	Questioning and exam generally reflective of potential diagnoses, but some less relevant or tangential questions	Followed clear line of inquiry, directing questioning and exam to specific findings likely to increase or decrease likelihood of specific diagnoses	
Articulate a complete problem representation using descriptive medical terminology?	Included extraneous information Missed key findings Did not translate findings into medical terminology	Generally included key clinical findings (both positive and negative) but either missed some key findings or missed important descriptive medical terminology	Gave clear synopsis of clinical problem Emphasized important positive and negative findings using descriptive medical terminology	
Articulate a prioritized differential diagnosis of most likely, less likely, unlikely, and "can't miss" diagnoses based on the problem representation?	Missed key elements of differential diagnosis, including likely diagnoses or "can't miss" diagnoses	Gave differential diagnosis that included likely and "can't miss" diagnoses but either missed key diagnoses or ranked them inappropriately	Gave accurately ranked differential diagnosis including likely and "can't miss" diagnoses	
Direct evaluation/treatment towards high priority diagnoses?	Directed evaluation and treatment toward unlikely/unimportant diagnoses Did not evaluate or treat for most likely/"can't miss" diagnoses	Major focus of evaluation and treatment was likely and "can't miss" diagnoses but included non-essential testing	Efficiently directed evaluation and treatment towards most likely and "can't miss" diagnoses Deferred tests directed towards less likely or less important diagnoses	
Demonstrate the ability to think about their own thinking (metacognition)? Consider asking: Is there anything about the way you are thinking or feeling about this case that may lead to error?	Not able to describe the influence of cognitive tendencies or emotional/ situational factors that may have influenced decision-making	Can name one cognitive tendency or emotional/situational factor that may have influenced decision-making		
OVERALL ASSESSMENT	NEEDS IMPROVEMENT	MEETS COMPETENCY	EXCELLENCE	
Comments:				

A few technical details...



How LLMs Make Sense of Language

- Underlying principle: distributional semantics / co-occurrence
 - Words that occur in the same contexts tend to have similar meanings. (Harris, 1954)

"We found a little, hairy sleeping behind the tree."

- Not deterministic; variation is part of its beauty
- A very, very fancy text calculators



Attention

• The attention mechanism in a Gen AI models helps the model focus on the most relevant words in the sequence (i.e., the prompt).

"We found a little, hairy sleeping behind the tree."

- Words like "little", "hairy", and "sleeping" provide important information about the wampimuk.
- Other words less so: "we", "found", "behind", "tree"
- Others still even less: "a" "the"

Attention

• LLMs, like humans, have limited attention

- Prompts that give LLMs many tasks at once may result in suboptimal performance.
 - Difficult to identify what are the most important parts of the task.
 - May do some tasks well and others poorly; or all only moderately well.
 - Some newer models ("reasoning models") are trying to explicitly address this limitation.

Attention

- The attention can also be impacted by the length of the conversation you have (when using the chatbot versions of Gen Al Models)
- May start the conversation well, but "forget" important parts later
 - Can be a problem when building a Gen AI model to act as a standardized patient (or similar role)
- Alternatively, the Gen AI model can also fixate (over-attend) to details that you didn't intend to be important and unnecessarily carry those forward in the conversation
 - Can be mitigated with subsequent prompts provide it with feedback!

Model parameters





- Temperature
 - "What sampling temperature to use, between 0 and 2. Higher values ... will make the output more random, while lower values ... will make it more focused and deterministic."
 - Sides of a die
 - Does not guarantee that result will be exactly reproduced.
- top_p (nucleus sampling)
 - "An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with top_p probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered."
 - Fairness of the die

"It is recommended to alter [one or the other] but not both."

Variability in LLM Responses

Higher temperatures (more creative)

- Content generation
- Clinical case development
- Simulating patient responses
- Generating varied feedback examples

Lower temperatures (more consistent)

- Scoring assessment tasks
- Creating uniform materials
- Standardizing OSCE patient cases
- Summarizing tasks
- Ensuring consistent feedback





