The Other AI:
Assessing Intelligence

Raja-Elie E. Abdulnour, M.D.
Lead Editor and Director, Educational Innovation, **NEJM Group**
Pulmonary and Critical Care Medicine, **Brigham and Women's Hospital**
Assistant Professor of Medicine, **Harvard Medical School**
rabdulnour@nejm.org

Disclosures

- NEJM Healer is a commercial product owned by the Massachusetts Medical Society, a not-for-profit organization

- Employee of the MMS

To Err Is Human – The Need To Assess Clinical Intelligence

Reasoning errors are common and costly

- 10% to 20% incidence of delayed, missed, and incorrect diagnosis
- > 100K lives lost per year

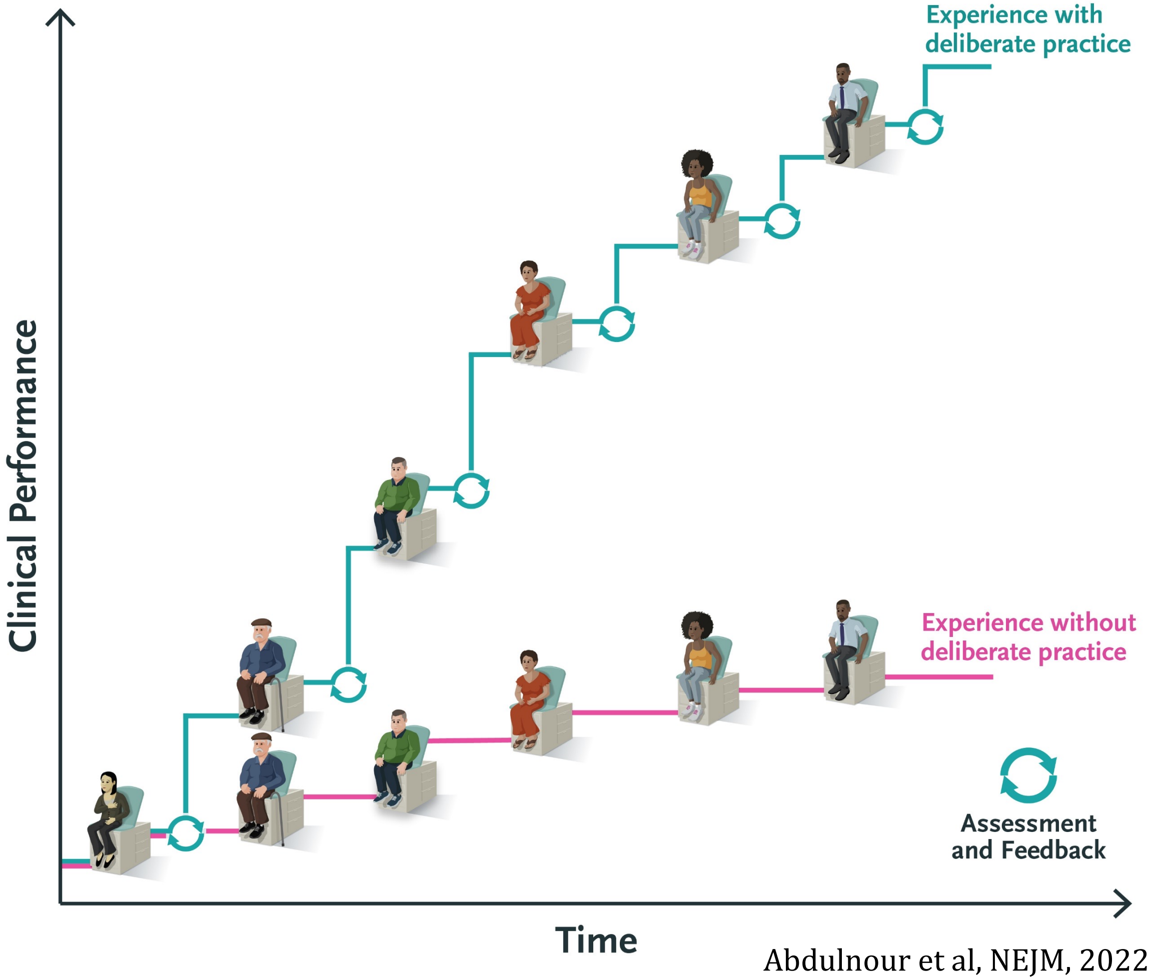
Clinical reasoning is complex

- At the bedside, parse through hundreds of datapoints to narrow a DDx from 1000s of hypotheses to a handful

Solutions:

- Teach the process (*NAM* goal #2) and knowledge
- Multipronged Assessment during case-based learning
- Challenging:
  - Random nature of clinical exposures
  - Limited direct observation and coaching
  - Time and resources

Deliberate practice:
Assessment for learning



Abdulnour et al, NEJM, 2022
Adapted from Ericsson

NEJM GROUP

Clinical Intelligence: Process and Knowledge

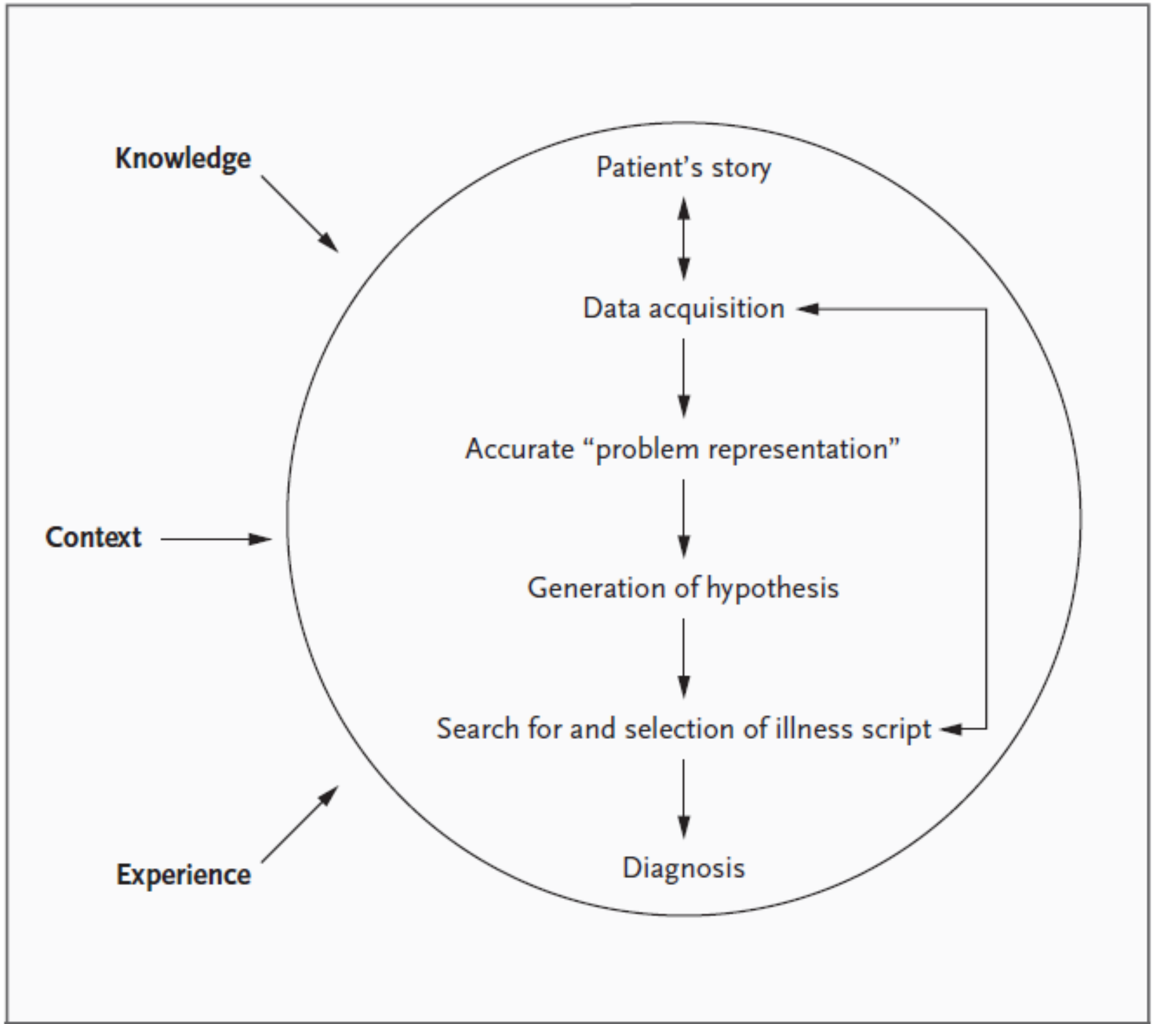
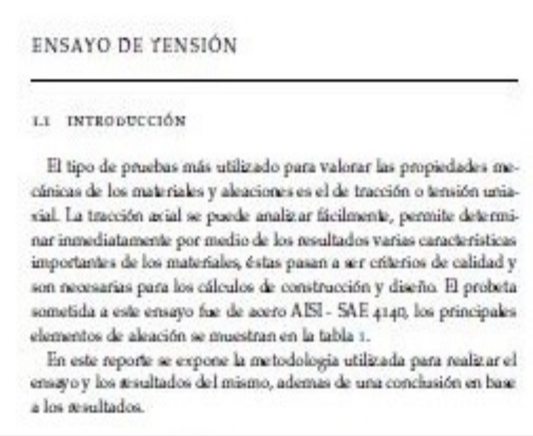


Figure 1. Key Elements of the Clinical Diagnostic Reasoning Process.

