

A Primer on Using Generative AI in Medical Education

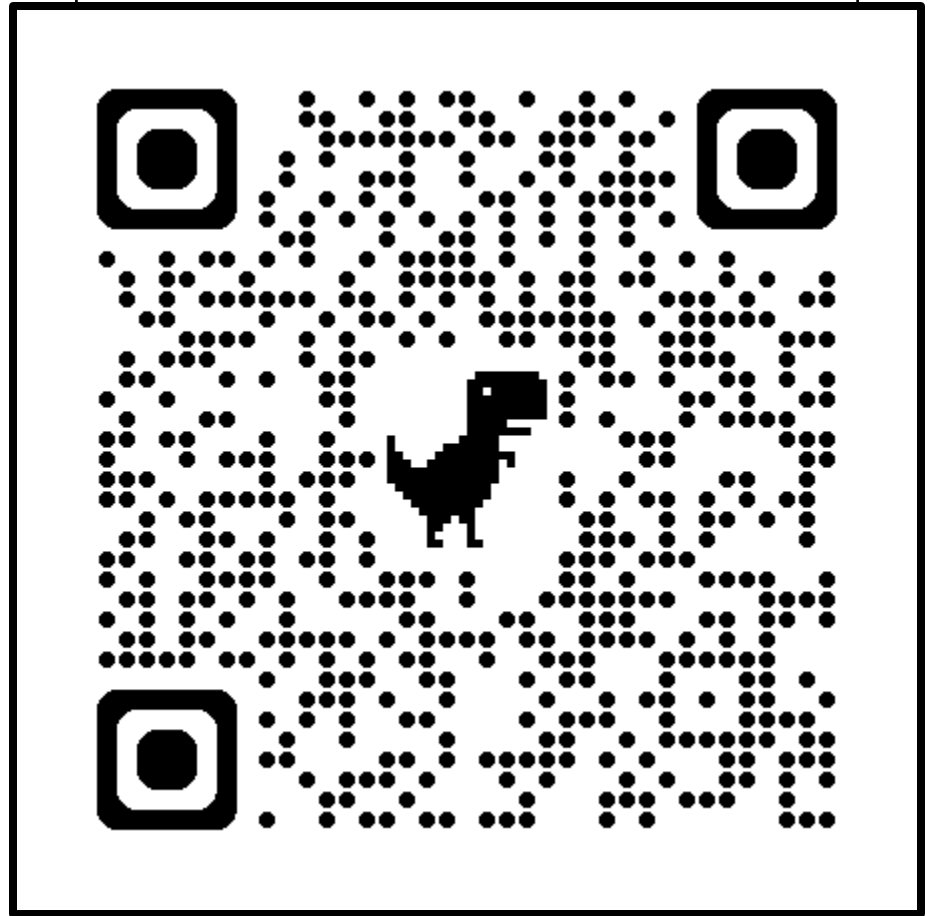
Christopher Runyon

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

- Explain the fundamental architecture of Generative AI models
- Describe the advantages and limitations of using Gen AI in various applications within medical education
- Design and evaluate effective prompts that leverage Gen AI strengths for creating educational content and assessments
- Demonstrate the ability to integrate Gen AI into existing instructional strategies to improve the quality and efficiency of medical education