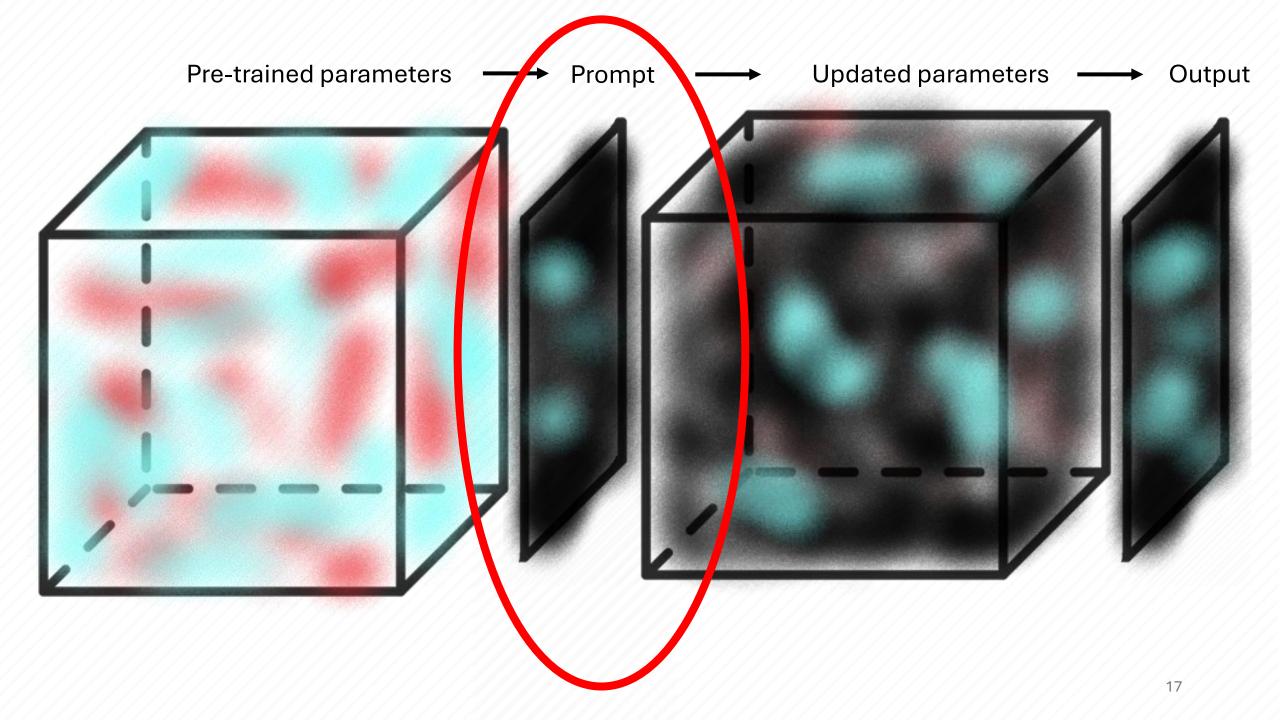
Variability in LLM Responses

Model	ChatGPT 4o
n replications	100
Temperature	0

- Replicated pizza prompt
- 5 different responses:
 - **(89)** "Delicious, versatile, customizable, and universally loved comfort food with endless topping possibilities."
 - (7) "Pizza combines diverse flavors and textures, customizable toppings, and universal appeal in one dish."
 - (1) "Delicious, versatile, and customizable; perfect for any occasion and loved by almost everyone."
 - (1) "Delicious, versatile, and customizable; pizza satisfies cravings with endless topping combinations."
 - (1) "Delicious, versatile, customizable, satisfying, and perfect for sharing; pizza pleases almost everyone."
 - (1) "Delicious, versatile, customizable, satisfying, and universally loved comfort food with endless topping possibilities."

Prompt Engineering

- There are ways to interact with GPTs that can result in more useful responses
 - Provide more context for what you would like the GPT to output
 - Help activate more relevant connections in the network
- We're going to focus on two general strategies for prompting
 - Prompt formulas
 - Prompt refinement
 - Prompt chaining / chain-of-thought prompting



"Degrees of Freedom" in LLM Prompts

- When prompting an LLM, any detail you don't specify explicitly will be inferred by the model.
 - Good = want general creative help
 - Bad = have a specific assessment point in mind

A vacation is when	you take a trip to some _	place
with your	family. Usually y	ou go to some place
that is near a/an	or up on a	ı/an NOUN
A good vacation pla	ice is one where you can	ride
or play	or go hunting for _	PLURAL NOUN . I like
to spend my time _	VERB ENDING IN "ING"	VERB ENDING IN "ING"

"Degrees of Freedom" in LLM Prompts

• When prompting an LLM, any detail you don't specify explicitly will be inferred by the model.

• Example: "Develop an assessment question about an illness that could be caused by two different organ systems."