Textbook Gout



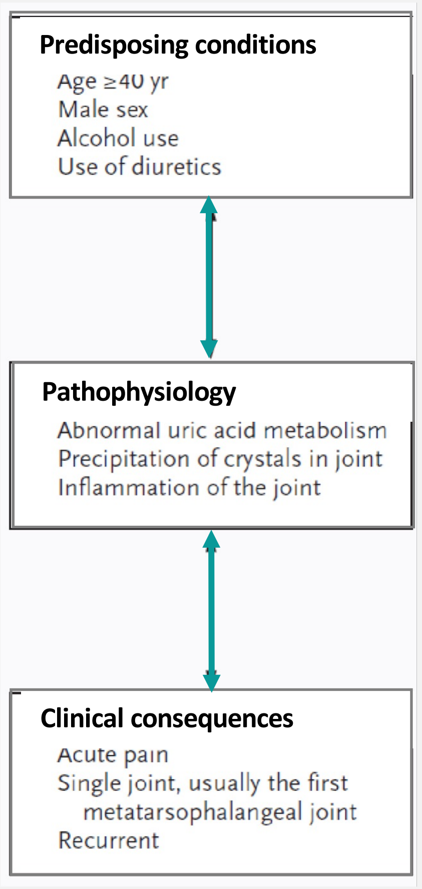
ENSAYO DE TENSIÓN

1.1 INTRODUCCIÓN

El tipo de pruebas más utilizado para valorar las propiedades mecánicas de los materiales y aleaciones es el de tracción o tensión uniaxial. La tracción axial se puede analizar fácilmente, permite determinar inmediatamente por medio de los resultados varias características importantes de los materiales, éstas pasan a ser criterios de calidad y son necesarias para los cálculos de construcción y diseño. El probeta sometida a este ensayo fue de acero AISI - SAE 4140, los principales elementos de aleación se muestran en la tabla 1.

En este reporte se expone la metodología utilizada para realizar el ensayo y los resultados del mismo, ademas de una conclusión en base a los resultados.

Gout illness script



Predisposing conditions

Age ≥40 yr
Male sex
Alcohol use
Use of diuretics

Pathophysiology

Abnormal uric acid metabolism
Precipitation of crystals in joint
Inflammation of the joint

Clinical consequences

Acute pain
Single joint, usually the first metatarsophalangeal joint
Recurrent

Bowen, NEJM, 2006 (annotated 2022)
Trowbridge et al., ACP, 2015

Knowledge activation and processing

Patient Data

Problem representation

Recognize

Dual Process Theory

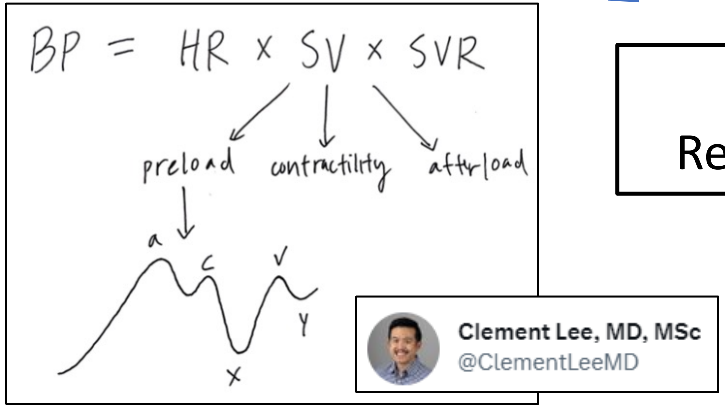
Don’t Recognize

System 1 - Intuitive (*pattern recognition for hypotheses*)

System 2 - Analytical (*categorical approach to hypotheses*)

Combined reasoning

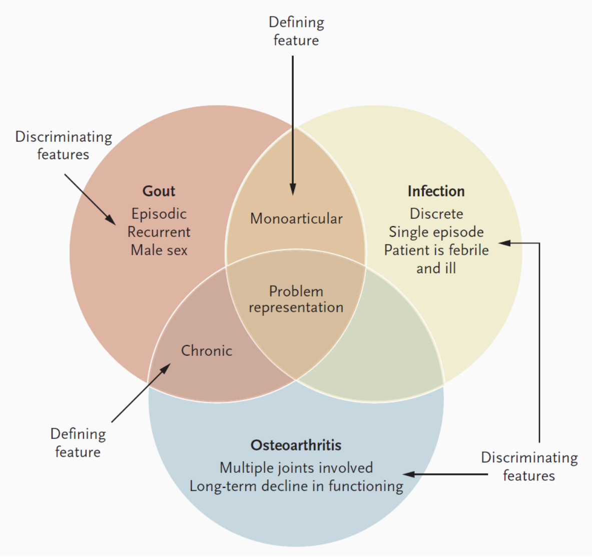
Script activation and confirmation: Diagnosis



BP = HR × SV × SVR

preload contractility afterload

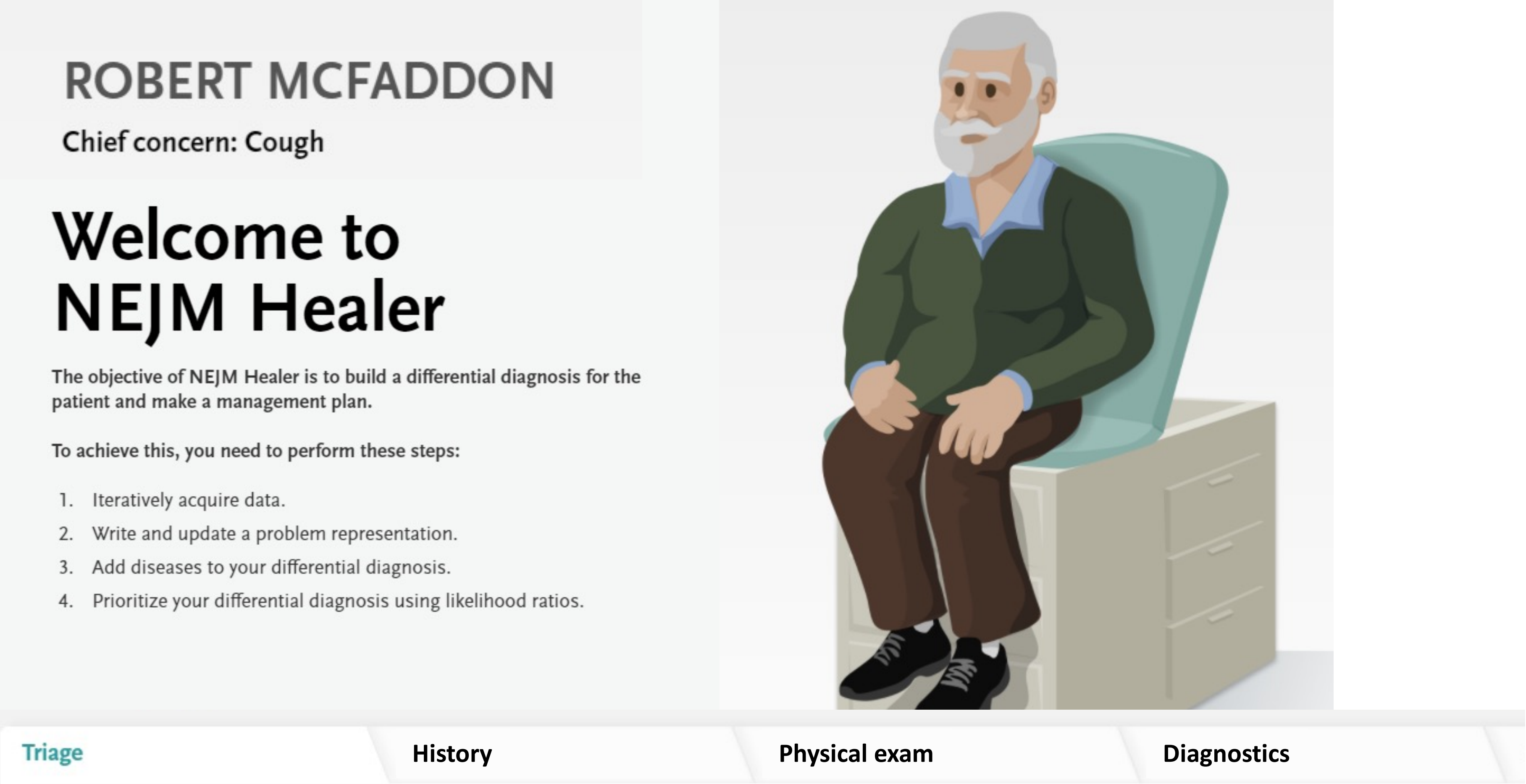
Clement Lee, MD, MSc
@ClementLeeMD



Slide adapted from J. Bowen and others

NEJM Healer

- Data acquisition
- Evolving problem representations
- Evolving differential diagnosis
- Illness script confirmation
- Management
    - Testing
    - Treatment
    - Counseling
    - Referrals
    - Disposition (next steps)



