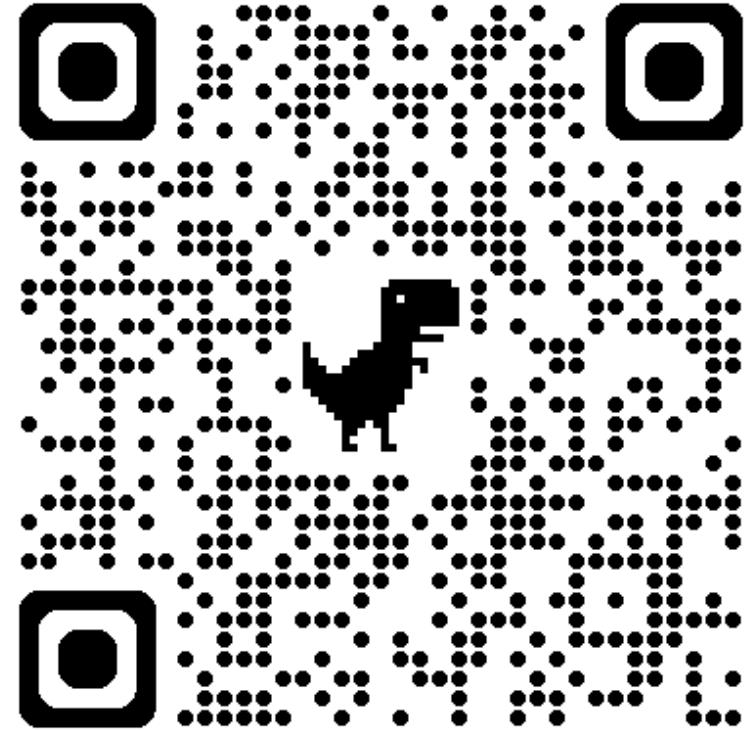




Image generated with ChatGPT



Using GPTs to Assist with Grading OSCE Notes

Christopher Runyon, PhD



I have no financial disclosure or conflicts of interest with the presented material in this presentation.

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.

Goals

Empower you with the necessary conceptual understanding to effectively use LLMs to assist with grading constructed response assessments (e.g., OSCE Notes)

We will not be doing hands-on demonstrations of using LLMs to score OSCE notes

- Materials available in Github repository to try
- Upcoming workshop (25 October) on integrating LLMs into R workflows

Learning Objectives

- Explain the foundational concepts of large language models (LLMs) to support informed use in educational contexts.
- Understand the differences between analytic and holistic rubrics and their implications for LLM-assisted grading.
- Apply effective prompt engineering techniques for assessing medical education tasks.
- Identify key considerations for implementing LLM-based grading systems.

Introduction to Generative Pretrained Transformers (GPTs)



Not just what to do,
but why to do it

“We write to advocate for dispelling the magical aura around AI in medical education... What is missing then is not more coursework on AI... By dispelling **AI's magic and showing [others] it is just several familiar concepts working in concert**, we can prepare them to understand both the AI we have now and whatever arrives in the years to come.”

McQuade CN, Wijesekera TP, Chartash D. **Dispelling the magic of artificial intelligence in medical education.** *Med Educ.* 2025;59(3):350-351.

What is a GPT?

- **G = Generative**
 - refers to the model's ability to create new content
 - will *always* produce a response
- **P = Pretrained**
 - model has undergone an initial extensive training phase on a vast corpus of text data; equips the model with a broad 'understanding' of language and its nuances
- **T = Transformer**
 - aspect of the model architecture (not discussed today)
 - attention layer (will be discussed today)

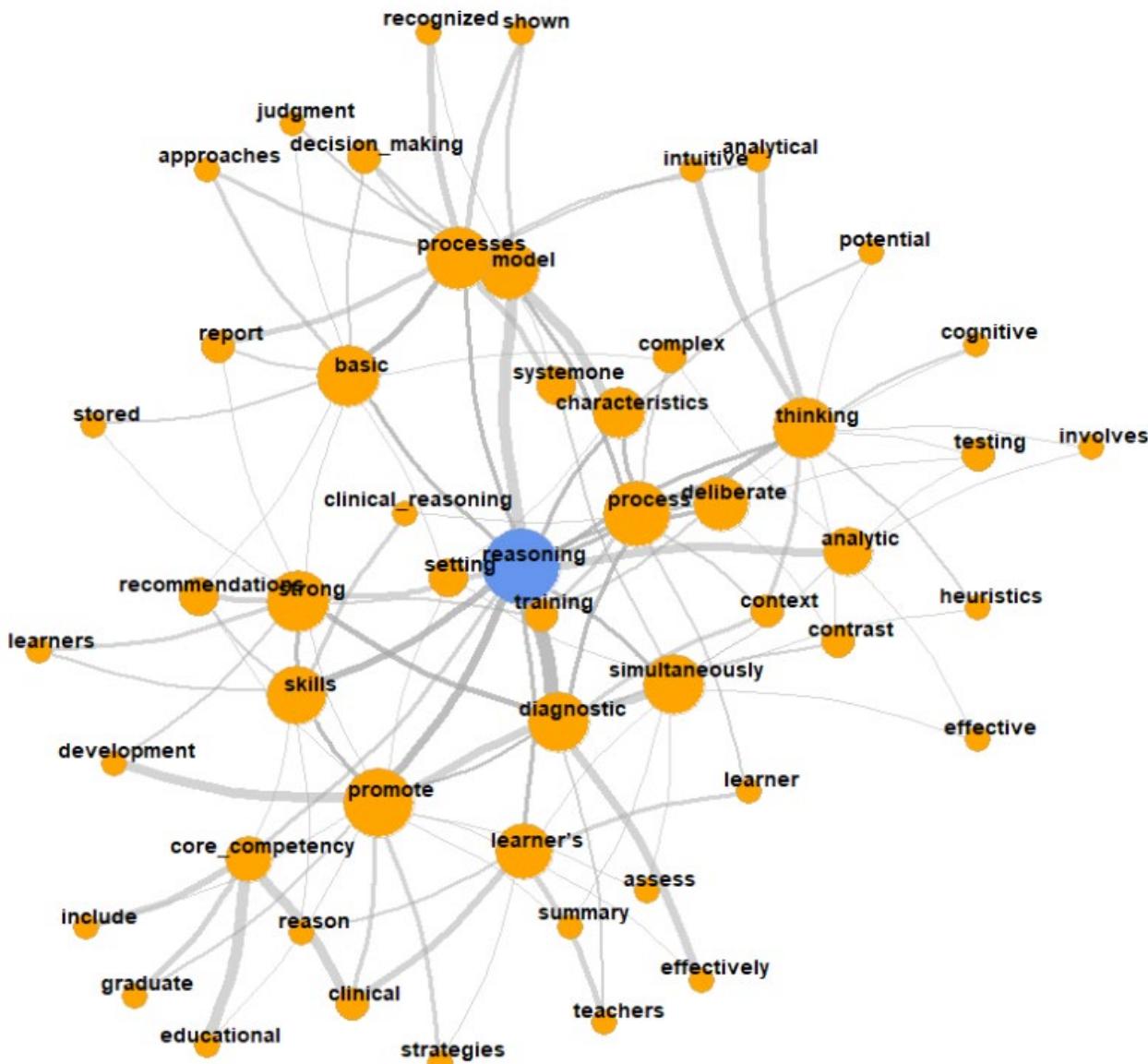
How LLMs Make Sense of Language

- 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 [REDACTED] sleeping behind the tree.”

- A very, *very* fancy text calculators
- *Always* will produce a response
 - Hallucinations = inaccuracies (all types)
- Not deterministic; variation is part of its beauty