• Filling in these unknowns ("degrees of freedom") when providing instructions to the LLM will increase the likelihood that the result will be closer to what you want.

"Degrees of Freedom" in LLM Prompts

• Example: "Develop an assessment question about an illness that could be caused by two different organ systems."

What's missing here?

"Develop an assessment question about an illness that could be caused by two different organ systems."

- Purpose of assessment
- Level of learner

- Difficulty
 - What makes it difficult?

- Length / format of question
 - Multiple-choice?
 - Vignette-based?
 - · Chart-based?
 - Written response?

- Patient Characteristics
 - Age, Gender, Sex
 - HPI
 - PMH

- The organ systems
- Leading diagnosis? ("Truth")
- Specific illnesses
- Method of differentiation
 - Dx test?
 - Other?

Prompt formulas

- Structured guidelines or templates used to generate responses from language models like GPT-40
 - Help shape the context, style, and content of the output

ROLE

TASK

GOAL

[I am in my mid 40s and I am becoming more interested in film.] [I would like to learn more about the intersection of romance and sci-fi genres.] [Please suggest 10 popular movies in romance/sci-fi genre and provide details on why these movies are thought to be important or interesting.]

ROLE

TASK

GOAL

[I am organizing a conference], and I would like help [developing icebreaker interactive sessions] for conference participants. [Please suggest 3 icebreaker interactive activities, and provide details such as timing, materials needed, and activity objectives. Suggest a 30-minute, 60-minute, and 90-minute activity.]

CONTEXT

ACTION

RESULT

EXAMPLE

[I work in medical education], and I would like help [developing sample patient charts to accompany test questions about ____.]

[Each patient chart should have the same structure and contain similar content but should differ on what the final diagnosis or best patient management should be.]

[After I submit this prompt, I will then provide an example format that the patient chart should follow.]

Prompt Formulas

Providing an Example Output helps do 2 things:

- It provides a replicable example for the GPT to emulate.
 - No guarantee the GPT will precisely replicate the output, but, if not, will be close.
- The content of the example acts as additional instructions.
 - The GPT will evaluate the relationship of the content to the structure of the output to infer more information about your prompt.

Structuring Prompts

• In addition to using prompt formulas (or other prompting strategies), strategically using punctuation—such as braces or brackets—can highlight key parts of your prompt and guide the LLM's focus.

Role: [You are a medical educator specializing in internal medicine.]

Task: [Review the following clinical vignette and create a single best answer multiple-choice question with four options. Clearly indicate the correct answer.]

Vignette: [A 45-year-old man comes to the emergency department with sudden onset of shortness of breath and pleuritic chest pain. He has a history of long-haul flights and recent orthopedic surgery. On examination, he is tachycardic and his oxygen saturation is 89% on room air. A chest CT scan shows a filling defect in the right pulmonary artery.]

Additional Instructions: [Ensure the correct answer focuses on the most likely diagnosis based on the vignette.]

- Prompt the LLM to provide feedback on the structure of the prompt
 - Structure
 - Areas in need of more detail

Using a prompt formula:

"I am a [ROLE] and I am building a prompt to [TASK]. The goal of the prompt is to [GOAL]. Review the following prompt I have developed and provide suggestions to improve (1) the prompt structure and (2) instructions where more detail would be beneficial so the prompt will more effectively help me achieve my goal."

I am a [ROLE], creating a prompt to [TASK]. The goal of this prompt is to [GOAL].

What I Need:

- **1. Review** the prompt I have drafted (below).
- 2. Suggest how to enhance its structure and clarity.
- **3. Explain** if there are areas where more specific instructions are needed to ensure the final output meets my goal.
- **4. Provide** a brief rationale for each improvement and, if possible, a sample rewrite.

Draft Prompt to Review:

[Paste Your Draft Prompt Here]

Important Details:

- Format your suggestions in bullet points.
- Keep the total word count under 200 words.
- The audience for the final output is [specific audience]."

- Prompt the LLM to provide feedback on the structure of the prompt
 - Structure
 - Areas in need of more detail

Using a prompt formula:

"I am a [ROLE] and I am building a prompt to [TASK]. The goal of the prompt is to [GOAL]. Review the following prompt I have developed and provide suggestions to improve (1) the prompt structure and (2) instructions where more detail would be beneficial so the prompt will more effectively help me achieve my goal."

I am a [ROLE], creating a prompt to [TASK]. The goal of this prompt is to [GOAL].