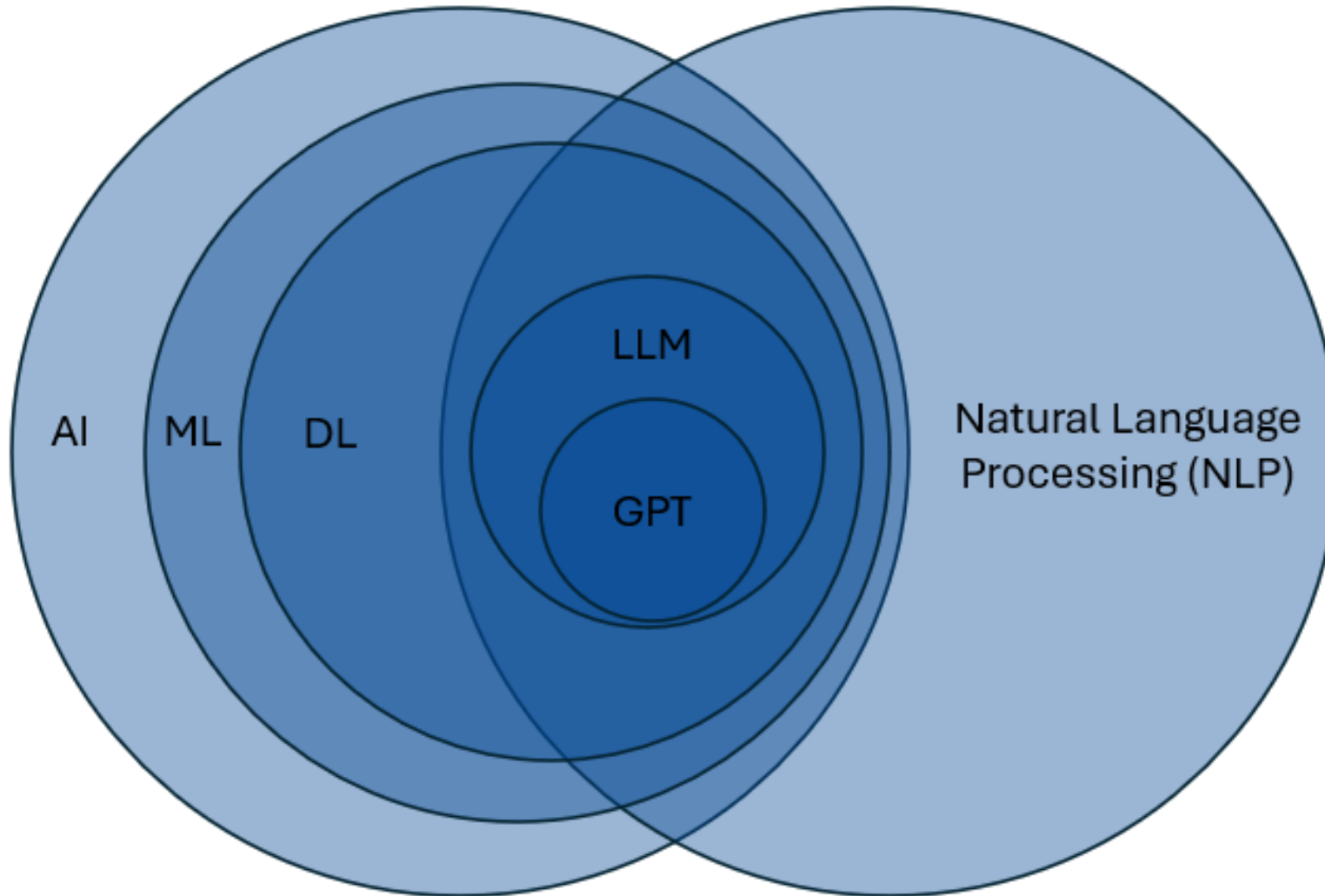
Why learn technical aspects of LLMs?

- LLMs vs Microwaves

Microwaves	LLMs
Been around for decades	Recently developed
Technology (largely) unchanged	Technology is constantly evolving
Best practices are well-known	Best practices still being developed

- Helps with understanding why prompt engineering is important and how you may need to adjust your prompts.
- Helpful for independent problem solving

What are Gen AI models?



AI = Artificial Intelligence

ML = Machine Learning

DL = Deep Learning

LLM = Large Language Models

GPT = Generative Pretrained Transformers

Variations in Generative AI Models

- [OpenAI's ChatGPT](#) (several models)
- [Anthropic's Claude](#)
- [Google's Gemini](#)
- [DeepSeek](#)
- Meta's [Llama](#) (Llama 4 models just released!)
- ...

Many freely-available models available through huggingface.co

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.”

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

Harris ZS. Distributional structure. *Word*. 1954;10(2-3):146-162.