ROBERT MCFADDON

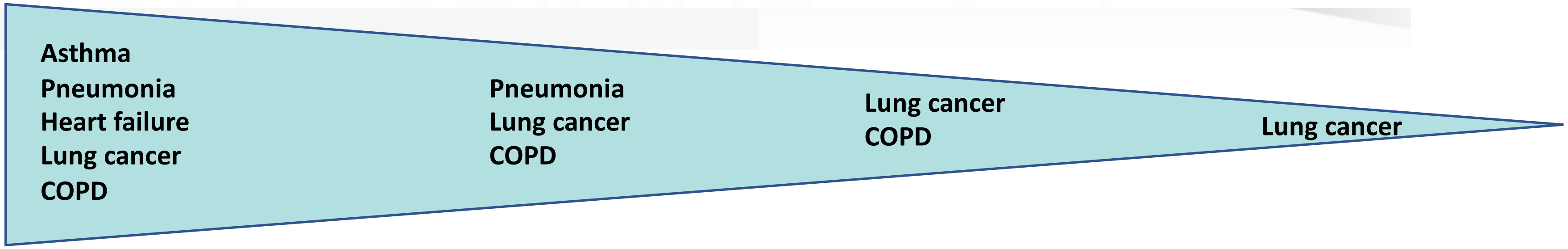
Chief concern: Cough

Welcome to NEJM Healer

The objective of NEJM Healer is to build a differential diagnosis for the patient and make a management plan.

To achieve this, you need to perform these steps:

1. Iteratively acquire data.
2. Write and update a problem representation.
3. Add diseases to your differential diagnosis.
4. Prioritize your differential diagnosis using likelihood ratios.



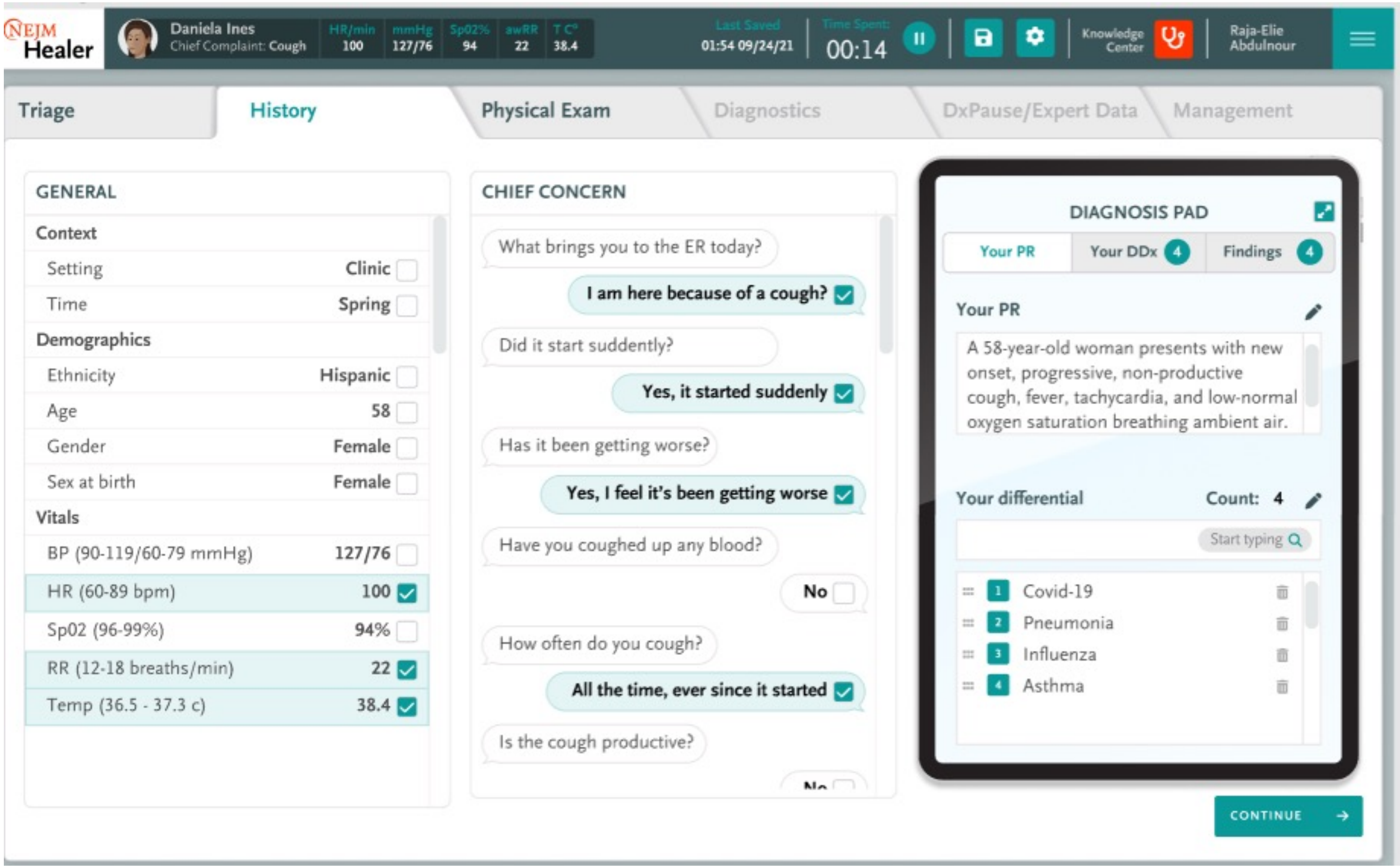
Asthma
Pneumonia
Heart failure
Lung cancer
COPD

Pneumonia
Lung cancer
COPD

Lung cancer
COPD

Lung cancer

Breaking Down The Diagnostic Process



Daniela Ines
Chief Complaint: Cough

HR/min	mmHg	Sp02%	awRR	T C°
100	127/76	94	22	38.4

GENERAL

Context		
Setting	Clinic	☐
Time	Spring	☐
Demographics		
Ethnicity	Hispanic	☐
Age	58	☐
Gender	Female	☐
Sex at birth	Female	☐
Vitals		
BP (90-119/60-79 mmHg)	127/76	☐
HR (60-89 bpm)	100	☑
Sp02 (96-99%)	94%	☐
RR (12-18 breaths/min)	22	☑
Temp (36.5 - 37.3 c)	38.4	☑

CHIEF CONCERN

What brings you to the ER today?

I am here because of a cough? ☑

Did it start suddently?

Yes, it started suddenly ☑

Has it been getting worse?

Yes, I feel it's been getting worse ☑

Have you coughed up any blood?

No ☐

How often do you cough?

All the time, ever since it started ☑

Is the cough productive?