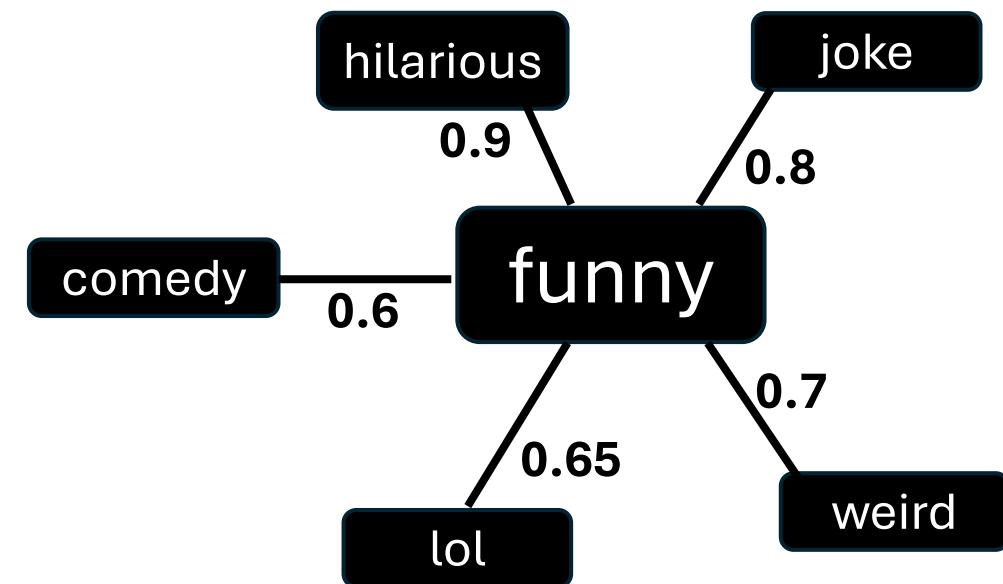




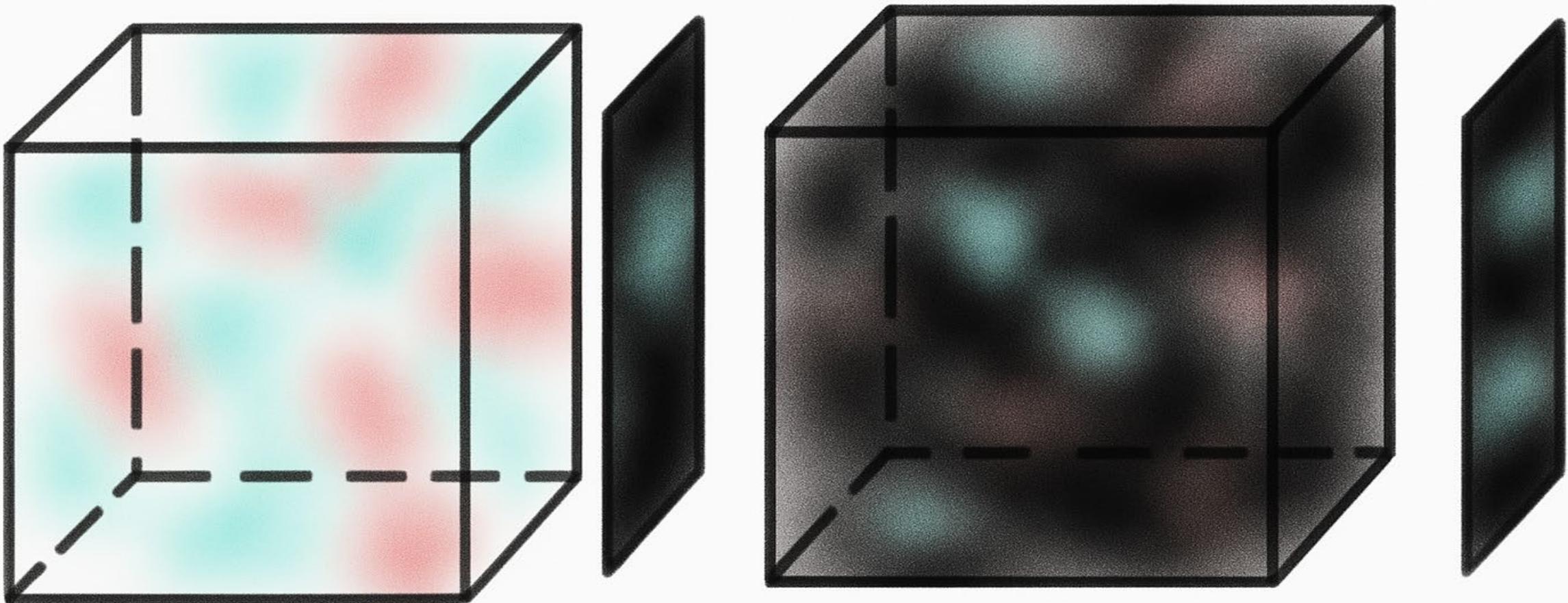
Llama: 8 – 405 Billion Parameters (varies based on model)

DeepSeek R1: 671 Billion

Semantic Similarity



Pre-trained parameters → Prompt → Updated parameters → Output

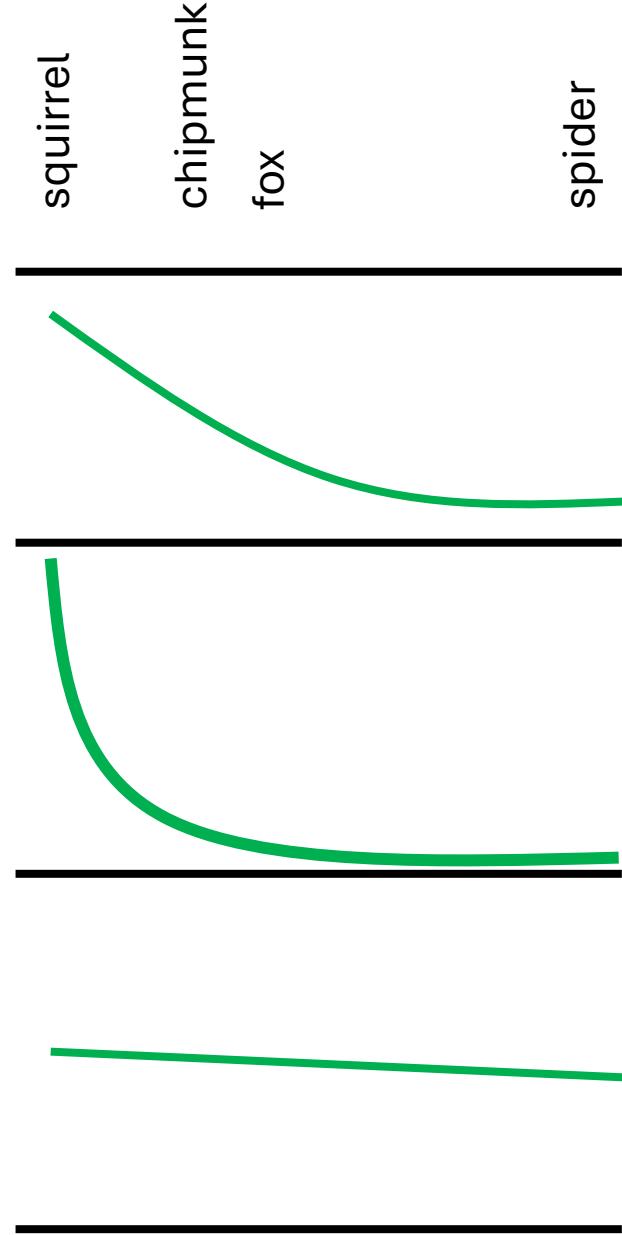


Model parameters

Temperature

- Lower values
 - Makes the probability distribution of the next word **sharper/skewed**
 - Model is **more** likely to pick the **most** likely words
- Higher values
 - Makes the probability distribution of the next word chosen **flatter**
 - Model is **more** likely to pick **less** likely words

Probability of Selection



In less than 15 words, please describe why pizza is so good

Model	ChatGPT 4o
Temperature	1 (default)

- Pizza combines savory, cheesy, and versatile toppings on a delicious, soft, or crispy crust.
- Pizza combines savory cheese, flavorful toppings, and a crispy crust for a satisfyingly delicious experience.
- Pizza's delicious due to its savory combination of cheese, sauce, and various toppings.
- Pizza combines savory flavors, melty cheese, and a satisfying crust in every bite. Irresistibly delicious!
- Pizza's delicious due to its perfect blend of cheese, sauce, and various toppings.

In less than 15 words, please describe why pizza is so good

Model	ChatGPT 4o
Temperature	1 (default)

- Pizza combines savory, cheesy, and versatile toppings on a delicious, soft, or crispy crust.
- Pizza combines savory cheese, flavorful toppings, and a crispy crust for a satisfyingly delicious experience.
- Pizza's delicious due to its savory combination of cheese, sauce, and various toppings.
- Pizza combines savory flavors, melty cheese, and a satisfying crust in every bite. Irresistibly delicious!
- Pizza's delicious due to its perfect blend of cheese, sauce, and various toppings.

In less than 15 words, please describe why pizza is so good

- 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.”

Model	ChatGPT 4o
Temperature	0 (lowest)
Replications	100

In less than 15 words, please describe why pizza is so good

Model	ChatGPT 4o
Temperature	2 (highest)

- Delicious, customizable, comforting dish suitable for any occasion; loved globally for its diverse flavors.
- Pizza combines diverse flavors and textures, offering endless customization to satisfy any craving.
- Deliciously versatile, universally loved, customizable flavor combinations, comforting, and fun to share.
- Versatile ingredients, customizable toppings, perfect balance of flavors, and universally loved comfort food.
- Delicious flavors, endless toppings, crispy crust, shareable, comfort food, universally loved, quick meal option.

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



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 [REDACTED] 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.