Wampimuk sentence from: <https://aclanthology.org/P14-1132>

Image generated with ChatGPT 4o



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.

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 that are possible
 - 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 / limiting the sides that are possible

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

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.”

Variability in LLM Responses

Higher temperatures (more creative)

- Variation in 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  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
 - 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!

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

Pre-trained parameters



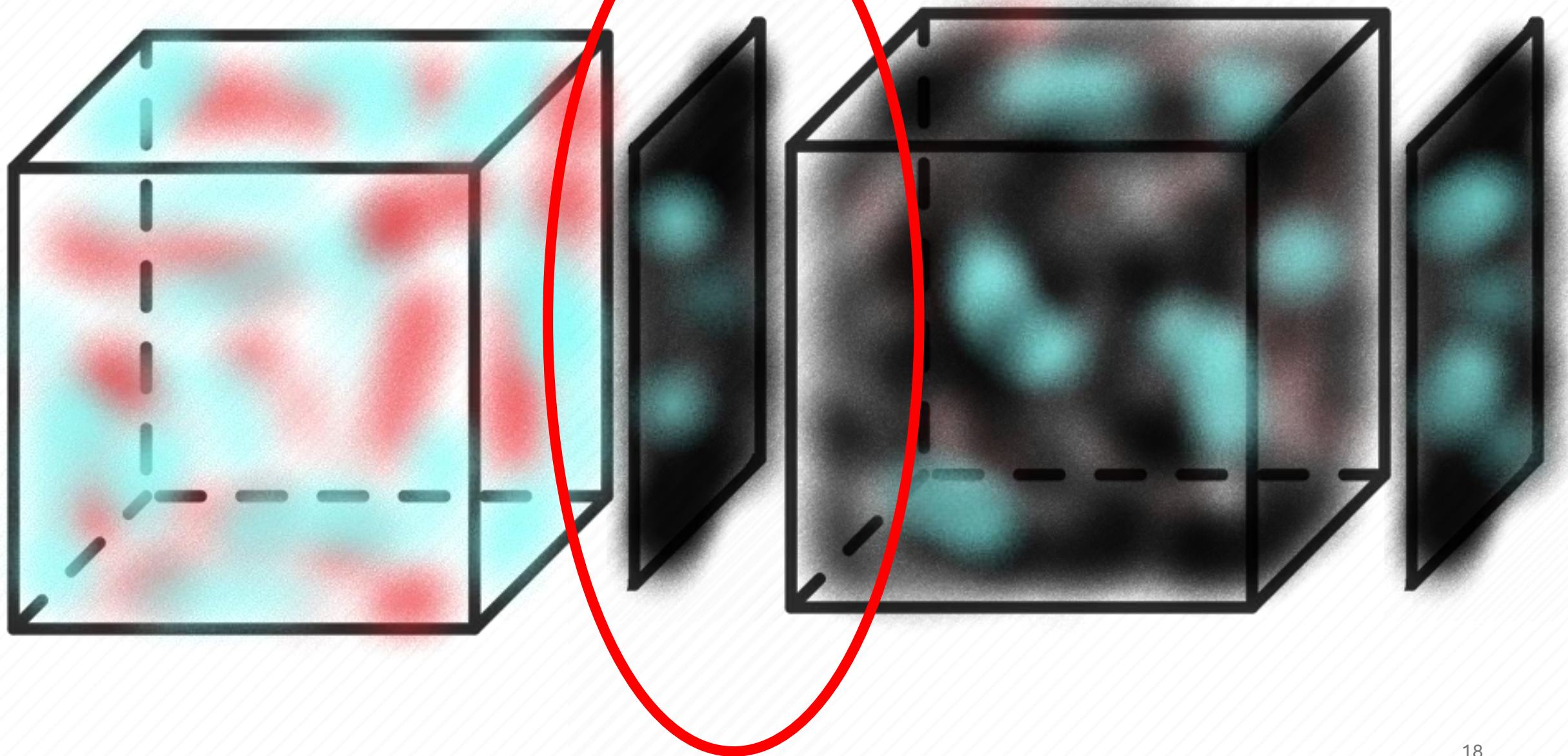
Prompt



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"

“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?