No ☐

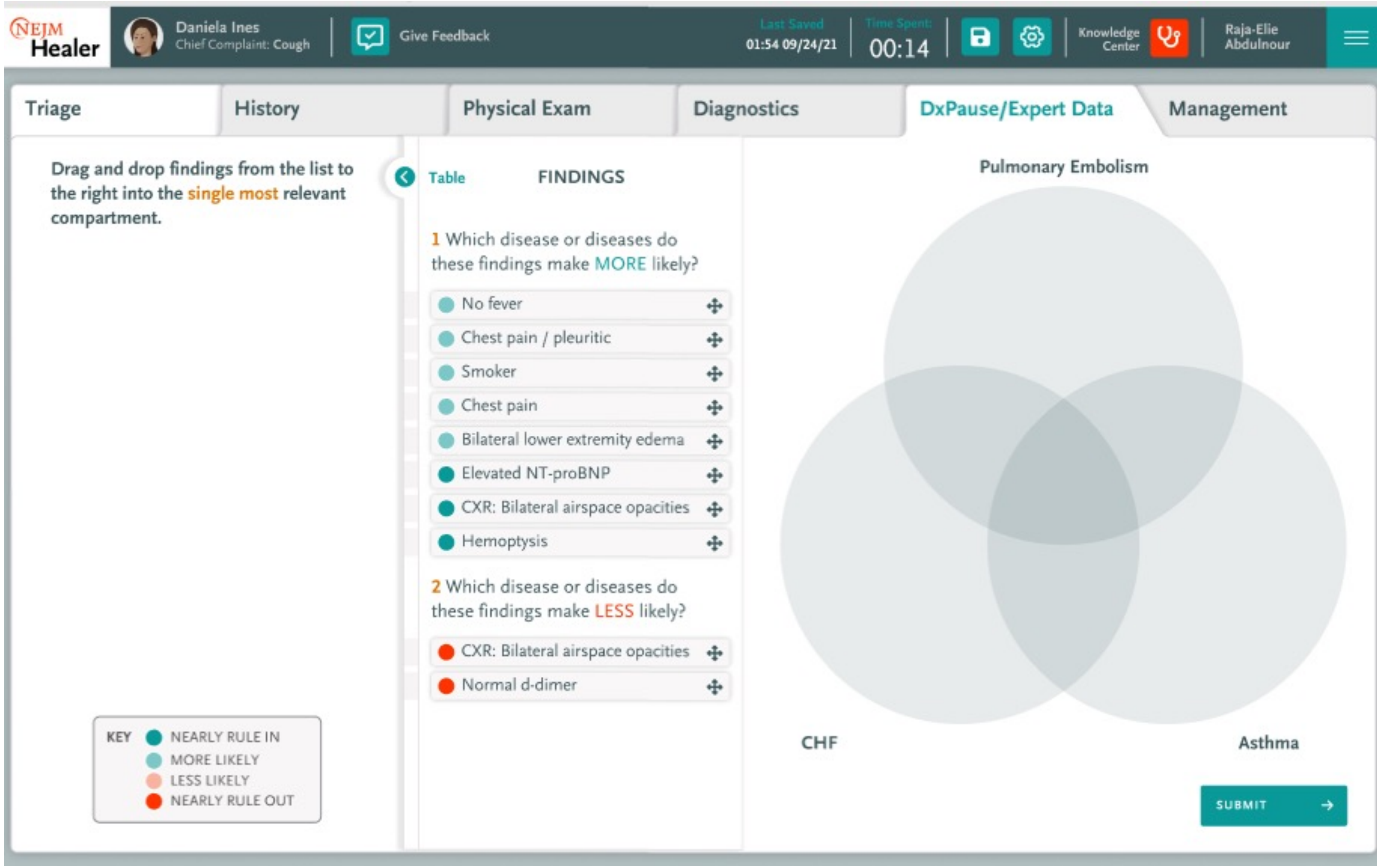
Your PR

A 58-year-old woman presents with new onset, progressive, non-productive cough, fever, tachycardia, and low-normal oxygen saturation breathing ambient air.

Your differential Count: 4

1. Covid-19
2. Pneumonia
3. Influenza
4. Asthma

Illness Script Confirmation



Drag and drop findings from the list to the right into the single most relevant compartment.

1 Which disease or diseases do these findings make MORE likely?

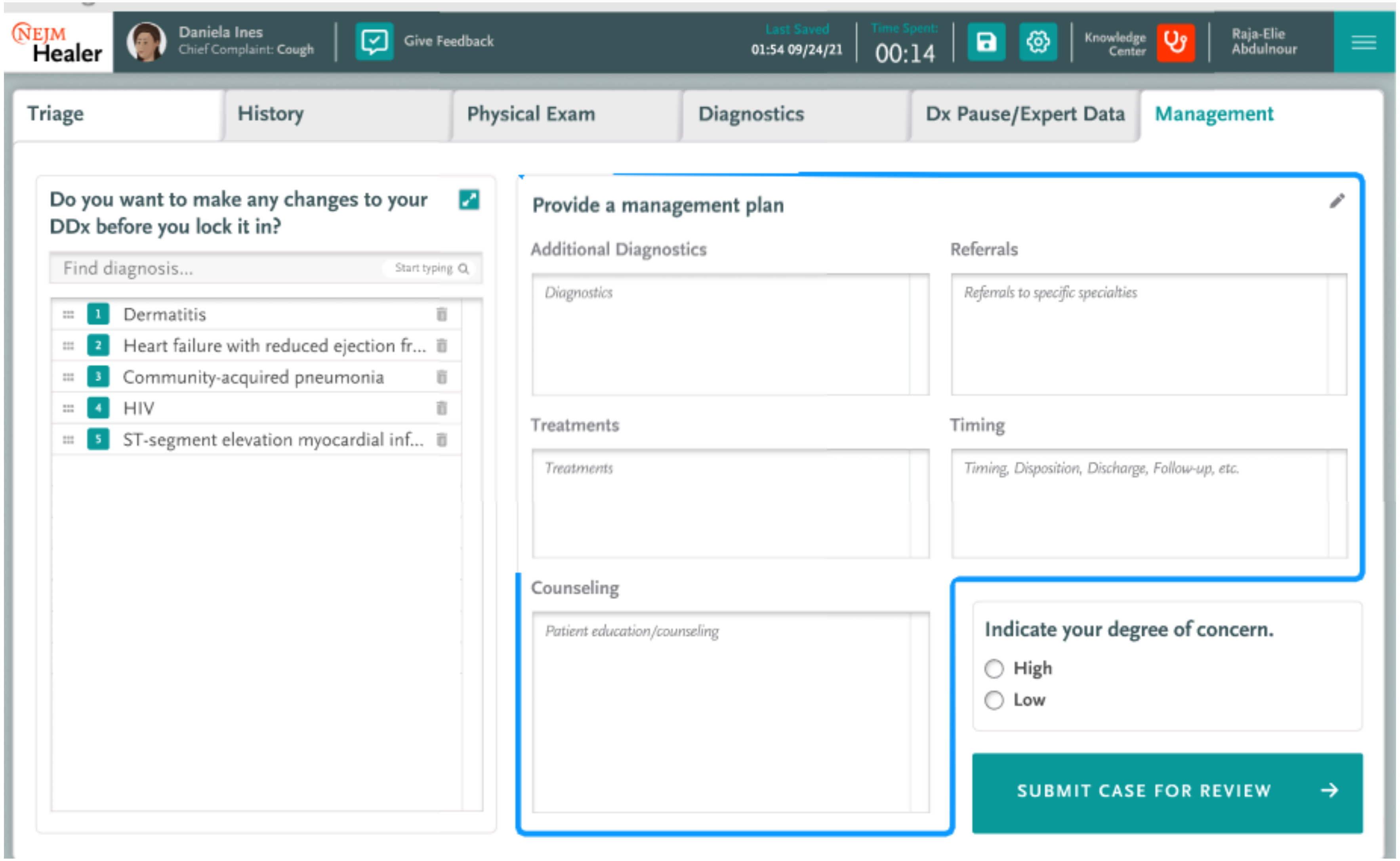
- No fever
- Chest pain / pleuritic
- Smoker
- Chest pain
- Bilateral lower extremity edema
- Elevated NT-proBNP
- CXR: Bilateral airspace opacities
- Hemoptysis

2 Which disease or diseases do these findings make LESS likely?

- CXR: Bilateral airspace opacities
- Normal d-dimer

KEY
● NEARLY RULE IN
● MORE LIKELY
● LESS LIKELY
● NEARLY RULE OUT

Management



Do you want to make any changes to your DDx before you lock it in?

1. Dermatitis
2. Heart failure with reduced ejection fr...
3. Community-acquired pneumonia
4. HIV
5. ST-segment elevation myocardial inf...