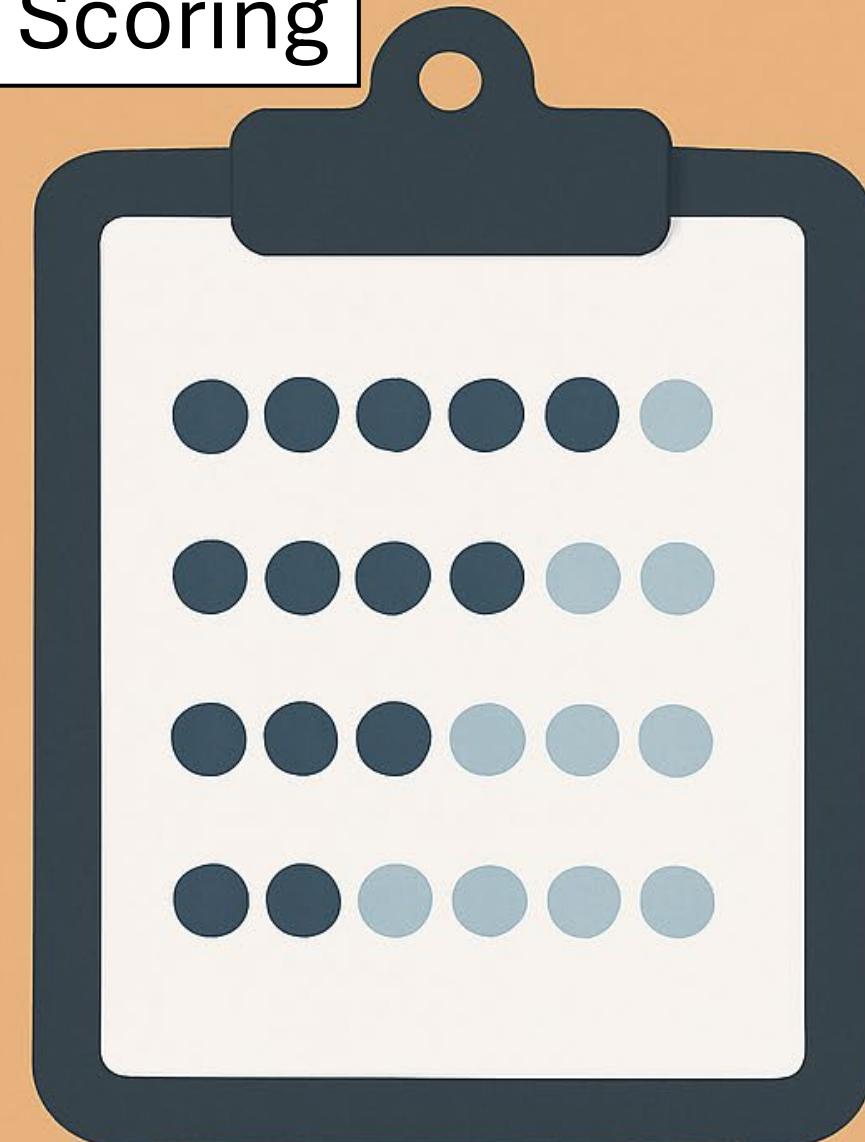
Attention

- The attention can also be impacted by the length of the conversation you have (when using the chatbot versions of Gen AI Models)
- May start the conversation well, but “forget” important parts later
- Alternatively, the Gen AI model can also fixate (over-attend) to less important details
- Solution: Provide the model feedback! (when using chatbot)

Using Rubrics for Scoring



Analytic



Holistic

Scoring Rubrics

- **Ensure Consistency:** same standard of assessment is applied to each student's work
- **Streamline Grading:** criteria for performance evaluations of work are made a priori
- **Support Feedback:** feedback can be linked to rubric elements
- **Instructional Alignment:** can ensure criteria are aligned with learning objectives



Scoring Rubrics

Analytic Rubrics

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 ⁷



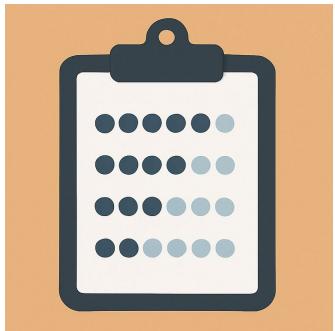
Scoring Rubrics

Analytic Rubrics

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



Scoring Rubrics



Analytic Rubrics

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

Holistic Rubrics

ASSESSMENT of REASONING TOOL

Learner: _____

Evaluator: _____



SOCIETY to
IMPROVE
DIAGNOSIS in
MEDICINE

Did the Learner...	Assessment		
	Minimal	Partial	Complete
Collect/report history and examination data in a hypothesis-directed manner?	<ul style="list-style-type: none"> Non-directed in questioning and exam Asked questions without clear focus on potential diagnoses 	<ul style="list-style-type: none"> Questioning and exam generally reflective of potential diagnoses, but some less relevant or tangential questions 	<ul style="list-style-type: none"> 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?	<ul style="list-style-type: none"> Included extraneous information Missed key findings Did not translate findings into medical terminology 	<ul style="list-style-type: none"> Generally included key clinical findings (both positive and negative) but either missed some key findings or missed important descriptive medical terminology 	<ul style="list-style-type: none"> 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?	<ul style="list-style-type: none"> Missed key elements of differential diagnosis, including likely diagnoses or "can't miss" diagnoses 	<ul style="list-style-type: none"> Gave differential diagnosis that included likely and "can't miss" diagnoses but either missed key diagnoses or ranked them inappropriately 	<ul style="list-style-type: none"> Gave accurately ranked differential diagnosis including likely and "can't miss" diagnoses
Direct evaluation/treatment towards high priority diagnoses?	<ul style="list-style-type: none"> Directed evaluation and treatment toward unlikely/unimportant diagnoses Did not evaluate or treat for most likely/"can't miss" diagnoses 	<ul style="list-style-type: none"> Major focus of evaluation and treatment was likely and "can't miss" diagnoses but included non-essential testing 	<ul style="list-style-type: none"> 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?	<ul style="list-style-type: none"> Not able to describe the influence of cognitive tendencies or emotional/situational factors that may have influenced decision-making 	<ul style="list-style-type: none"> Can name one cognitive tendency or emotional/situational factor that may have influenced decision-making 	
OVERALL ASSESSMENT	NEEDS IMPROVEMENT <input type="checkbox"/>	MEETS COMPETENCY <input type="checkbox"/>	EXCELLENCE <input type="checkbox"/>

Comments:

Thammasitboon S, Rencic JJ, Trowbridge RL, Olson APJ, Sur M, Dhaliwal G. The Assessment of Reasoning Tool (ART): structuring the conversation between teachers and learners. *Diagnosis*. 2018;5(4):197-203.

ASSESSMENT of REASONING TOOL

Learner: _____

Evaluator: _____



SOCIETY to
IMPROVE
DIAGNOSIS in
MEDICINE

Did the Learner...	Assessment		
	Minimal	Partial	Complete
Collect/report history and examination data in a hypothesis-directed manner?	<ul style="list-style-type: none">Non-directed in questioning and examAsked questions without clear focus on potential diagnoses	<ul style="list-style-type: none">Questioning and exam generally reflective of potential diagnoses, but some less relevant or tangential questions	<ul style="list-style-type: none">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?	<ul style="list-style-type: none">Included extraneous informationMissed key findingsDid not translate findings into medical terminology	<ul style="list-style-type: none">Generally included key clinical findings (both positive and negative) but either missed some key findings or missed important descriptive medical terminology	<ul style="list-style-type: none">Gave clear synopsis of clinical problemEmphasized important positive and negative findings using descriptive medical terminology

Thammasitboon S, Rencic JJ, Trowbridge RL, Olson APJ, Sur M, Dhaliwal G. The Assessment of Reasoning Tool (ART): structuring the conversation between teachers and learners. *Diagnosis*. 2018;5(4):197-203.