“Degrees of Freedom” in LLM Prompts

- When prompting an LLM, any detail you don't explicitly specify 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.

Prompt formulas

- Structured guidelines or templates used to generate responses from language models like GPT-4o
 - 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].”

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

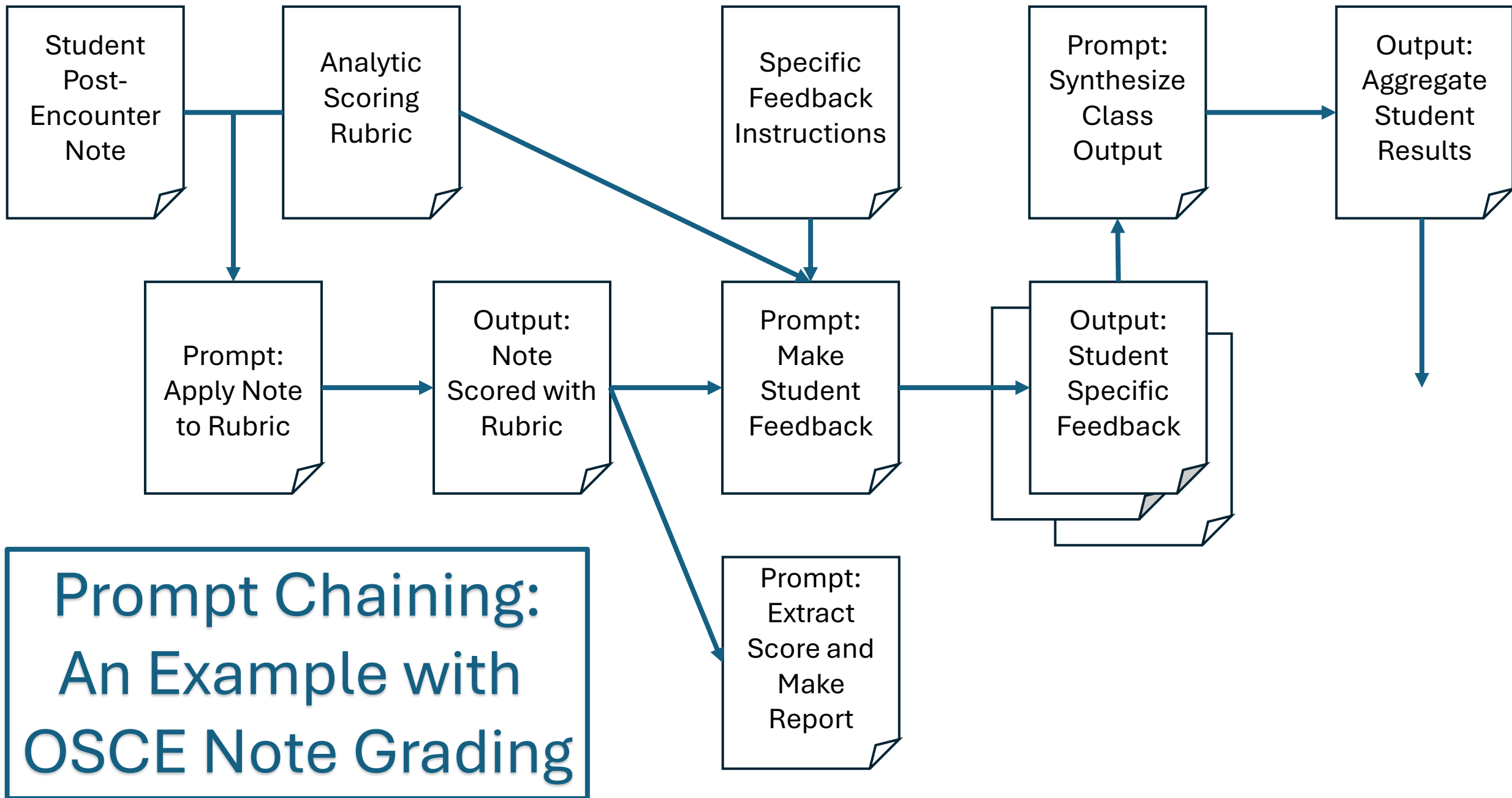
Prompt Chaining

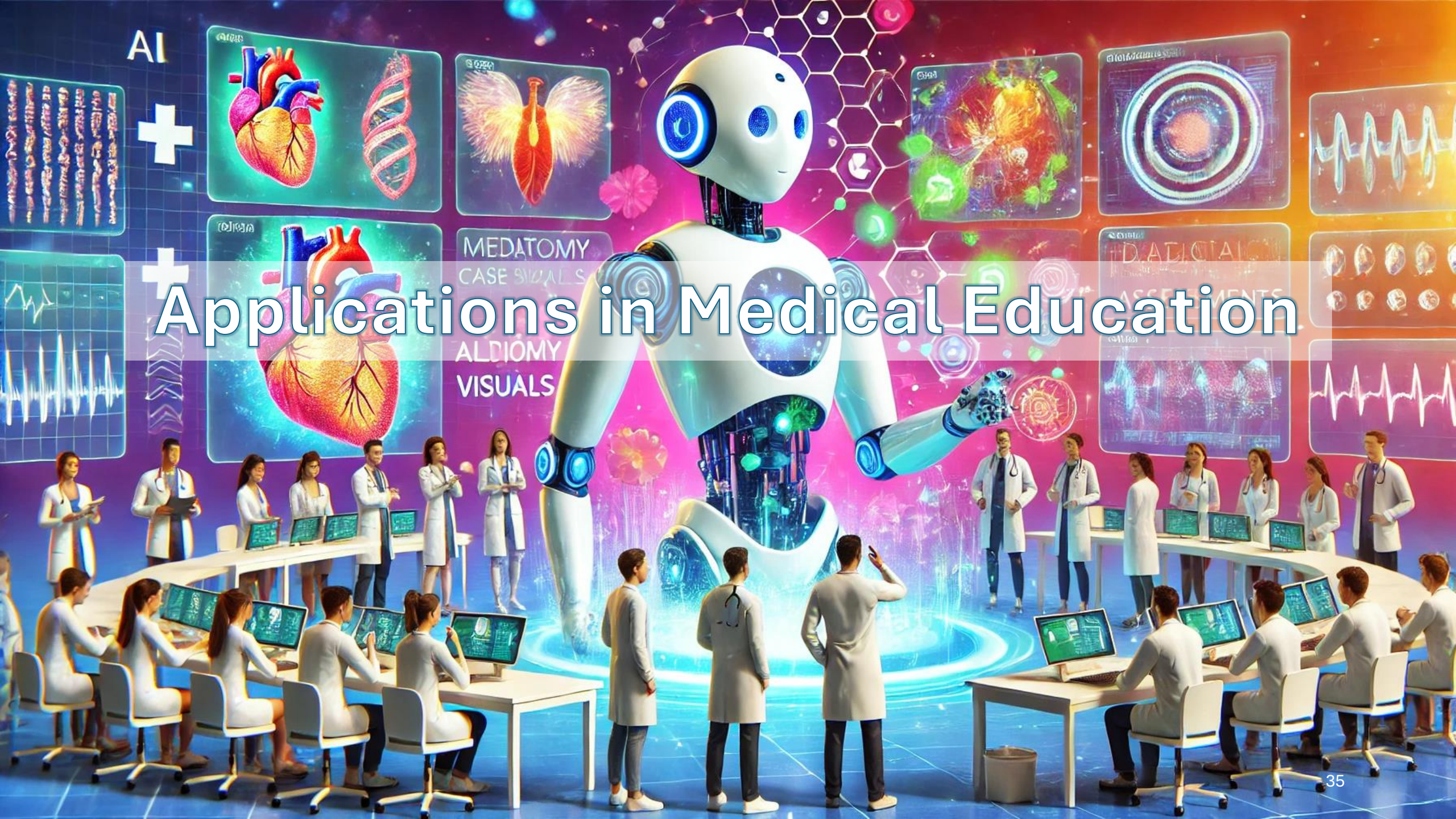
Strengths

- Modular approach
- Reduced complexity of instructions
- Error isolation
- Clear, structured 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 single-prompt process