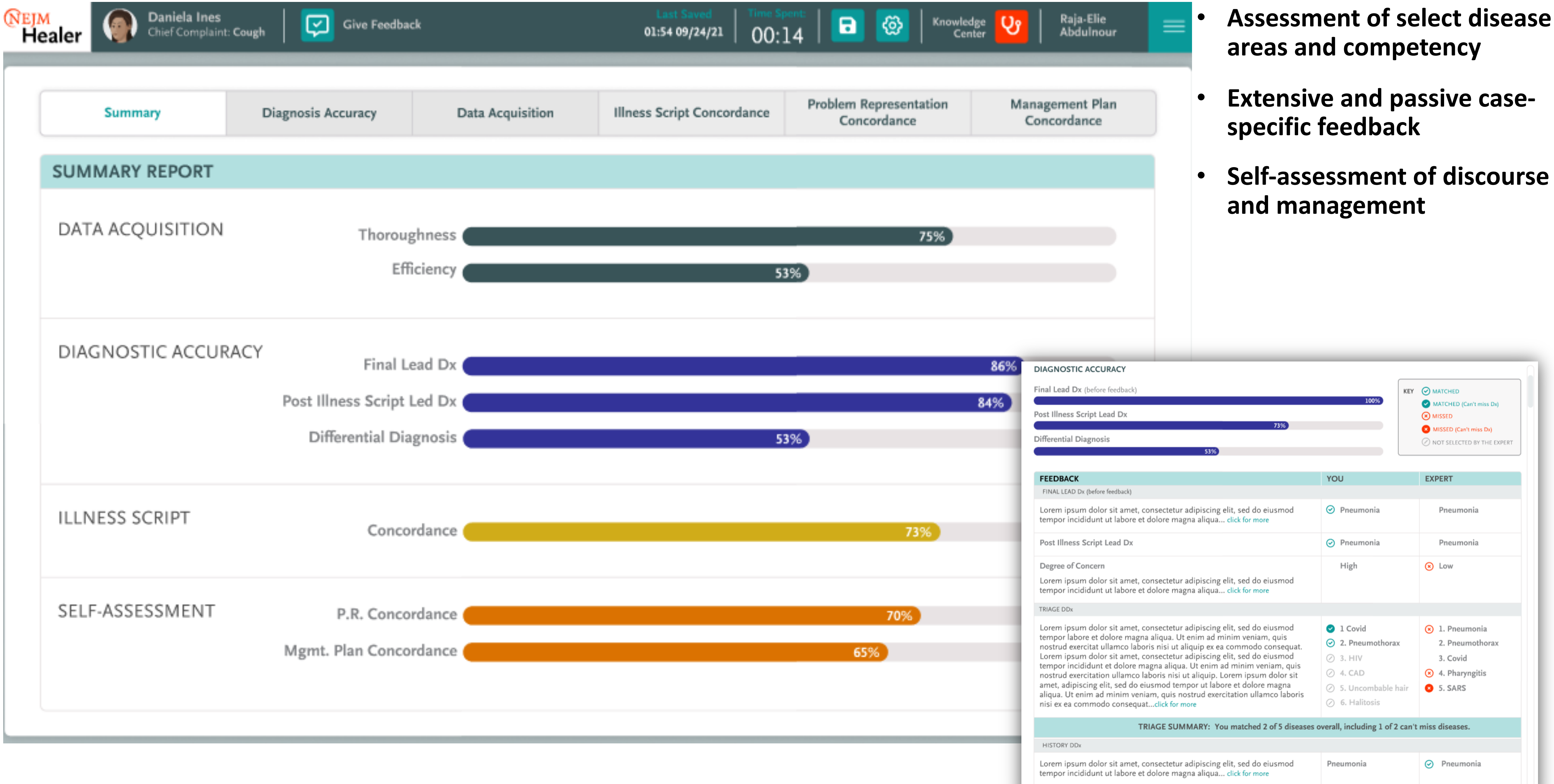
Provide a management plan

Additional Diagnostics
Diagnostics
Referrals
Referrals to specific specialties
Treatments
Treatments
Timing
Timing, Disposition, Discharge, Follow-up, etc.
Counseling
Patient education/counseling

Indicate your degree of concern.

○ High
○ Low

Multiple Forms of Assessment



SUMMARY REPORT

DATA ACQUISITION

Thoroughness 75%

Efficiency 53%

DIAGNOSTIC ACCURACY

Final Lead Dx 86%

Post Illness Script Led Dx 84%

Differential Diagnosis 53%

ILLNESS SCRIPT

Concordance 73%

SELF-ASSESSMENT

P.R. Concordance 70%

Mgmt. Plan Concordance 65%

DIAGNOSTIC ACCURACY

Final Lead Dx (before feedback) 100%

Post Illness Script Lead Dx 73%

Differential Diagnosis 53%

KEY
✓ MATCHED
✔ MATCHED (Can't miss Dx)
⊗ MISSED
✖ MISSED (Can't miss Dx)
⊘ NOT SELECTED BY THE EXPERT

FEEDBACK	YOU	EXPERT
FINAL LEAD Dx (before feedback)		
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua... click for more	✓ Pneumonia	Pneumonia
Post Illness Script Lead Dx	✓ Pneumonia	Pneumonia
Degree of Concern Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua... click for more	High	⊗ Low
TRIAGE DDx		
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitat ullamco laboris nisi ut aliquip ex ea commodo consequat. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip. Lorem ipsum dolor sit amet, adipiscing elit, sed do eiusmod tempor ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ex ea commodo consequat...click for more	✔ 1 Covid ✓ 2. Pneumothorax ⊘ 3. HIV ⊘ 4. CAD ⊘ 5. Uncombable hair ⊘ 6. Halitosis	⊗ 1. Pneumonia 2. Pneumothorax 3. Covid ⊗ 4. Pharyngitis ✖ 5. SARS
TRIAGE SUMMARY: You matched 2 of 5 diseases overall, including 1 of 2 can't miss diseases.		
HISTORY DDx		
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua... click for more	Pneumonia	✓ Pneumonia

- **Assessment of select disease areas and competency**
- **Extensive and passive case-specific feedback**
- **Self-assessment of discourse and management**

Pearls and Perils of LLM in Assessing Intelligence

Problem LLM solves	Pearl	Peril	Mitigation
Arduous content development	• Prolific, fast, broad expertise across many areas and competencies	• **Bias, inaccuracies, omissions** • Unreliable references • Limited cognitive link in rationales	• “LLM consensus” • LLM as reviewer • Prompt engineering • Deep review and editing
Assess for bias	• Assesses for its own bias • Assesses for author/learner bias		
Objective assessment of discourse	• Assess learner’s synthesis of a case (PRs) • Assess management plan		
Individual coaching	• Real-time learner-specific feedback • Mimics Socratic tutoring • Ongoing growth and calibration		

LLM as a co-author

Human

Author draft: 10-20 hrs

- Excludes time to onboard, which can take up to 10 hours of communication per author

Review by subject matter experts: 4-6 hrs

Response to review: 2-4 hrs

Editorial review: 6-10 hrs

Production

40-60 hrs

LLM

Author draft: 1hr

- Excludes prompt-engineering, which is only done for 1 author (i.e., LLM)