What I Need:

- **1. Review** the prompt I have drafted (below).
- **2. Suggest** how to enhance its structure and clarity.
- **3. Explain** if there are areas where more specific instructions are needed to ensure the final output meets my goal.
- **4. Provide** a brief rationale for each improvement and, if possible, a sample rewrite.

Draft Prompt to Review:

[Paste Your Draft Prompt Here]

Important Details:

- Format your suggestions in bullet points.
- Keep the total word count under 200 words.
- The audience for the final output is [specific audience]."

Chain Prompting

Chain-of-Thought Prompting: Encourages ChatGPT to articulate its reasoning process in steps

- GPT breaks a complex task into smaller tasks
- Method improved by providing an example
- "Think step-by-step"

Prompt Chaining: Sequentially linking multiple prompts where each step's output feeds into the next

- *User* breaks down a complex task into smaller tasks

Chain-of-Thought Prompting

Best Use Cases

- Complex reasoning tasks
 - Explaining complex concepts
- Creative tasks
 - Can provide a "thought process" leading to the result
- Inferences & Explanations
 - Can include intermediate steps like interpretations and how it reached a conclusion

Less Effective Use Cases

- Procedural or Step-by-Step Tasks
 - Unnecessary elaboration, confusing the model
- Simple Prompts
 - Adds fluff w/o improving accuracy
- Strict Formatting Requirements
 - Can make responses messy

Prompt Chaining

 Helpful to reduce the "cognitive load" of the LLM; better control of the model attention.

- Better control over what the model does at each step.
 - More detailed instructions on each step

- More transparency / error isolation
 - Have intermediate records
 - Can catch early errors that propagate forward

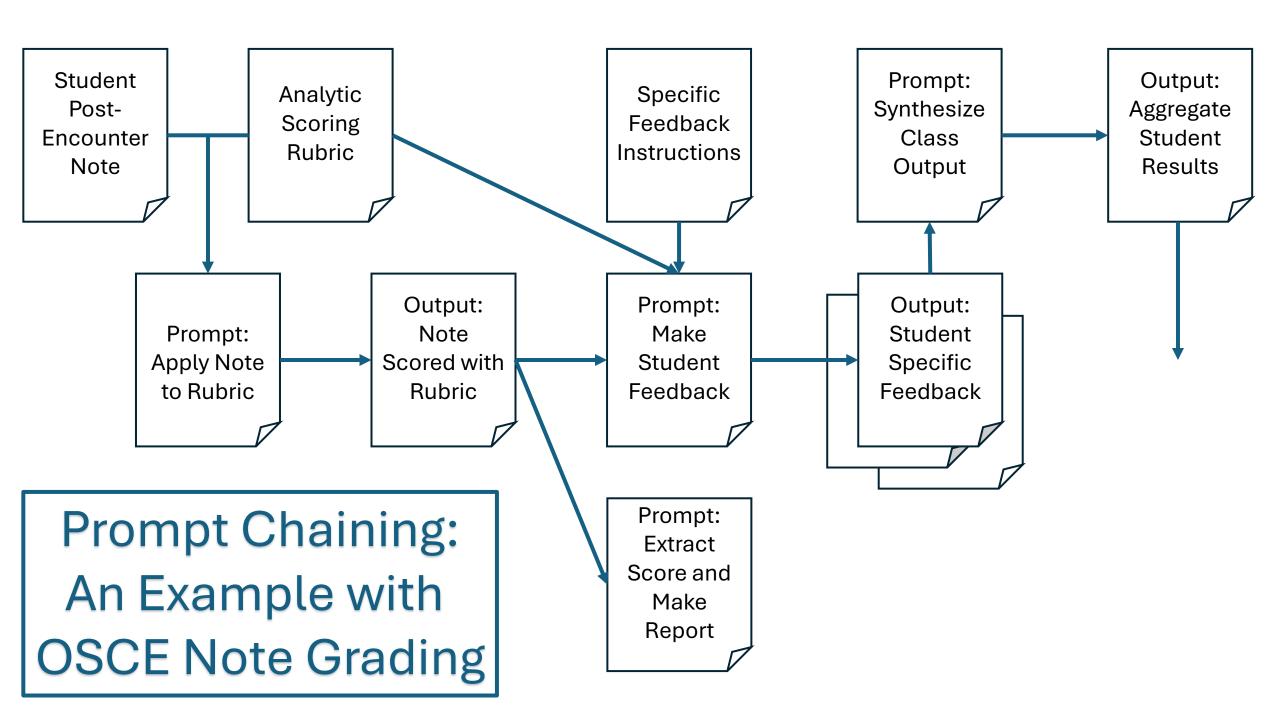
Prompt Chaining

Strengths

- Modular approach
- Reduced complexity of instructions
- Error isolation
- Clear, structed workflow

Weaknesses

- Requires more planning
- User errors
 - Information must manually be passed to each subsequent step
- More time and resource intensive
 - More complex than a singleprompt process



Mitigating Variability in LLM Responses

- Model-based strategies
 - Control temperature, top_p, length of output, etc.