Scoring Rubrics



Analytic Rubrics

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

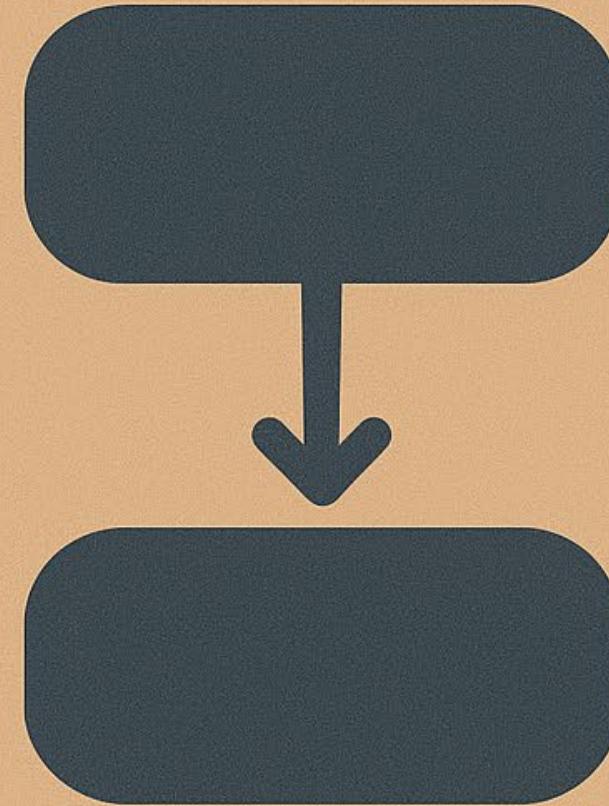
Holistic Rubrics

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

Prompt Engineering

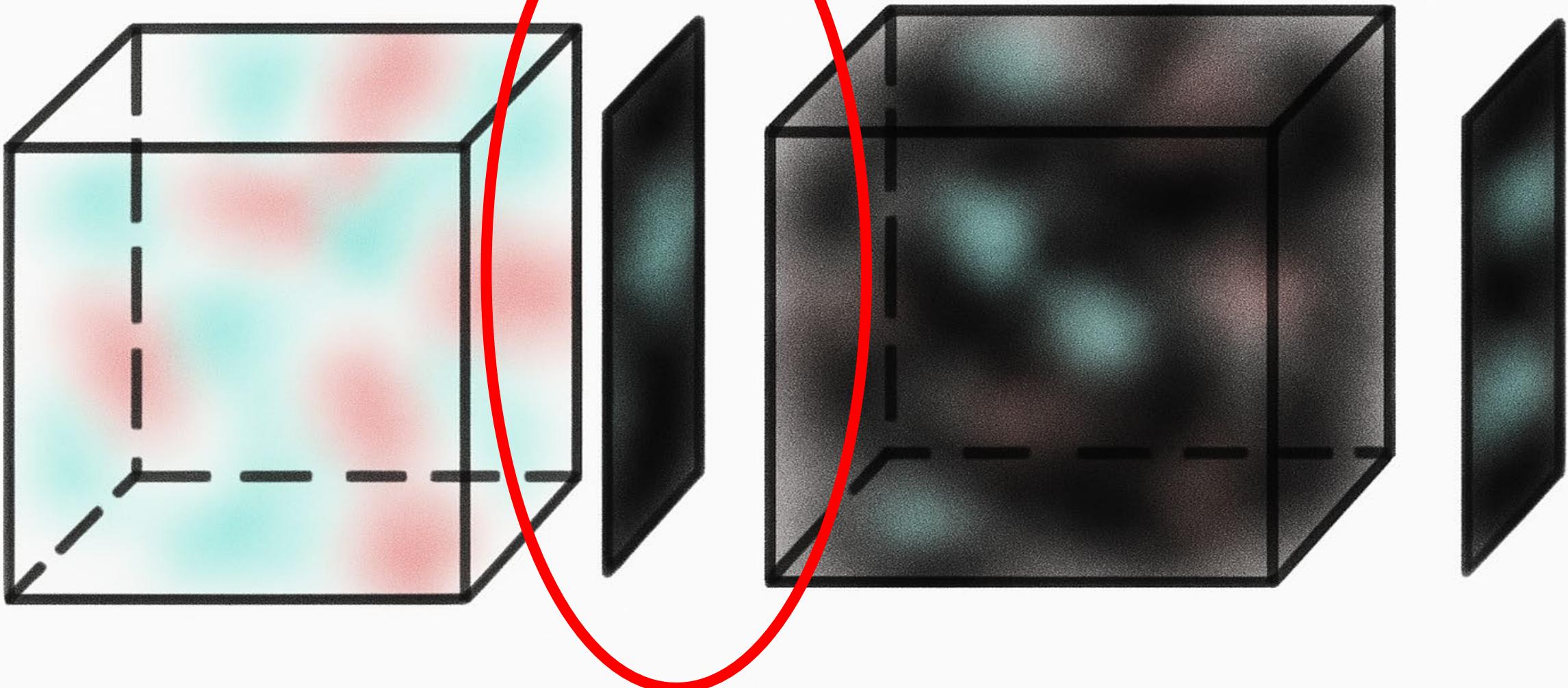


Formulas



Chaining

Pre-trained parameters → Prompt → Updated parameters → Output



“Degrees of Freedom” in LLM Prompts

- When prompting an LLM, any detail you don't explicitly specify 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
ADJECTIVE
with your _____ family. Usually you go to some place
ADJECTIVE
that is near a/an _____ or up on a/an _____.
NOUN NOUN

A good vacation place is one where you can ride _____
PLURAL NOUN
or play _____ or go hunting for _____. I like
GAME PLURAL NOUN
to spend my time _____ or _____.
VERB ENDING IN "ING" VERB ENDING IN "ING"

Activity : “Degrees of Freedom”

- 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?

Caveat

- Even if you know exactly what you want the model to do, try submitting the same prompt with fewer constraints.
- Highly-specified prompts will constrain the model's creativity, and you could miss something interesting.

“Test counterintuitive things because no one else will”

- Rory Sutherland

Prompt Engineering

- There are ways to interact with LLMs that can result in more useful responses
- Two general strategies we'll discuss today:
 - Prompt formulas
 - Prompt Chaining

Goals of Prompt Engineering:

1. Help activate more relevant connections in the network
2. Appropriately leverage the attention mechanism of the model

“Degrees of Freedom” in LLM Prompts

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

Prompt formulas

- Structured guidelines or templates used to generate responses from generative language models
 - 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 Refinement w LLMs

- 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.”

Prompt Refinement w LLMs

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].”

Chaining

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

- Happens automatically in newer “reasoning” models
- Invisible Instructions

Anthropic Research System



Screenshot of Anthropic's Chatbot
Captured June 30, 2025

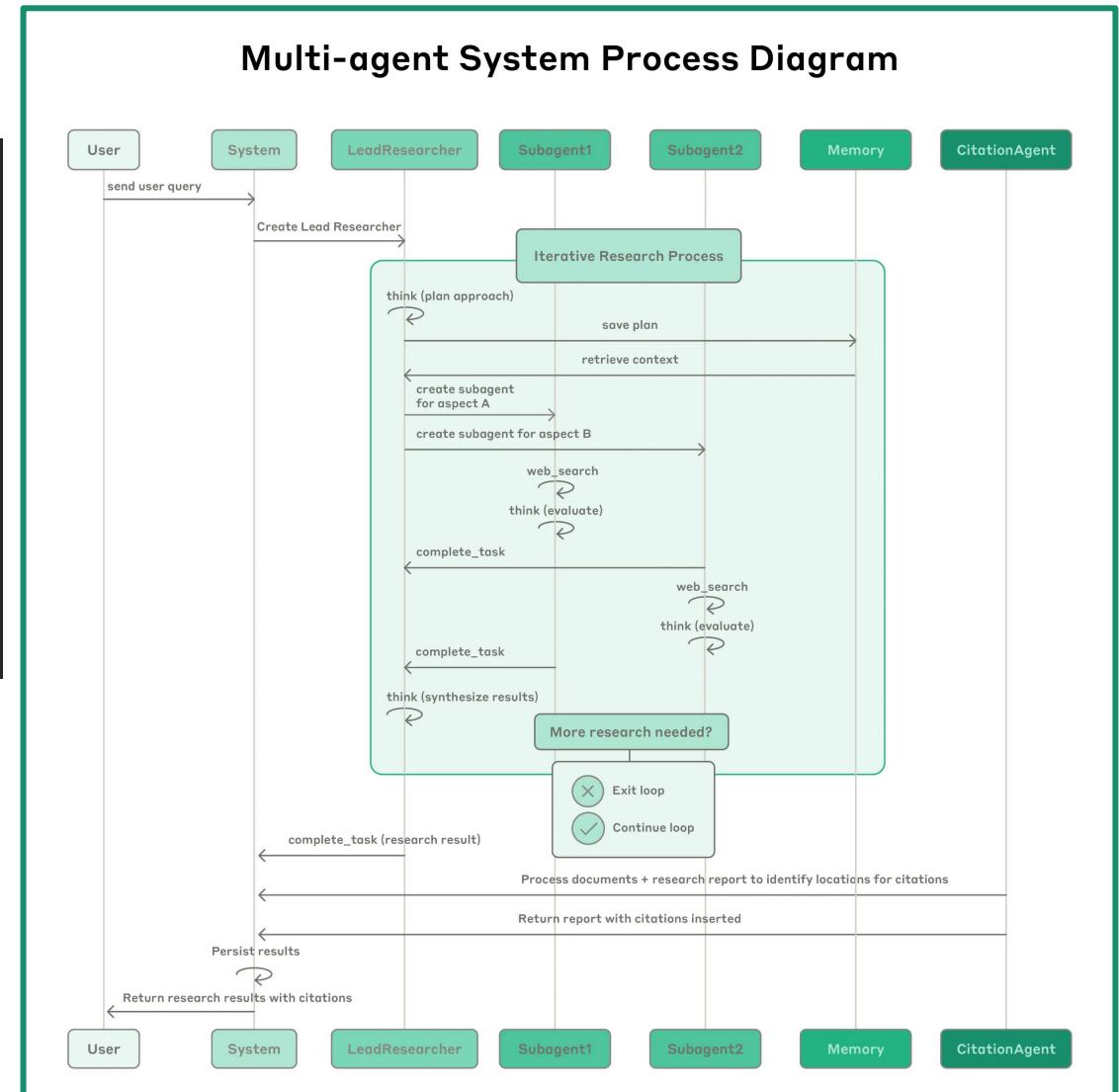


Image from
<https://www.anthropic.com/engineering/built-multi-agent-research-system>

Chaining

Chain-of-Thought : Encourages GenAI model to articulate its reasoning process in steps

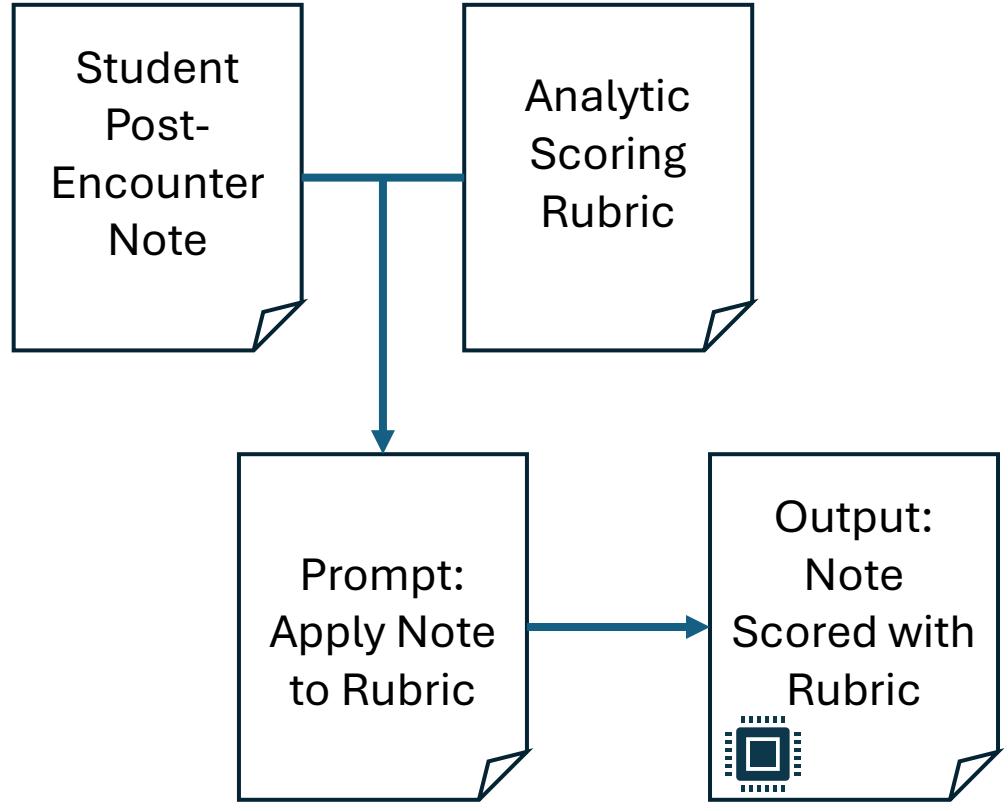
- Happens automatically in newer “reasoning” models
- Invisible Instructions