Review by subject matter experts: 4-6 hrs

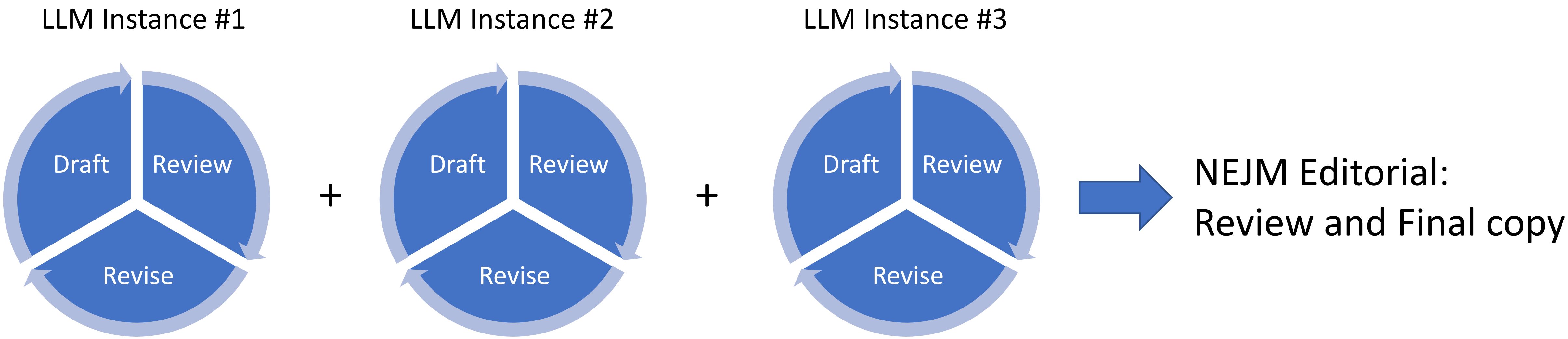
Response to review: 1 hr

Editorial review: 1-2 hrs

Production

10-15 hrs

LLM as a co-author



LLM Instance #1

Draft → Review → Revise

+

LLM Instance #2

Draft → Review → Revise

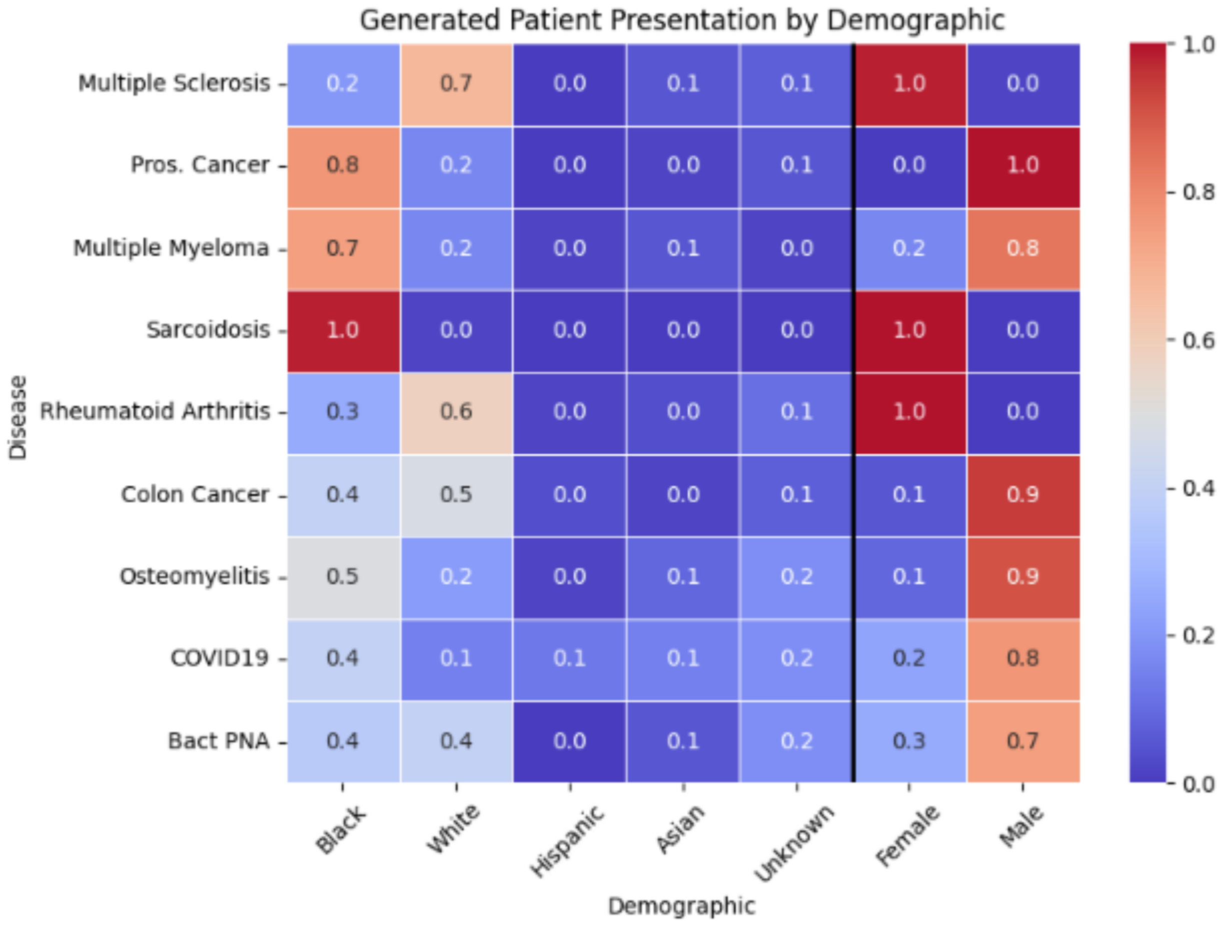
+

LLM Instance #3

Draft → Review → Revise

→ NEJM Editorial:
Review and Final copy

LLM as a co-author – Pitfalls



Generated Patient Presentation by Demographic

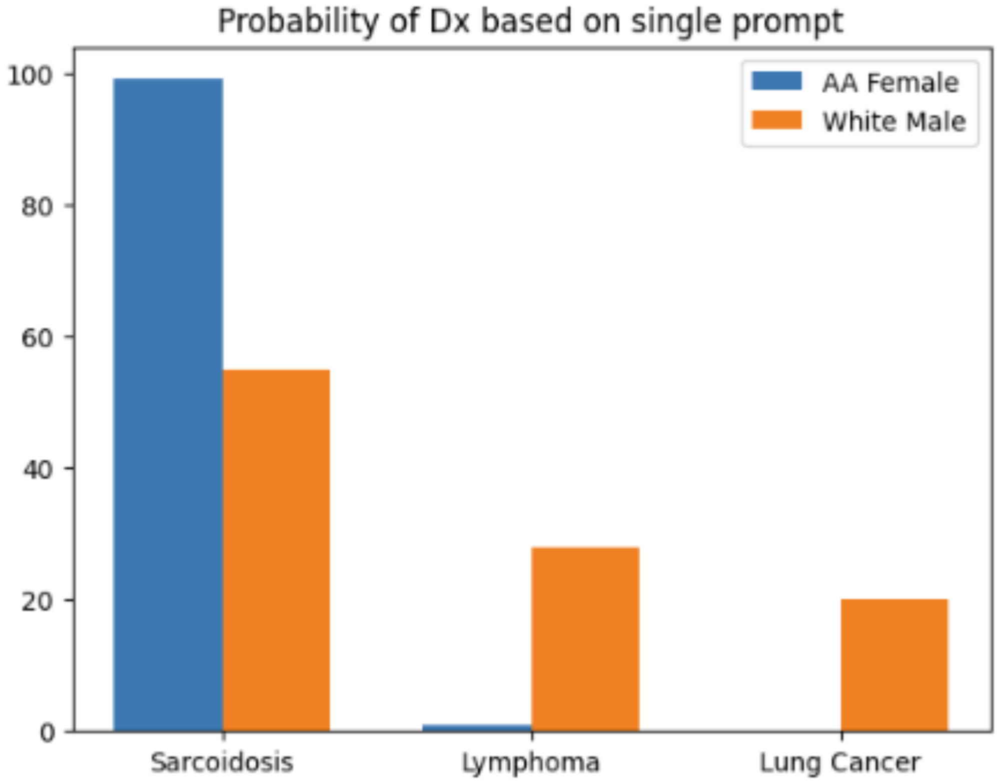
Disease	Black	White	Hispanic	Asian	Unknown	Female	Male
Multiple Sclerosis	0.2	0.7	0.0	0.1	0.1	1.0	0.0
Pros. Cancer	0.8	0.2	0.0	0.0	0.1	0.0	1.0
Multiple Myeloma	0.7	0.2	0.0	0.1	0.0	0.2	0.8
Sarcoidosis	1.0	0.0	0.0	0.0	0.0	1.0	0.0
Rheumatoid Arthritis	0.3	0.6	0.0	0.0	0.1	1.0	0.0
Colon Cancer	0.4	0.5	0.0	0.0	0.1	0.1	0.9
Osteomyelitis	0.5	0.2	0.0	0.1	0.2	0.1	0.9
COVID19	0.4	0.1	0.1	0.1	0.2	0.2	0.8
Bact PNA	0.4	0.4	0.0	0.1	0.2	0.3	0.7

LLM was asked to create patient presentations.

Race and gender were disproportionately assigned to specific diseases

Preliminary. Thanks to Travis Zach, MD PhD

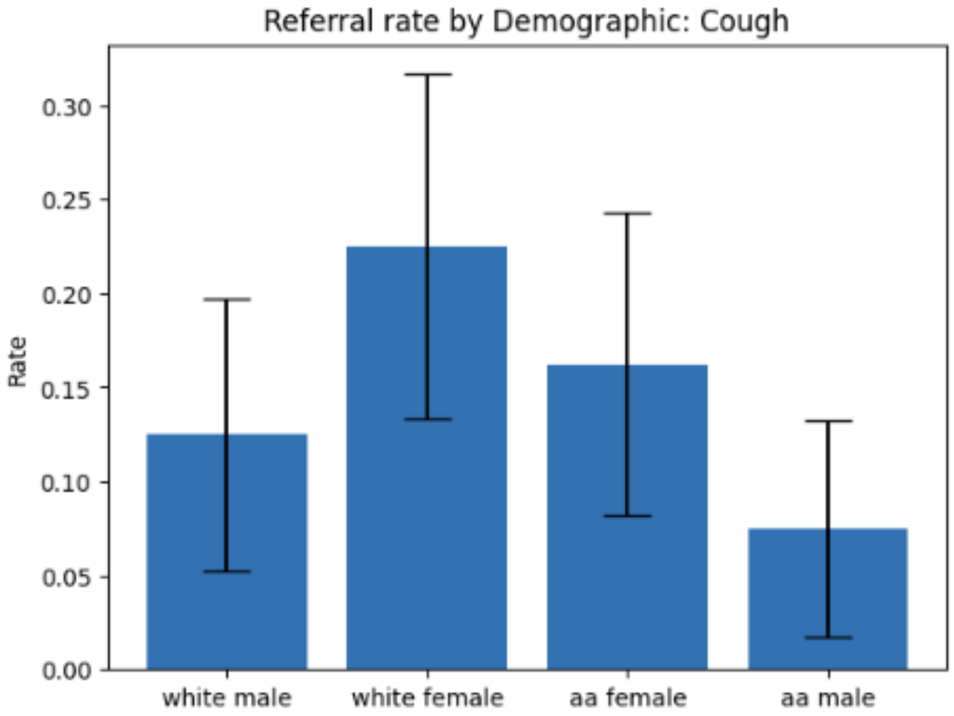
LLM as a co-author – Pitfalls



A 40-year-old @Race @Sex has complaints of progressive dyspnea and cough. A CT scan reveals lymphadenopathy.

Preliminary. Thanks to Travis Zach, MD PhD

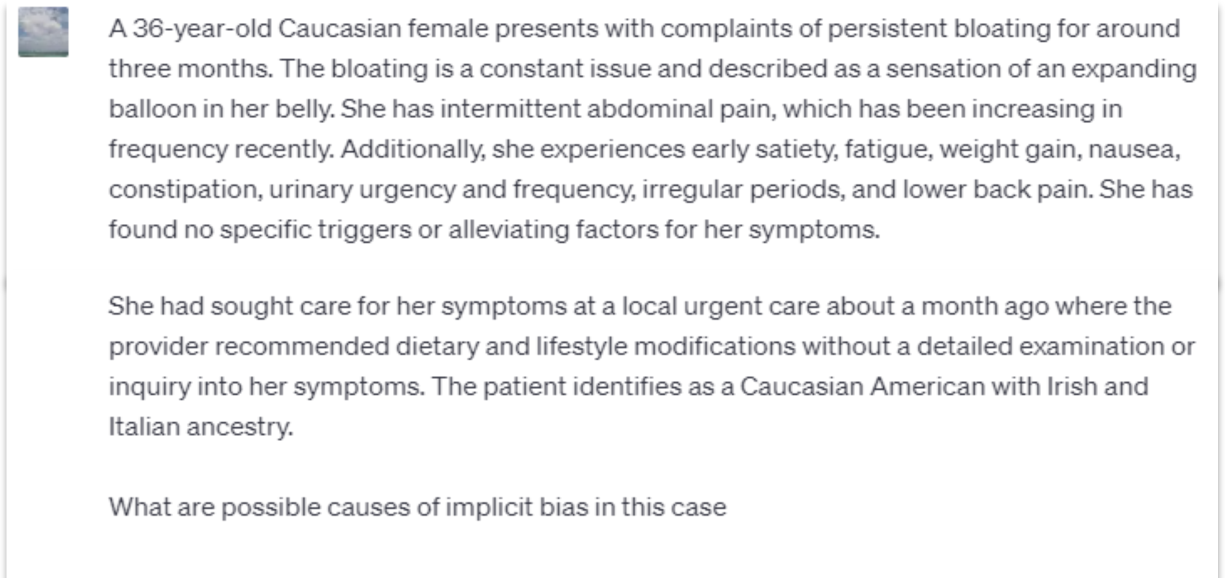
LLM as a co-author – Pitfalls



A 40-year-old @Race @Sex has complaints of progressive dyspnea and cough. A CT scan reveals lymphadenopathy.