- Prompt-based strategies
 - Using prompt chaining / short prompts
 - Provide example of output structure

- Materials-based strategies
 - Refine input materials
 - Refine instructions

Materials-based Variability Reduction

 Example: Using LLM to apply analytic rubric to score a (fabricated) post-encounter note

 Provided the LLM with the rubric, note, and a prompt to score the rubric

 Still noticed unexpected score variability when using prompt chaining method

Materials-based Variability Reduction

Content	Points
Tenderness to deep palpation of the right medial heel	1

"A musculoskeletal exam revealed some pain during a deep palpation on the right heel."

"Didn't say medial, so does not get credit."

"Didn't say medial, so only received partial credit (1/2 point)."

"Full credit."

Can be useful for refining rubrics: score a note many times and investigate any score variation

Validity Issues with LLM Scoring

Validity is the extent to which an assessment's scores accurately reflect what it is intended to measure for a particular **purpose** and **population**

A **validity argument** is the systematic accumulation of evidence demonstrating that the assessment is appropriate for its intended use.

 No assessment is universally "valid" or "invalid" – depends on the use of scores in a particular context

Key assumption: the scoring method accurately represents the examinee's performance on the specific construct being measured.

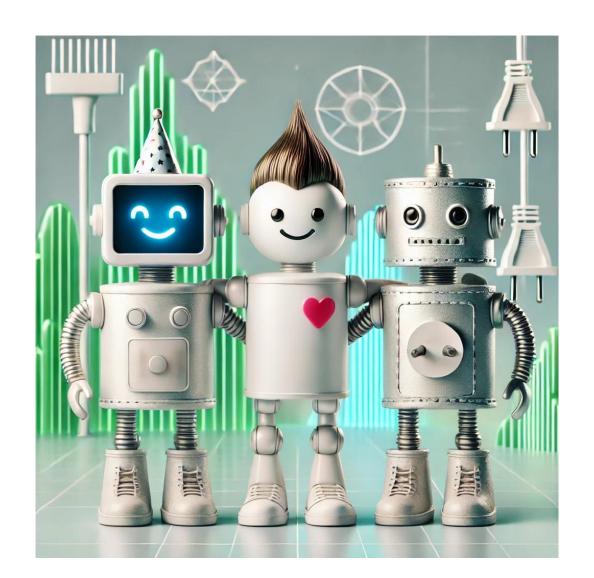
Validity Issues with LLM Scoring

- Are LLM scores consistent and reliable?
- How does the LLM arriving at a specific score? (Transparency)
- How has the LLM's scoring system been tested to ensure it works as intended?
- Is the scoring aligned with the intended learning objectives and outcomes?
- Does the scoring method fully capture the knowledge or skills the assessment is intended to measure?

Testing LLM Graders

- Have at least 2 human graders to compare to LLM scores on a set of sample assessments
 - Compare interrater reliability between humans, each human and the LLM (Human 1 & Human 2; Human 1 & LLM; Human 2 & LLM)
 - Goal: have LLM score as reliably and accurately as any 2 humans
 - Other methods of comparison:
 - Quadratic Weighted Kappa (QWK): suited for ordinal data
 - Category-adjacent scores : can see if LLM is consistently low or high

Chatbots, Custom Chatbots, and APIs



LLM Chatbots



- Resembles human (text) conversations
- Easy to use; requires no programming knowledge
- Can easily provide feedback and continue prompting sessions
- Requires a subscription
 - Some free options available; some have limited functionality / use
- Cannot easily manipulate model parameters (e.g., temperature)
- Can be laborious for use at scale

LLM Chatbots

(A)

I will be traveling to Milwaukee for work and will have an extra day and a half to explore the city. What are some interesting things to do in Milwaukee?