AI
+

+

Applications in Medical Education

MEDATOMY
CASE SIGNALS

ALDIOMY
VISUALS

DIAGNOSTIC
ASSESSMENTS

Applications in Medical Education

[Scoping Review of AI in Medical Education Assessment](#) (2024; BEME Guide)

[Supplemental Material – Articles included in the Guide](#)

- Admission / Selection
- Teaching / Instruction (Simulation)
- Assessment (Clinical Skills, MCQ generation, text analysis, prediction, surgical skills)
- Clinical Reasoning
- Automation of case /procedural logs
- Knowledge / Attributes of AI
- Performance of LLMs on Medical Examinations

+ Perspectives!

Applications in Medical Education

Artificial Intelligence in Health Professions Education assessment (AMEE Guide No 178) – 2025

Issue, Theory, Potential, Examples / Best Practices, Limitations

- Theories and Frameworks
- Human Responses
- Personalised and flexible summative assessment
- Assessment standardization
- Personalized and flexible formative assessment
- Take-home essay
- Other Assessment Types
- AI as Tutor
- AI as Learner
- Competency-based Assessment
- Difficult areas and ethical issues

LLM Applications in Assessment

- **MCQ / Item Generation**

- Can use a structured prompt to help generate new items for a specific parts of the curriculum.
- “You are a medical educator that is an expert in [topic]. You are developing an examination to assess student’s knowledge of [more specific topic]. Your are generating a USMLE-style item for this purpose. The item should be appropriate for [level of learner].”
- Even better: “I will post an example item; please mimic this style while adhering to USMLE guidelines.”

LLM Applications in Assessment

From the [Step 2 CK sample questions](#):

A hospitalized 57-year-old man has had severe progressive pain in his left knee since awakening 2 hours ago. He was admitted to the hospital 2 days ago for an acute myocardial infarction. Cardiac catheterization showed occlusion of the left anterior descending artery, and he underwent placement of a stent. Current medications include aspirin, metoprolol, lisinopril, simvastatin, clopidogrel, and heparin. Vital signs are within normal limits. Examination of the knee shows a large effusion. The knee is hot to touch and erythematous. He holds the knee in 30 degrees of flexion; the pain is exacerbated with further flexion or extension. Laboratory studies show:

Hematocrit	40%
Leukocyte count	13,000/mm ³
Serum	
Ca ²⁺	9.2 mg/dL
Urea nitrogen	15 mg/dL
Creatinine	1.0 mg/dL
Albumin	3.6 g/dL

An x-ray of the left knee shows calcification of the synovium. Which of the following is the most likely diagnosis?

- (a) Deep venous thrombosis
- (b) Gonorrhea
- (c) Gout
- (d) Hemarthrosis
- (e) Pseudogout**
- (f) Septic Arthritis

LLM Applications in Assessment

- “I would like for you to generate a new MCQ item in the same format with the same response options, and I want you to change the case content to make one of the incorrection options the correct answer.”
- “Change the item content so the correct answer is now [pick other option]”

LLM Applications in Assessment

A 60-year-old man comes to the emergency department with severe pain and swelling in his right big toe that began abruptly overnight. He has a history of hypertension and chronic kidney disease. Medications include lisinopril and hydrochlorothiazide. He drinks two to three beers nightly. Physical examination reveals a swollen, erythematous right first metatarsophalangeal joint that is warm to the touch. He has marked tenderness and pain with even light touch or passive motion. Laboratory studies show:

Hematocrit 38%

Leukocyte count 11,500/mm³

Serum Ca²⁺ 9.5 mg/dL

Urea nitrogen 28 mg/dL

Creatinine 1.8 mg/dL

Uric acid 9.0 mg/dL

An x-ray of the foot reveals no bony erosions. Which of the following is the most likely diagnosis?