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

- User creates the invisible instructions

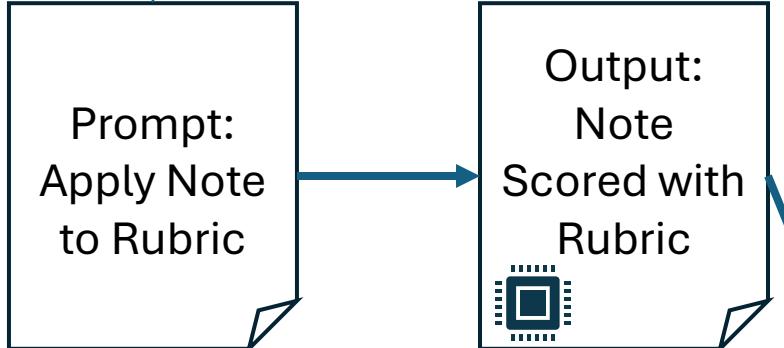
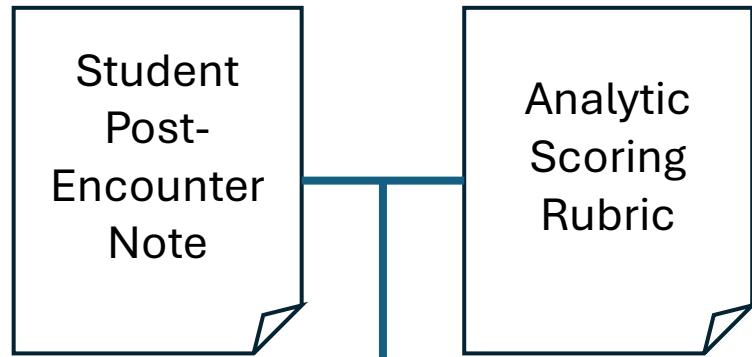
Prompt Chaining

- LLMs aren't just powerful in isolation - they become transformative when woven throughout your workflow
- 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



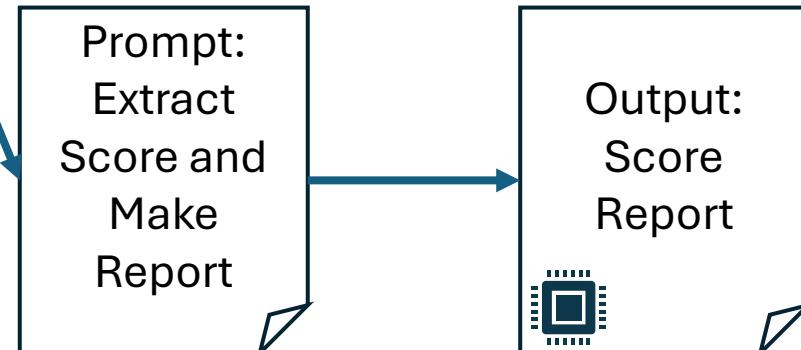
= LLM output

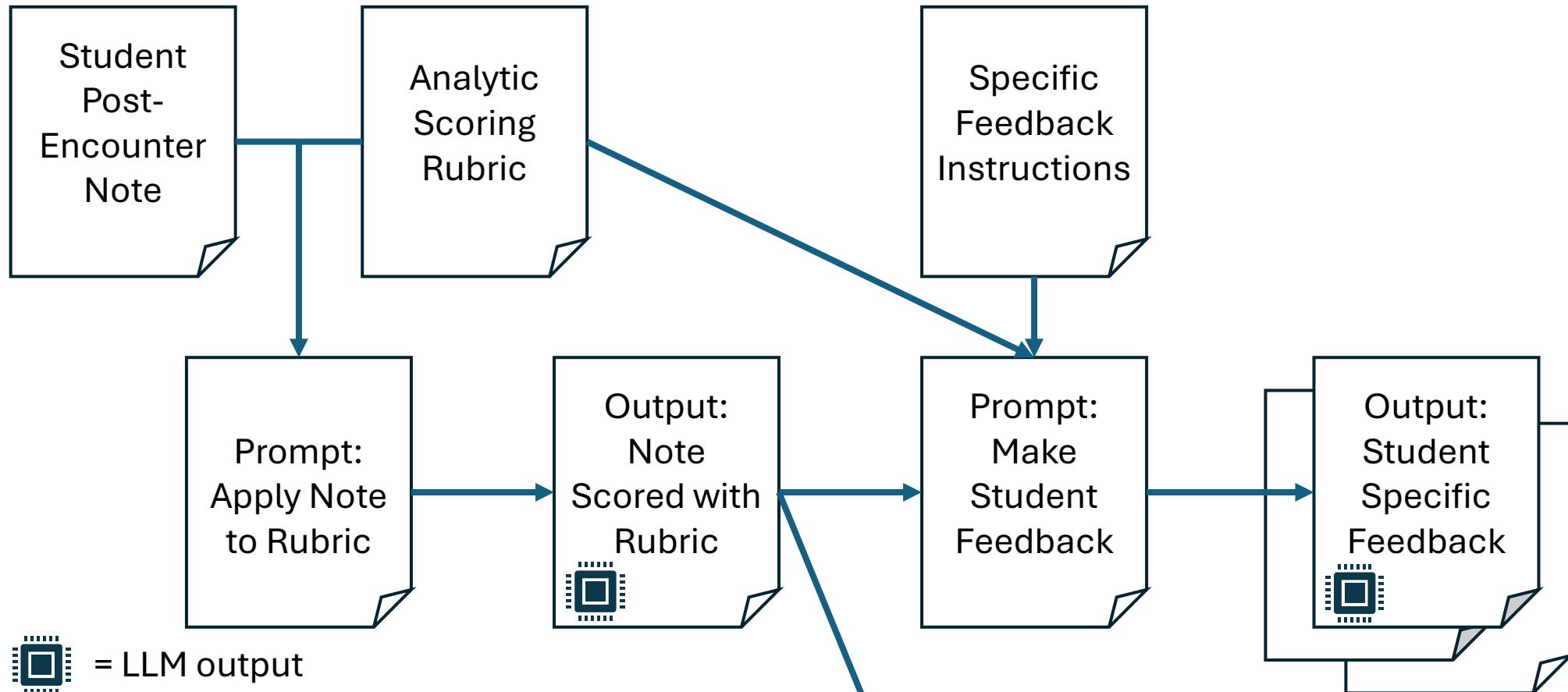
Prompt Chaining: An Example with OSCE Note Grading