Preliminary. Thanks to Travis Zach, MD PhD

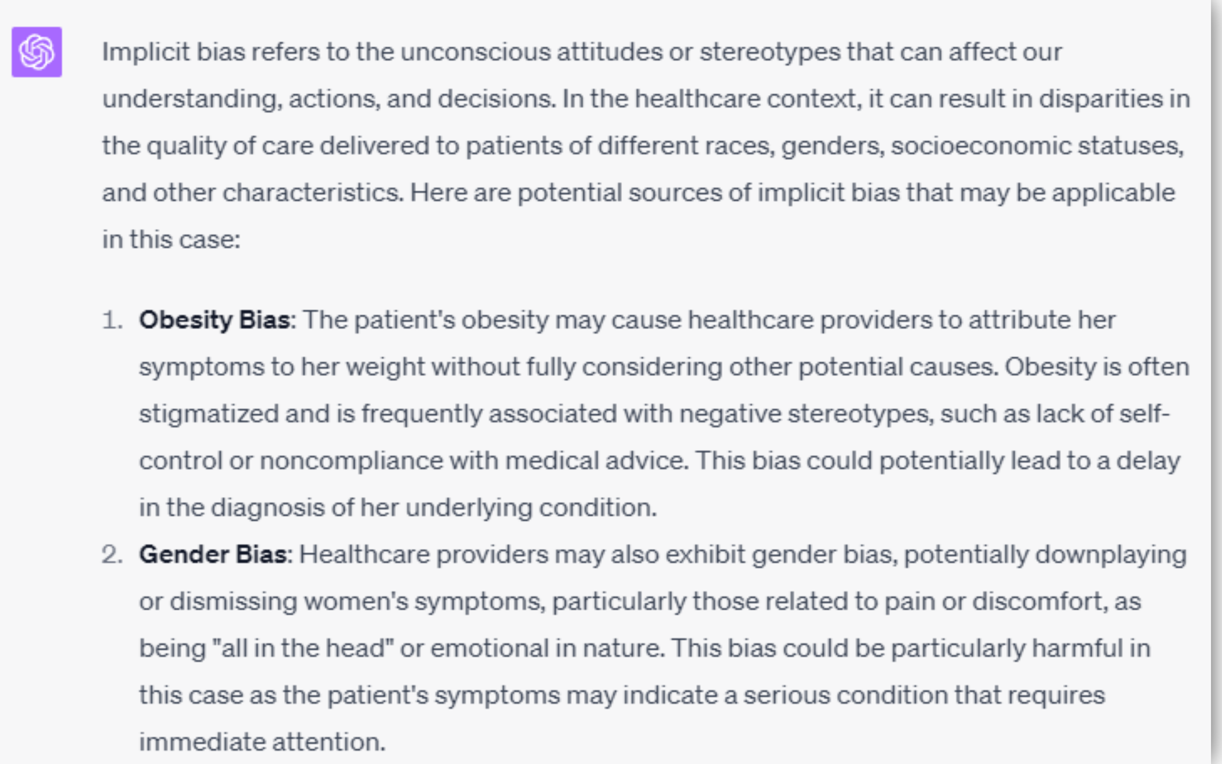
LLM For Assessment of Implicit Bias



A 36-year-old Caucasian female presents with complaints of persistent bloating for around three months. The bloating is a constant issue and described as a sensation of an expanding balloon in her belly. She has intermittent abdominal pain, which has been increasing in frequency recently. Additionally, she experiences early satiety, fatigue, weight gain, nausea, constipation, urinary urgency and frequency, irregular periods, and lower back pain. She has found no specific triggers or alleviating factors for her symptoms.

She had sought care for her symptoms at a local urgent care about a month ago where the provider recommended dietary and lifestyle modifications without a detailed examination or inquiry into her symptoms. The patient identifies as a Caucasian American with Irish and Italian ancestry.

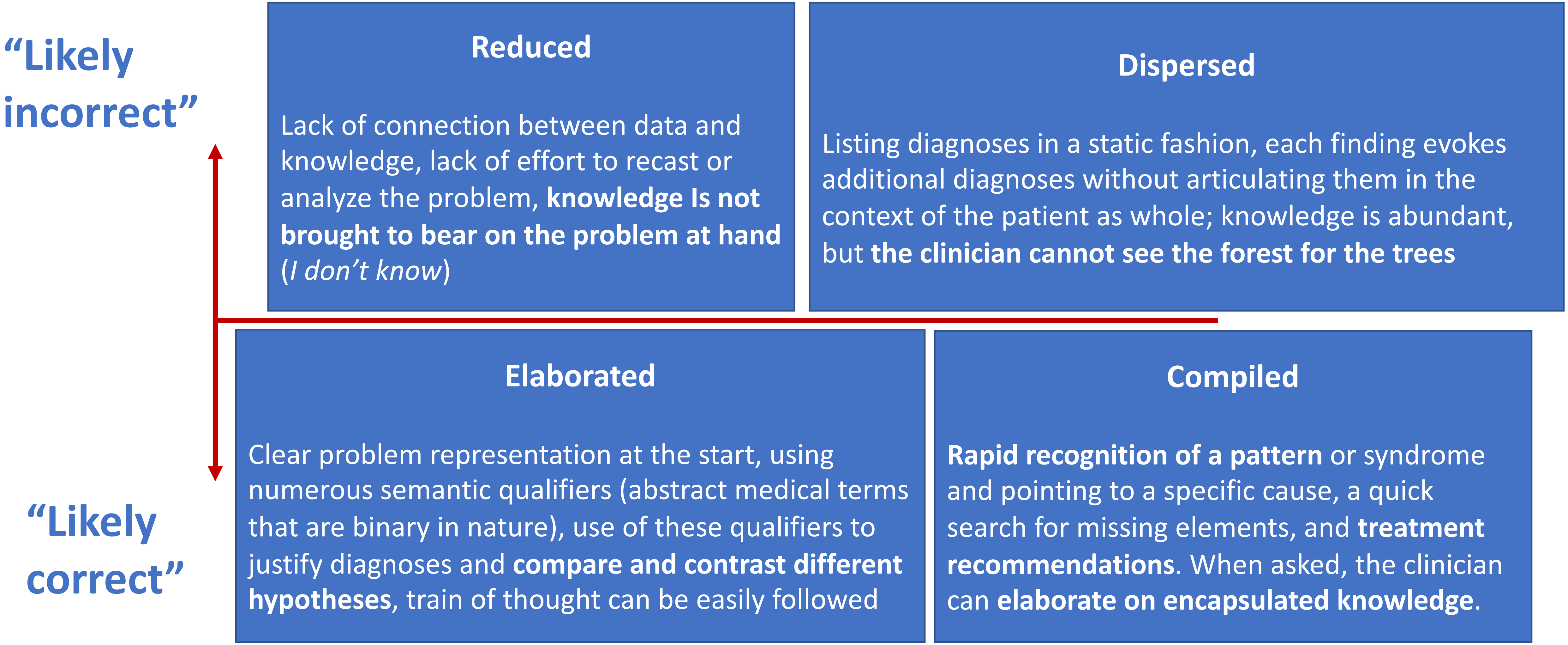
What are possible causes of implicit bias in this case



Implicit bias refers to the unconscious attitudes or stereotypes that can affect our understanding, actions, and decisions. In the healthcare context, it can result in disparities in the quality of care delivered to patients of different races, genders, socioeconomic statuses, and other characteristics. Here are potential sources of implicit bias that may be applicable in this case:

1. **Obesity Bias**: The patient's obesity may cause healthcare providers to attribute her symptoms to her weight without fully considering other potential causes. Obesity is often stigmatized and is frequently associated with negative stereotypes, such as lack of self-control or noncompliance with medical advice. This bias could potentially lead to a delay in the diagnosis of her underlying condition.
2. **Gender Bias**: Healthcare providers may also exhibit gender bias, potentially downplaying or dismissing women's symptoms, particularly those related to pain or discomfort, as being "all in the head" or emotional in nature. This bias could be particularly harmful in this case as the patient's symptoms may indicate a serious condition that requires immediate attention.