- (a) Deep venous thrombosis
- (b) Gonorrhea
- (c) Gout**
- (d) Hemarthrosis
- (e) Pseudogout
- (f) Septic Arthritis

LLM Applications in Assessment

- “I would like for you to generate a follow-up MCQ item about the best step in management for this condition.”

The 60-year-old man described above is diagnosed with acute gout affecting the right first metatarsophalangeal joint. Which of the following is the most appropriate next step in management?

- (A) Allopurinol
- (B) Colchicine
- (C) Hydrochlorothiazide dose adjustment
- (D) Indomethacin**
- (E) Intra-articular corticosteroid injection

LLM Applications in Assessment

The 60-year-old man described above is diagnosed with acute gout affecting the right first metatarsophalangeal joint. Which of the following is the most appropriate next step in management?

- (A) Allopurinol
- (B) Colchicine
- (C) Hydrochlorothiazide dose adjustment
- (D) Indomethacin**
- (E) Intra-articular corticosteroid injection

Rationale:

- **Acute Gout Management:** The priority is to control inflammation and relieve pain. First-line treatments for acute attacks include **nonsteroidal anti-inflammatory drugs (NSAIDs)** like indomethacin, colchicine, or corticosteroids.
- **Allopurinol (A):** Used for long-term urate-lowering therapy to prevent future attacks, but it is not initiated during an acute flare due to the risk of exacerbating symptoms.
- **Colchicine (B):** An alternative for patients who cannot tolerate NSAIDs, but NSAIDs are generally preferred due to quicker symptom relief.
- **Hydrochlorothiazide adjustment (C):** Thiazide diuretics can contribute to hyperuricemia and recurrent gout, but addressing this is part of long-term management rather than acute treatment.
- **Intra-articular corticosteroids (E):** Used if NSAIDs are contraindicated or if a single, large joint is involved and the patient cannot tolerate systemic treatment.

LLM Applications in Assessment

AI Patient Actor from
[Dartmouth Laboratories](#)



Neuroscience-Informed
Learning & Education
Lab

Geisel
School of Medicine
at Dartmouth



AI Patient Actor



Instructions

Settings

Mode selection

☒ Foundational

☐ OnDoc

Select case

A Demo Case

Select language

English

Doctor input

☒ Text-only ☐ Speech + Text

Patient response

☒ Text ☐ Speech

End Patient Encounter and Receive Feedback

Hi! I am Dr. X. What brings you to the clinic today?



Created by [NILE Lab](#) and [RDS @ Dartmouth Library](#). For questions, please contact [Thomas Thesen](#).

OSCE Grader

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

README

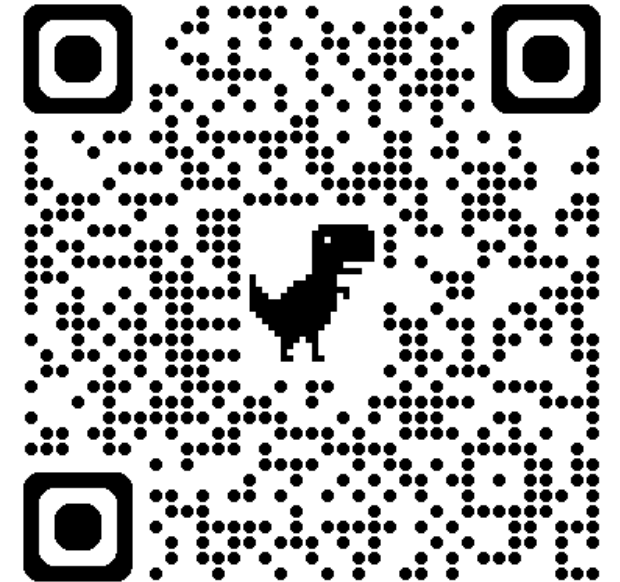
MIT license

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)