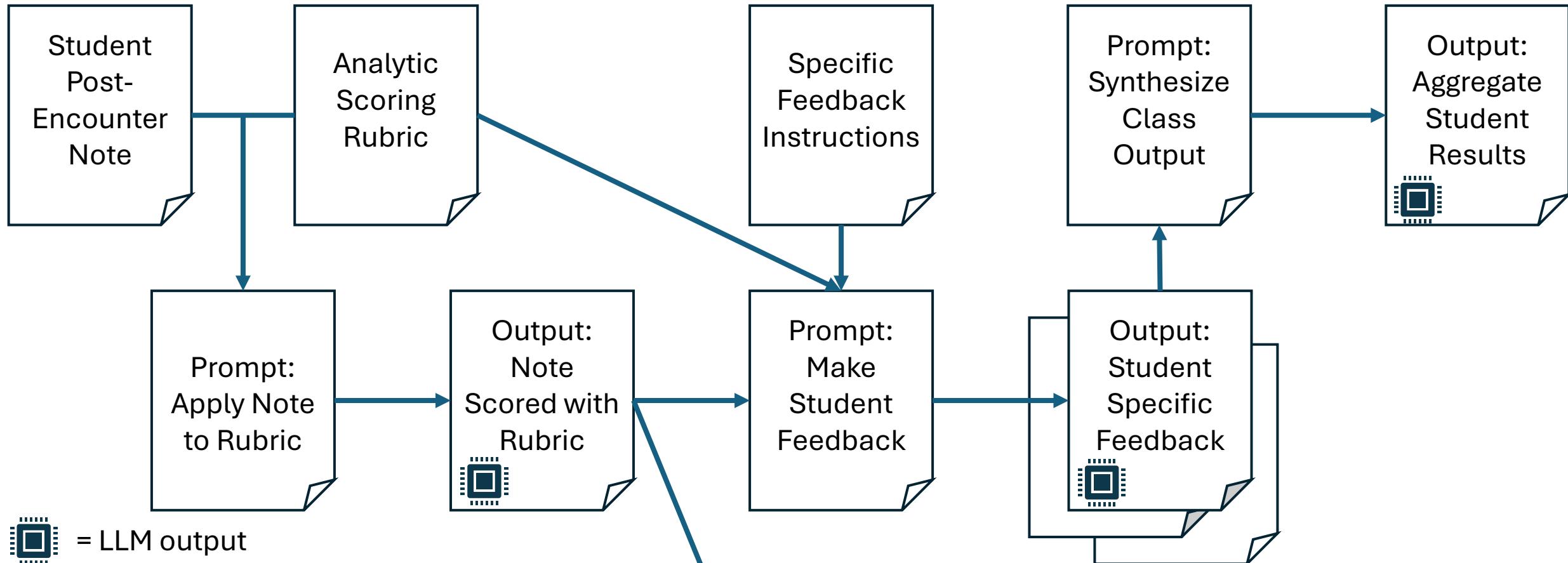
= LLM output

Prompt Chaining: An Example with OSCE Note Grading

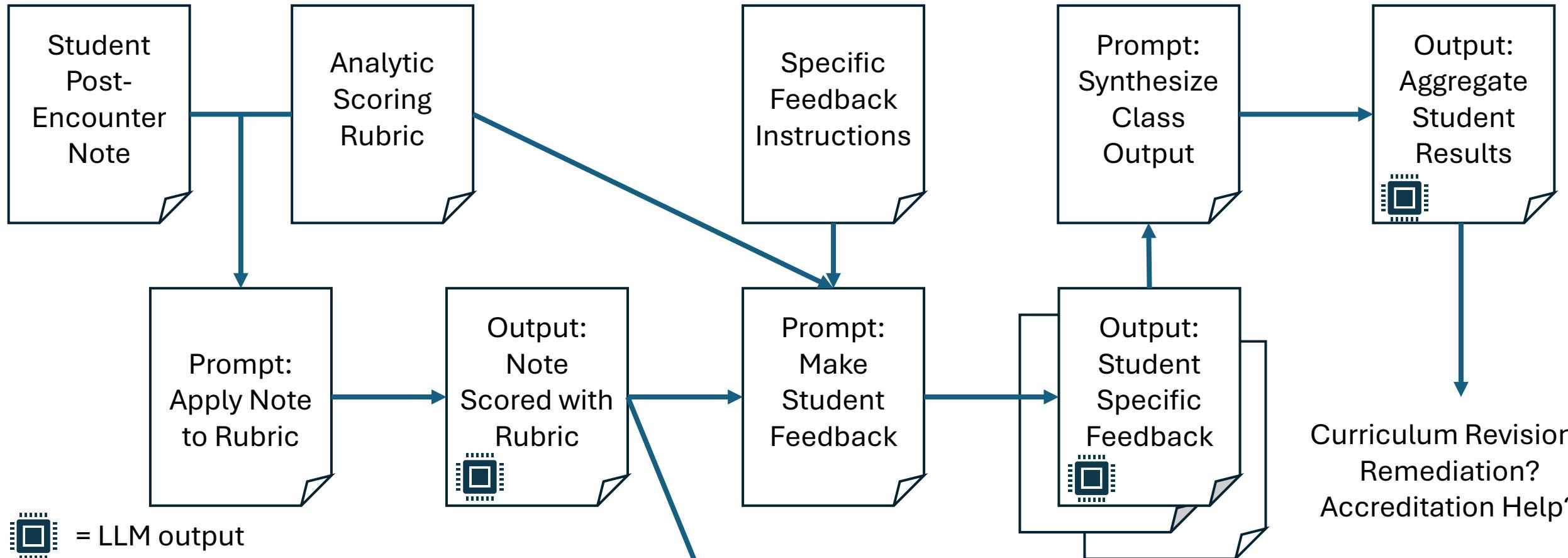




Prompt Chaining: An Example with OSCE Note Grading



Prompt Chaining: An Example with OSCE Note Grading



Prompt Chaining: An Example with OSCE Note Grading

Prompt Chaining

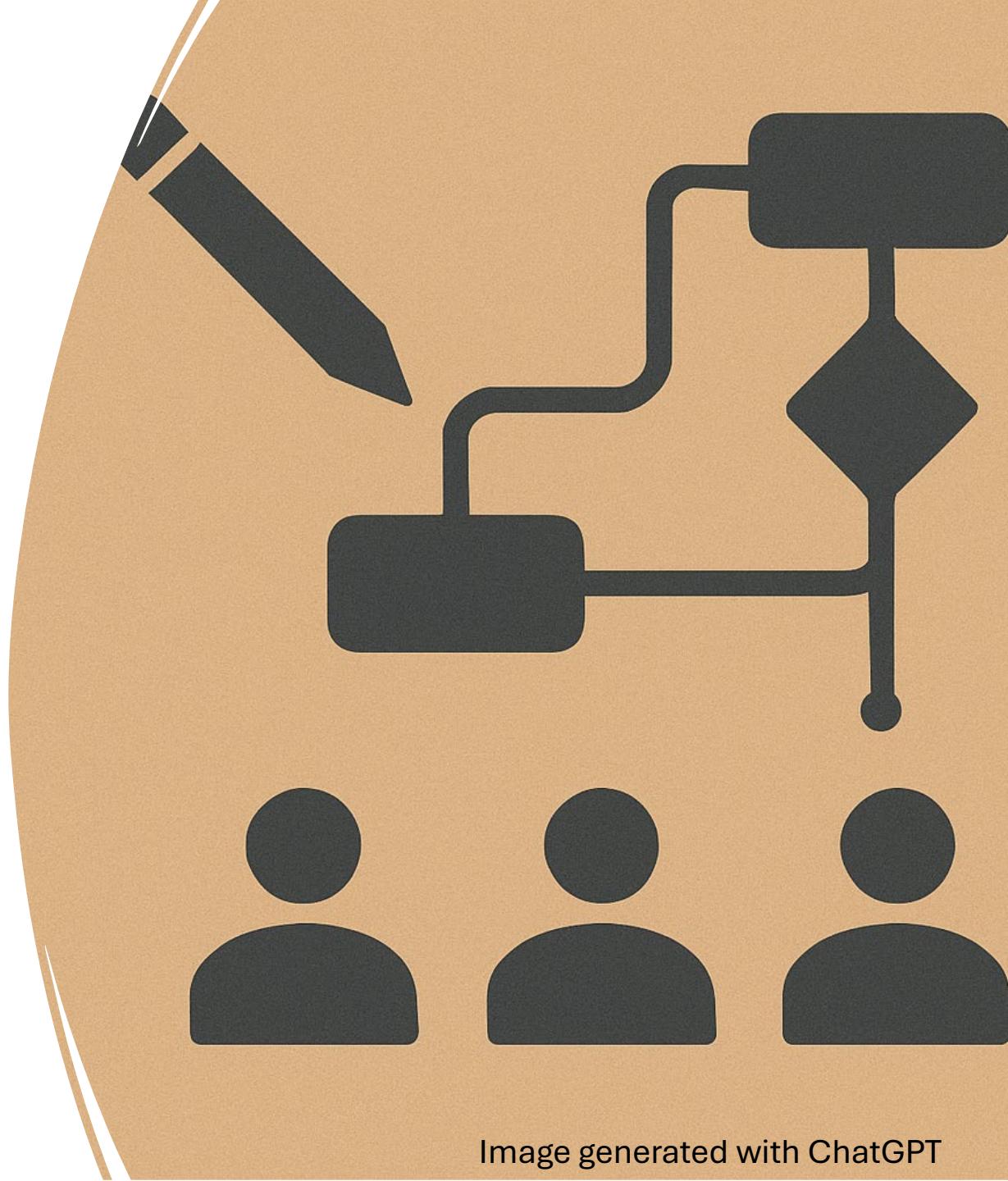
LLMs aren't just powerful in isolation - they become transformative when woven throughout your workflow

Activity: Prompt Chaining Workflows

In what ways can you envision using a multi-prompt system to leverage LLM scoring?

- Must include at least 3 steps
- What checks can you build into the system?

(Similar to multi-agentic system)



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 consistently and accurately represents the examinee's performance on the specific construct being measured.

Mitigating Variability in LLM Responses

- Model-based strategies
 - Controlling temperature, top_p, length of output, etc.
- Prompt-based strategies
 - Providing example of output structure
 - Using prompt chaining / short prompts
- Materials-based strategies
 - Refining input materials
 - Refining instructions (could be part of prompt-based strategies)

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

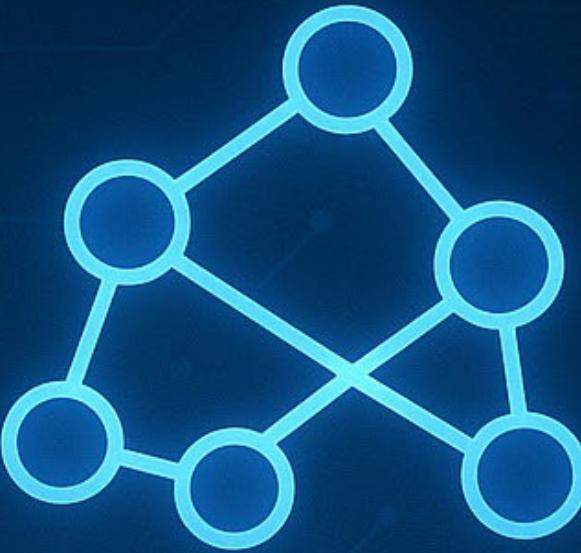
- Are LLM scores consistent and reliable?
- How does the LLM arriving at a specific score? (Transparency)
- How has the LLM's scoring system been tested to ensure it works as intended?
- Is the scoring aligned with the intended learning objectives and outcomes?
- Does the scoring method fully capture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