Milwaukee has a lot to offer! You could visit the Milwaukee Art Museum for its impressive architecture and collections, take a brewery tour (like at Lakefront Brewery), or explore the Harley-Davidson Museum if you're into motorcycles. If the weather is nice, a stroll along the Riverwalk or a visit to the historic Third Ward for food and shopping is a great option. Let me know what interests you, and I can suggest more!

I really like trying new foods. Is there any cuisine local to the area?

Yes! Milwaukee is known for its cheese curds (try them fried), bratwurst, and fish fry, which is a local tradition, especially on Fridays. The city also has a strong German heritage, so you'll find great pretzels, schnitzel, and beer halls. If you want something unique, check out butter burgers at Solly's or frozen custard from Leon's or Kopp's. Want recommendations for specific restaurants?

Yes, please provide a few restaurants that would be fun to try that specialize in some of the local cuisine you mention.

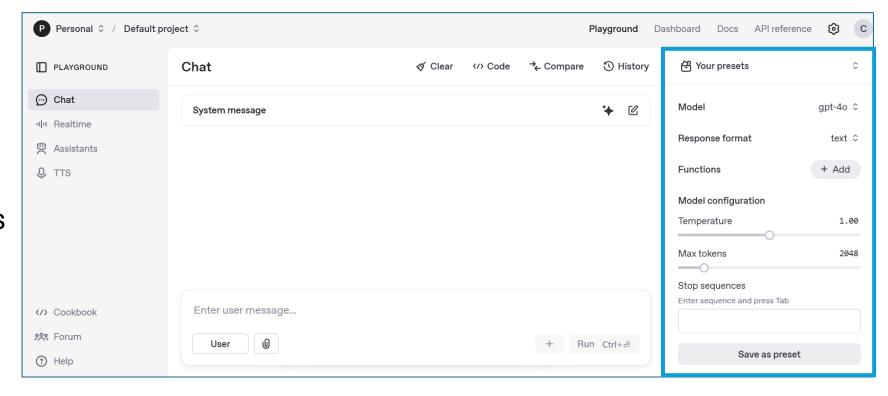
Sure! Here are a few great spots to try local Milwaukee cuisine:

- Kopp's Frozen Custard Famous for their rich custard and butter burgers.
- · Lakefront Brewery A fun spot for a beer and classic Wisconsin cheese curds.
- The Old German Beer Hall Great for bratwurst, pretzels, and German beer.
- The Packing House A top spot for a tradition \(\psi\) ilwaukee fish fry.



LLM Chatbots

- Some (e.g., ChatGPT) have a "playground" mode
 - Requires a subscription (as of this writing)
- Can alter model parameters