Other interesting use cases

- First-pass TA
 - Built a GPT that is further trained on materials for the course.
 - Students first ask the GPT their question about the course material.
 - GPT then either (a) provides the answer with where it can be found in the material or (b) provides where it can be found in the material [I forget]
 - If not a satisfactory answer, student indicates this and the question is passed on to the instructor
- Precision Medical Education (e.g., [NYU Grossman](#))
 - Custom tailored educational opportunities

Other interesting use cases

- Curriculum Design / Instructional alignment
 - Generate / refine learning objectives
 - Lesson plan structuring
 - Develop / refine assessment content aligned with the LOs
- Rubric refinement
 - See how the GPT interprets rubric rules
- [Clinical Mind AI \(Stanford\)](#) ; [MedSimAI](#)

**GPT
IMPLEMENTATION
CONSIDERATIONS**

**DATA
PRIVACY**

**USER
ETHICAL
ACCESSABILITY**

**USER
CONCERNITY**

**DATA
PRIVACY**

**DATA
PRIVACY**

ETHICALLY

**ETHICANCE PERFORMANCE
ACTIVIZATION OXTIMIZATION**



Chatbot vs. API Usage

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 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?

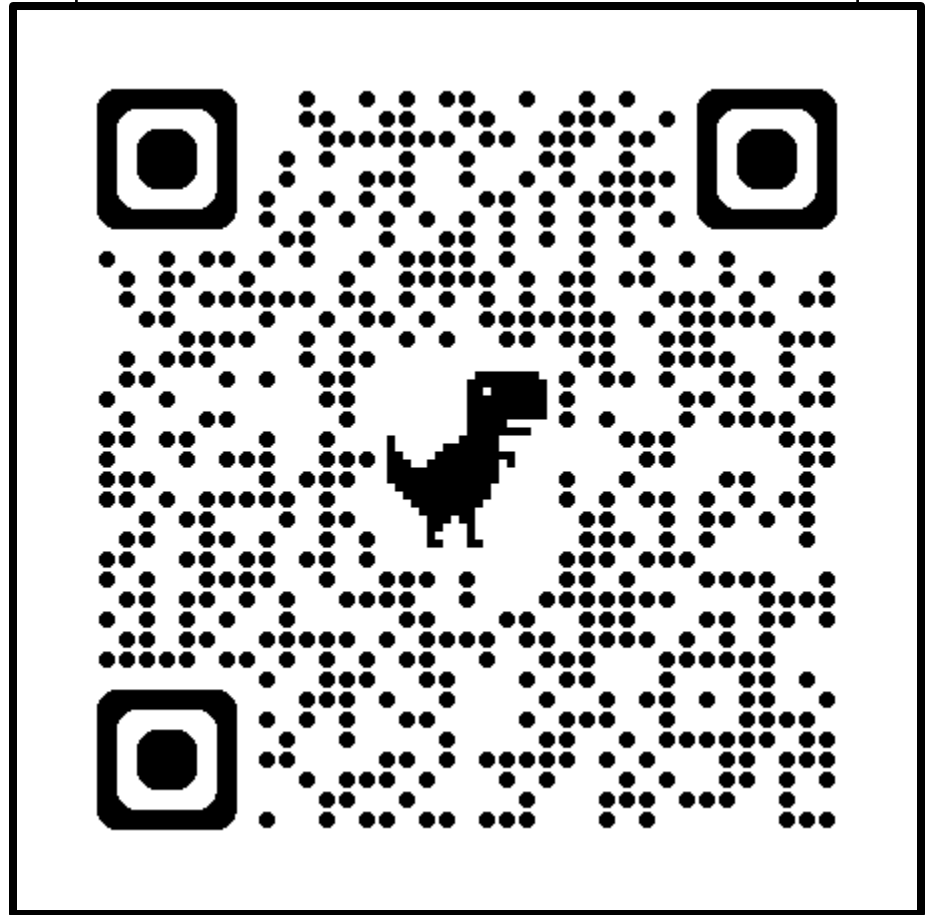
Christopher Runyon

CRunyon@nbme.org

or use this QR Code



Presentation Slides:



(both presentations available)