Testing LLM Graders

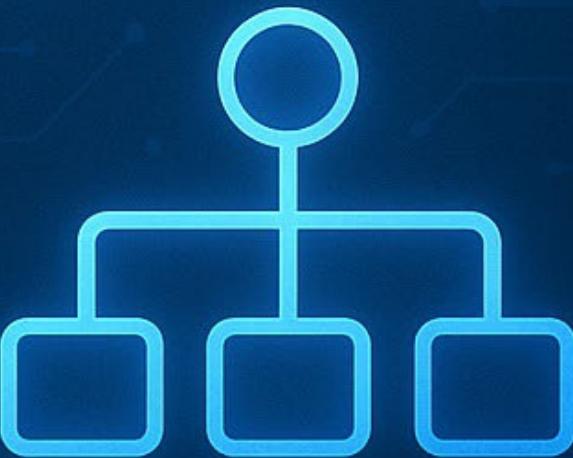
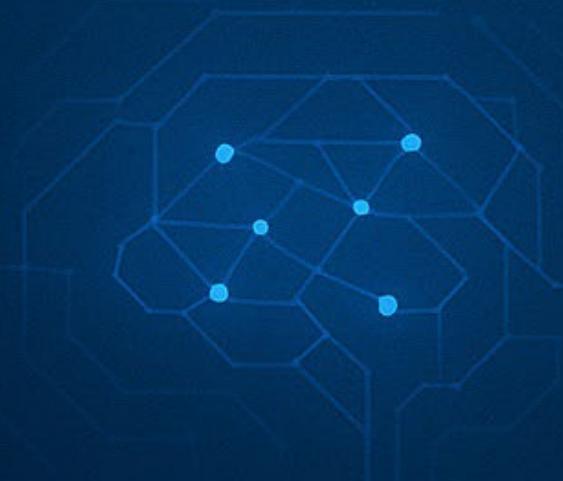
- Have at least 2 human graders to compare to LLM scores on a set of sample assessments
 - Compare interrater reliability between humans, each human and the LLM (Human 1 & Human 2; Human 1 & LLM; Human 2 & LLM)
 - Goal: have LLM score as reliably and accurately as any 2 humans
 - Just as good (or better)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

Summary of LLM Lessons

- Good at detecting lexical variants
- Needs extensive instructions to understand context
- Examples can help improve LLM performance
- Have limited attention; can only complete a small number of tasks well in a single prompt
- Prompt chaining can help improve transparency and isolate errors
- Variations in responses are expected, but can be mitigated
- The same note should be scored multiple times with the same LLM setup to learn about variability in response



IMPLICATIONS



Implications for Using Analytic Rubrics

- LLM architecture is well-suited to identify lexical variants (different ways of saying conceptually similar things)
- Need to test how many elements the model can accurately find in a single prompt; use prompt chaining if necessary
- May need to implement an extra step to ask the LLM to perform math; “you are a calculator and your job is to sum these numbers together”

Implications for Using Holistic Rubrics

- Need to have separate agents (prompts) for each of the holistic elements
- Having clear examples of the different score levels can help
 - Even better = *why* this example receives the score it does vs. others
- May still not recognize context; deep expertise is (currently) the LLM Achillies heel
 - E.g. pertinent negatives, missing information, pneumonics

(oh my!)

Chatbots, Custom Chatbots, and APIs

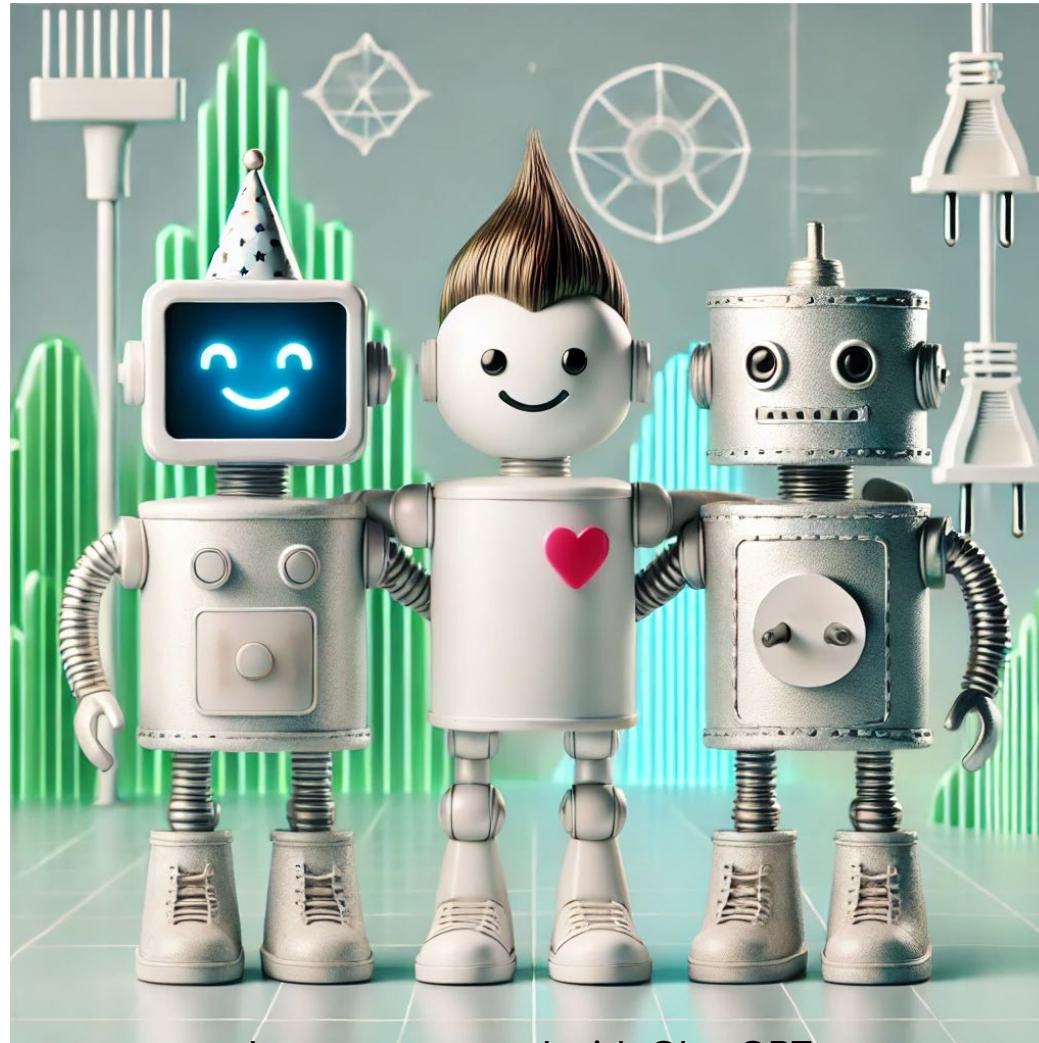


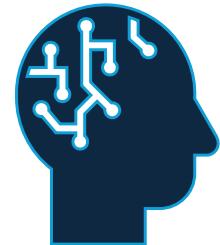
Image generated with ChatGPT

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



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.

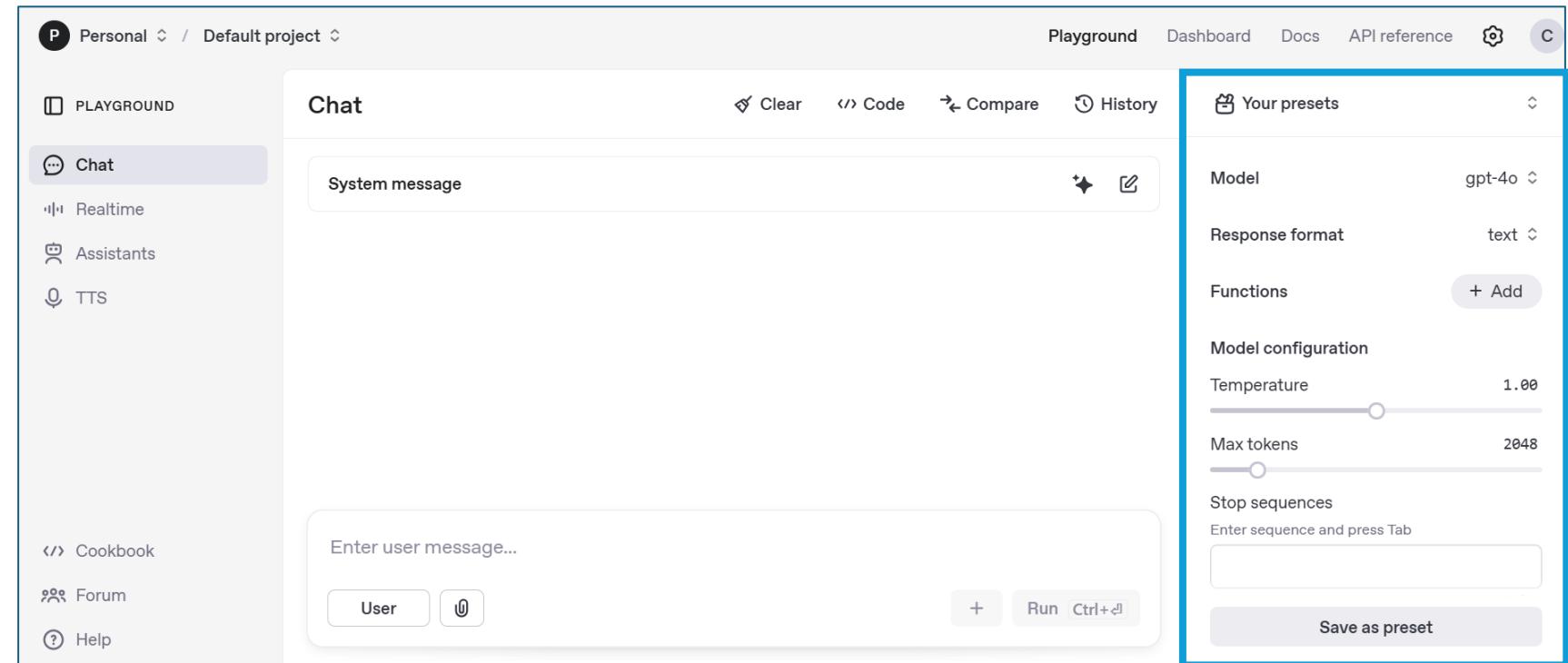
Sure! Here are a few great spots to try local Milwaukee cuisine:

- **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 traditional Milwaukee fish fry.