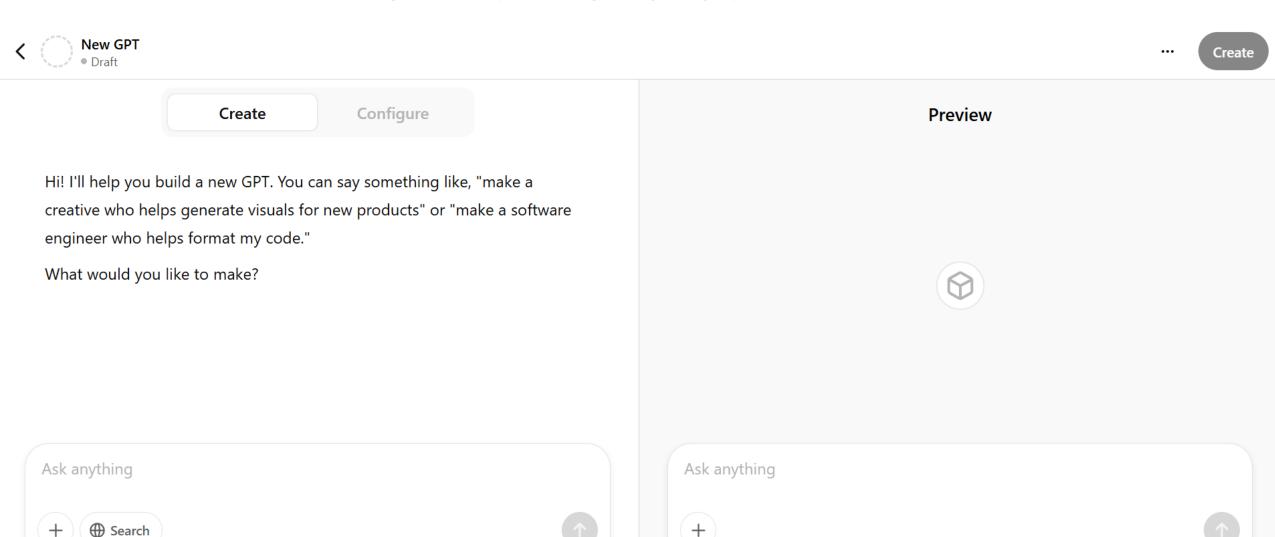
Custom Chatbot LLMs



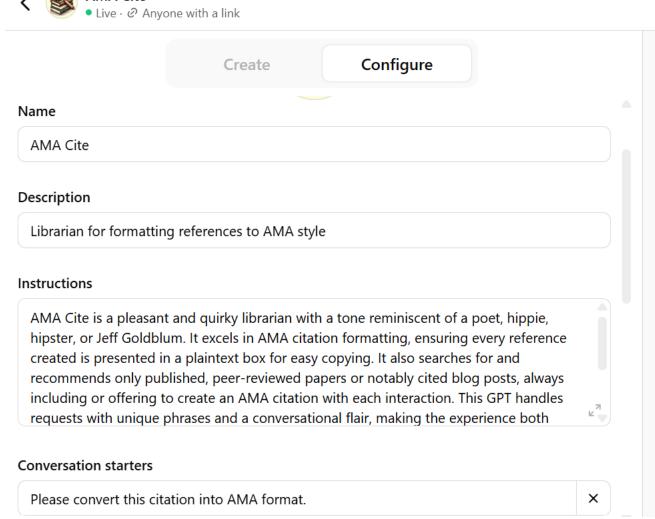
 More companies are offering the ability to easily make Custom GPTs

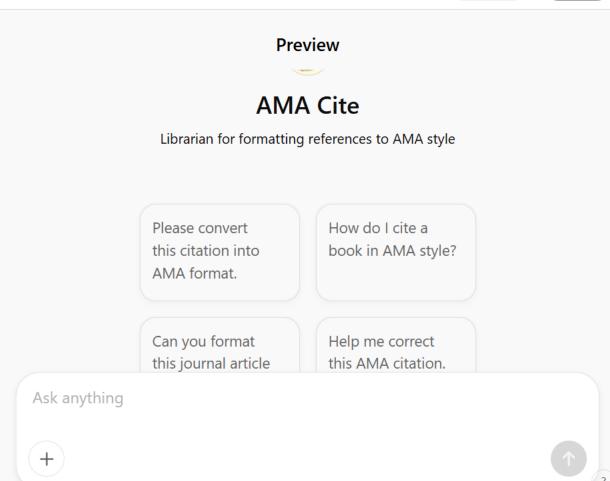
- Useful for repeating tasks; don't have to re-prompt the model or provide instructions again
 - Global email writer (clarity, conciseness, and replace idioms)
 - Citation helper (paste Bibtex format and indicate desired reference style)

Custom Chatbot LLMs



Custom Chatbot LLMs





Updates pending

Last edited Mar 3

⊘ Share

Update

Accessing LLMs via API



- Requires programming knowledge (e.g., Python, R, etc.)
- Can be build into a workflow for larger tasks
- Easier to manipulate model parameters
- Cost per use
 - Based on number of tokens submitted and number of tokens returned
- Cannot easily mimic conversations
 - Must include past exchanges in prompt requests
 - Must consider prompt token limits
 - Can summarize previous conversation or drop older messages

Accessing LLMs via API

```
☐ Copy  % Edit
python
import openai
# Set your OpenAI API key
openai.api key = "your-api-key-here"
# Define the prompt
user input = "Explain the concept of clinical reasoning in simple terms."
# Create a request to ChatGPT
response = openai.ChatCompletion.create(
    model="gpt-4-turbo", # Specify the model (GPT-4 or GPT-3.5-turbo)
    messages=[{"role": "system", "content": "You are a helpful medical assistant."},
              {"role": "user", "content": user input}]
# Extract and print the response
assistant_reply = response["choices"][0]["message"]["content"]
print("ChatGPT:", assistant reply)
```

API Key = specific to a user (or project)

A long series of letters and numbers:

'sk-1234567890abcdefGHIJKLMNOPQRSTUVWXYZ1234'