LLM Chatbots

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



Custom Chatbots



- 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)
 - Developing standardized instructions for creating rubrics for OSCEs / assessments

Custom Chatbots

New GPT
Draft

... Create

Create Configure Preview

Hi! I'll help you build a new GPT. You can say something like, "make a creative who helps generate visuals for new products" or "make a software engineer who helps format my code."

What would you like to make?

Ask anything

+ Search

Ask anything

?

?

This image shows a user interface for creating a custom AI model, specifically a GPT. The top navigation bar includes a back arrow, a circular icon labeled 'New GPT Draft', three dots for more options, and a prominent 'Create' button. Below the navigation, there are two tabs: 'Create' (which is selected) and 'Configure'. The main content area is divided into two sections: 'Create' on the left and 'Preview' on the right. The 'Create' section features a text input field containing a placeholder message about generating specific types of AI models. Below this is a question 'What would you like to make?'. At the bottom of this section are buttons for '+', 'Search', and a question mark. The 'Preview' section shows a simplified interface with an 'Ask anything' input field and its own 'Search' button. Both sections include a large question mark icon at the bottom right.

Custom Chatbots

 AMA Cite
Live · Anyone with a link

Last edited Mar 3 Updates pending

...

Share

Update

Create

Configure

Name

AMA Cite

Description

Librarian for formatting references to AMA style

Instructions

AMA Cite is a pleasant and quirky librarian with a tone reminiscent of a poet, hippie, hipster, or Jeff Goldblum. It excels in AMA citation formatting, ensuring every reference created is presented in a plaintext box for easy copying. It also searches for and recommends only published, peer-reviewed papers or notably cited blog posts, always including or offering to create an AMA citation with each interaction. This GPT handles requests with unique phrases and a conversational flair, making the experience both

Preview

AMA Cite

Librarian for formatting references to AMA style

Please convert
this citation into
AMA format.

How do I cite a
book in AMA style?

Can you format
this journal article

Help me correct
this AMA citation.

Ask anything

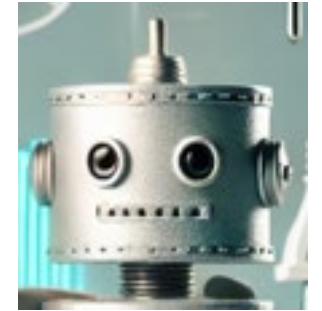


Conversation starters

Please convert this citation into AMA format.



Accessing LLMs via API



- Requires programming knowledge (e.g., Python, R, etc.)
- Can be built 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

```
import openai

# Initialize client
client = openai.OpenAI(api_key="your-api-key") ← API Key = specific to a user (or project)

# Send request
response = client.chat.completions.create(
    model="gpt-4",
    messages=[
        {"role": "system", "content": "You are a medical education assistant."},
        {"role": "user", "content": "Explain hypertension in simple terms."}
    ],
    max_tokens=150
)

# Get response
answer = response.choices[0].message.content
print(answer)
```

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>

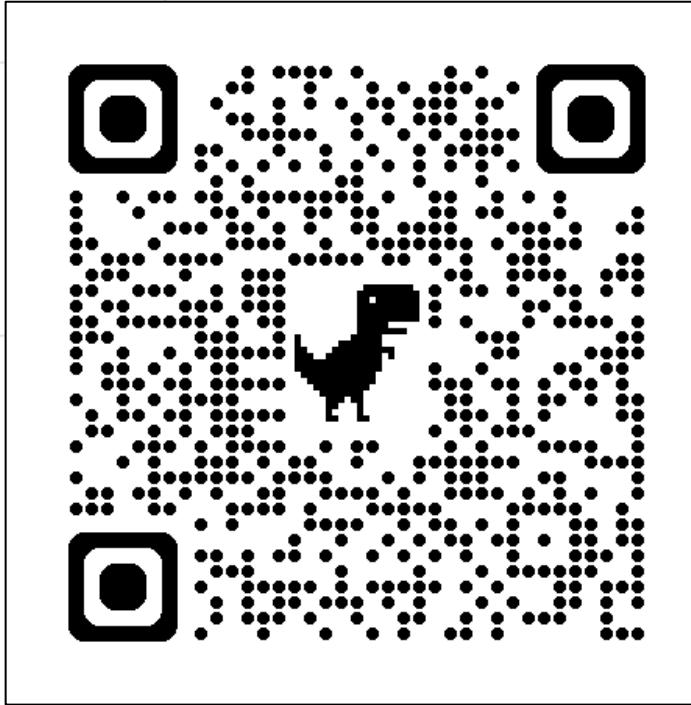
README MIT license

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)



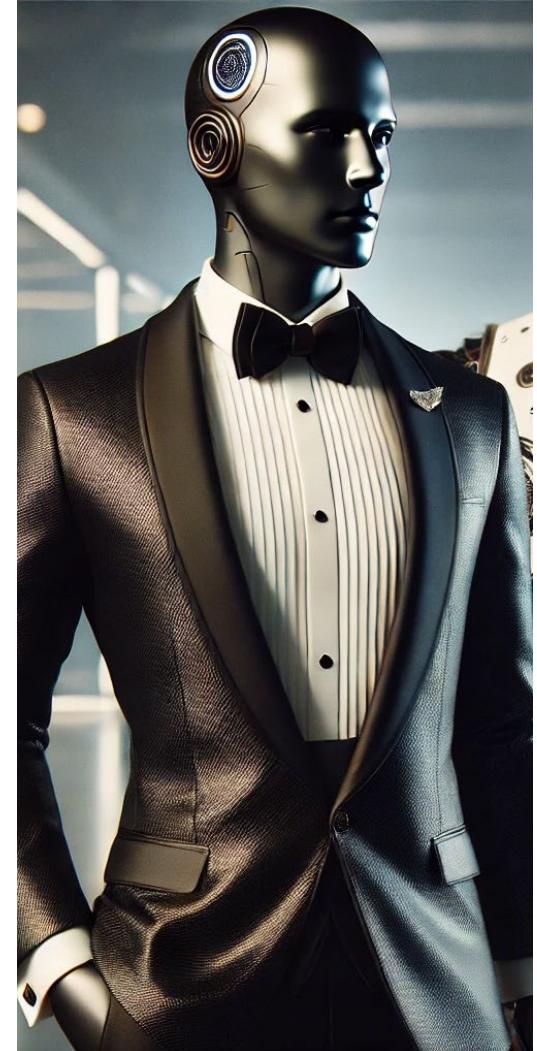
Proprietary and Open-Source Models



Image generated with ChatGPT

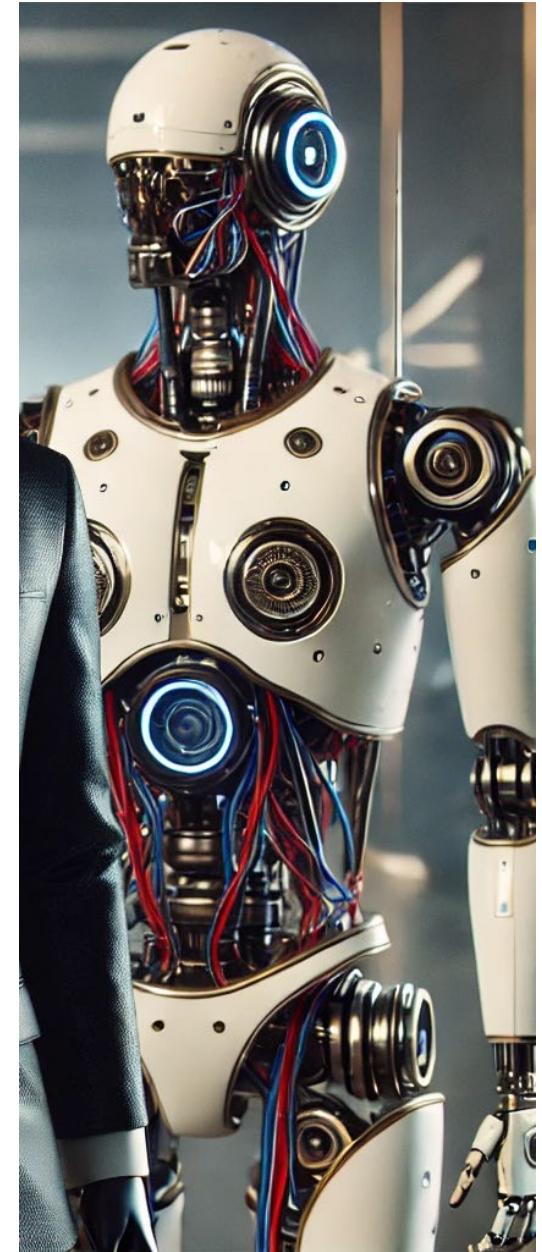
Proprietary Models

- Example: [Anthropic's Claude](#) or [OpenAI'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 huggingface.co; new Llama models
- Does not require any direct costs
 - Hardware and software required to use the models
- Requires 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 trained 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. OpenAI](#))
- 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

AI 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 AI 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 online asynchronous 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!





AMEE 2025
BARCELONA
HOW ARE EDUCATORS RELEVANT TO HEALTH?

Thank you for joining!
Your feedback is
important to us.



<https://www.surveymonkey.co.uk/r/LW-4T>