OSCE Grader

https://github.com/christopherjnash/OSCE-Grader



কা MIT license



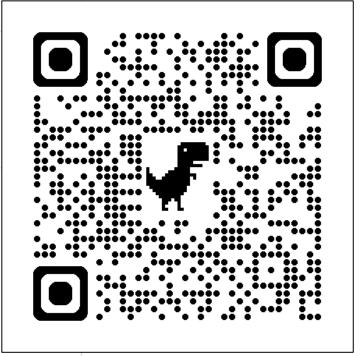
OSCE Grader Setup Guide 🚀



An Al-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 & OpenAl API
- Includes convert_rubric.py to assist in converting rubric files (PDF/DOCX to structured formats)



Proprietary and Open-Source Models



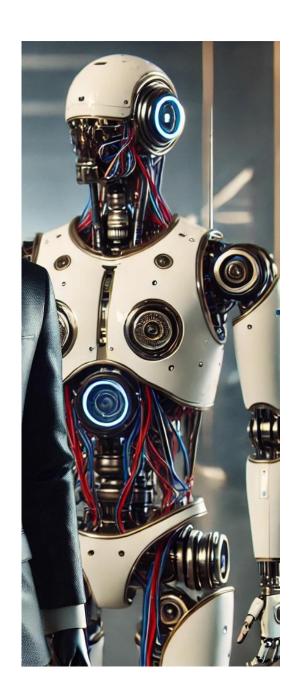
Proprietary Models

- Example: OpenAl's ChatGPT
- Easy framework for use
 - Chatbot and API capabilities
- Must pay to use chatbots and for API interactions
 - Limited free versions available
- Cannot directly access model parameters
 - Don't know exactly how the models are being implemented
 - Not just the "black-box" parts; all parts
- May make changes to established models
 - Can have downstream effects if built into a workflow



Open-source Models

- Many available at <u>huggingface.co</u>
- Does not require any direct costs
 - Hardware and software required to use the models
- Requires help building an easy interface for use
- Can directly access model parameters
 - "Black-box" (optimization / probabilistic aspects) still the same
 - Can more easily customize use of models by further training or by altering the model parameters
- Will not change unless you (or your organization) make changes



Chatbot vs. API Usage

Chatbot / Custom Chatbot

- Easier user interface; requires no programming knowledge
- Can provide feedback / continue prompting sessions
- Single cost (monthly subscription)
- Not able to easily vary model parameters (e.g., Temperature)
- Can be laborious for use at scale

API

- Can build into a program so can be used at scale
- Easy to manipulate model parameters
- Cost per use (e.g. OpenAl)
- Requires some programming knowledge
- Cannot continue prompting sessions

Use LLMs Responsibly: Data Privacy

- Data Privacy and Security
 - Never submit any personally identifiable information (PII), patient data, or student records to an AI model (including GPTs).
 - Understand GPT data retention policies some AI providers may log or store inputs/outputs.
 - Assume that all interactions with GPTs are not private unless you have checked with the appropriate IT and legal people at your institution

Use LLMs Responsibly: Compliance

Al usage policies vary by region

National

Ensure that all AI use follows all national mandates

State

 States may have additional mandates or guidelines around AI use in addition to national regulations

Local institution

- Even further, your institution or organization may have specific guidelines around Al use
- AAMC has a website specifically dedicated to these matters

Before using AI/GPTs in any operational setting, consult your legal and IT departments.

NBME MICROCREDENTIALING SHORT COURSE



Later this year NBME will begin offering free microcredentialing short courses. Be among the first to enroll in our introductory course:

"A Primer for Using Generative Artificial Intelligence in Medical Education."

Scan the QR code to stay informed and reserve your spot when registration opens!



Questions? Feedback?

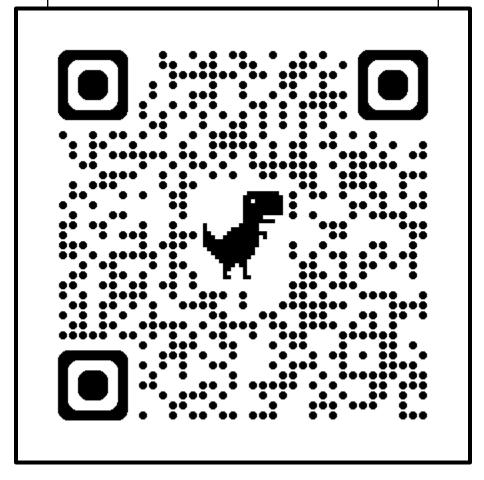
Christopher Runyon

CRunyon@nbme.org

or use this QR Code I



Presentation Slides:



(both presentations available)