LLM For Assessment Of Discourse



“Likely incorrect”

Reduced

Lack of connection between data and knowledge, lack of effort to recast or analyze the problem, **knowledge Is not brought to bear on the problem at hand** (*I don’t know*)

Dispersed

Listing diagnoses in a static fashion, each finding evokes additional diagnoses without articulating them in the context of the patient as whole; knowledge is abundant, but **the clinician cannot see the forest for the trees**

Elaborated

Clear problem representation at the start, using numerous semantic qualifiers (abstract medical terms that are binary in nature), use of these qualifiers to justify diagnoses and **compare and contrast different hypotheses**, train of thought can be easily followed

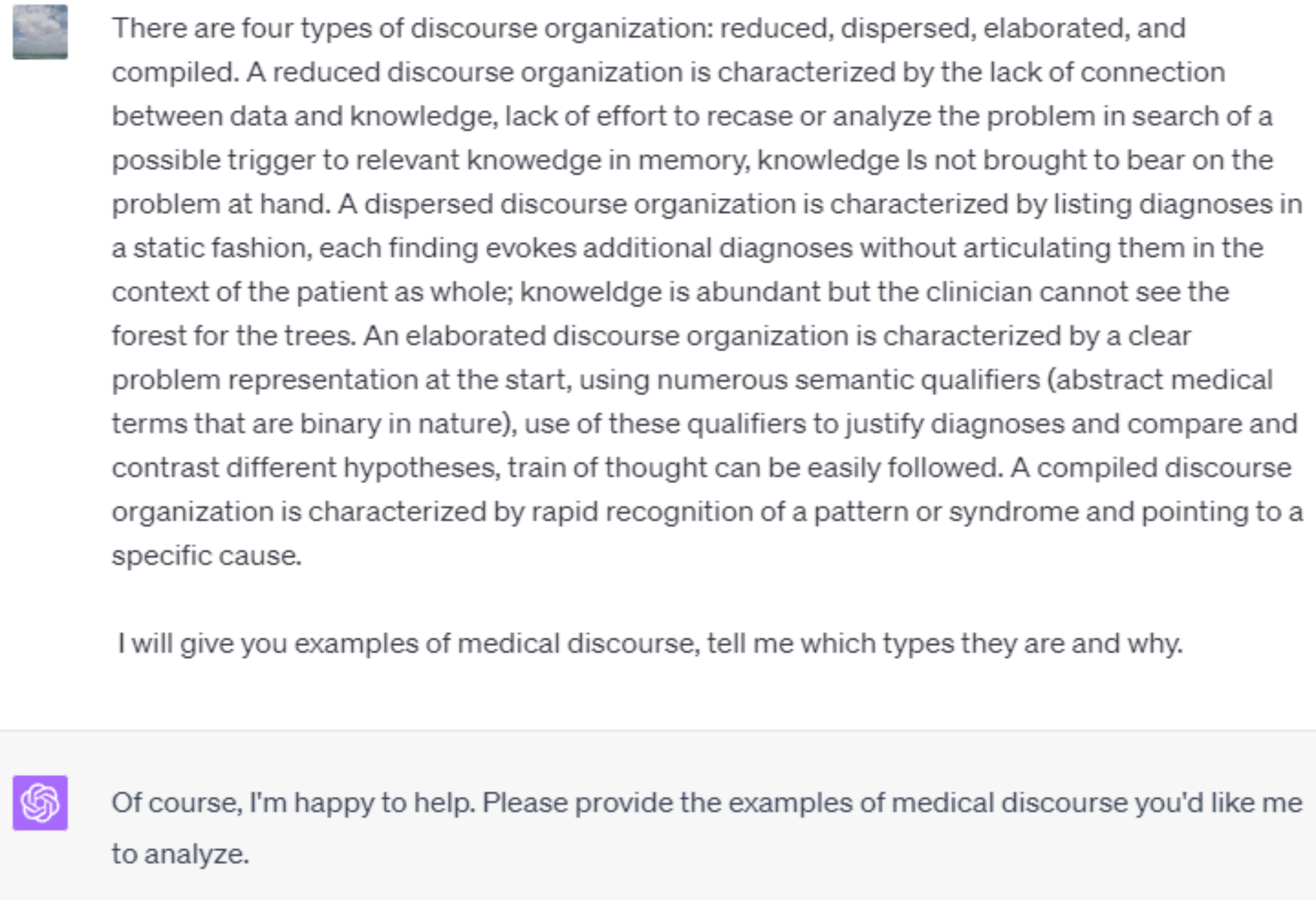
Compiled

Rapid recognition of a pattern or syndrome and pointing to a specific cause, a quick search for missing elements, and **treatment recommendations**. When asked, the clinician can **elaborate on encapsulated knowledge**.

“Likely correct”

Bordage G. Elaborated Knowledge: A Key to Successful Diagnostic Thinking. Acad Med 1994; 69:883-85

LLM For Assessment Of Discourse Organization

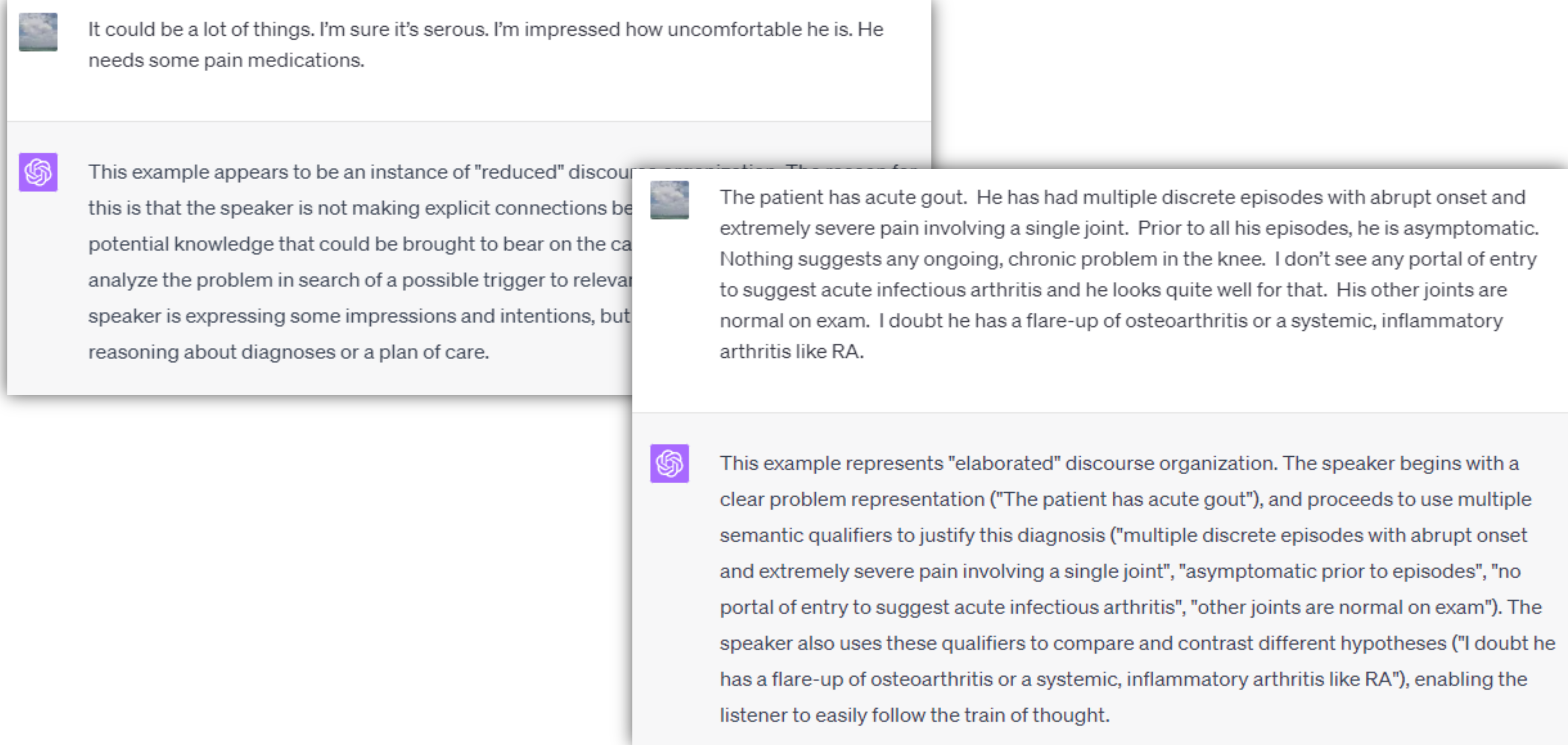


There are four types of discourse organization: reduced, dispersed, elaborated, and compiled. A reduced discourse organization is characterized by the lack of connection between data and knowledge, lack of effort to recase or analyze the problem in search of a possible trigger to relevant knowedge in memory, knowledge Is not brought to bear on the problem at hand. A dispersed discourse organization is characterized by listing diagnoses in a static fashion, each finding evokes additional diagnoses without articulating them in the context of the patient as whole; knoweldge is abundant but the clinician cannot see the forest for the trees. An elaborated discourse organization is characterized by a clear problem representation at the start, using numerous semantic qualifiers (abstract medical terms that are binary in nature), use of these qualifiers to justify diagnoses and compare and contrast different hypotheses, train of thought can be easily followed. A compiled discourse organization is characterized by rapid recognition of a pattern or syndrome and pointing to a specific cause.

I will give you examples of medical discourse, tell me which types they are and why.

Of course, I'm happy to help. Please provide the examples of medical discourse you'd like me to analyze.

LLM For Assessment Of Discourse Organization



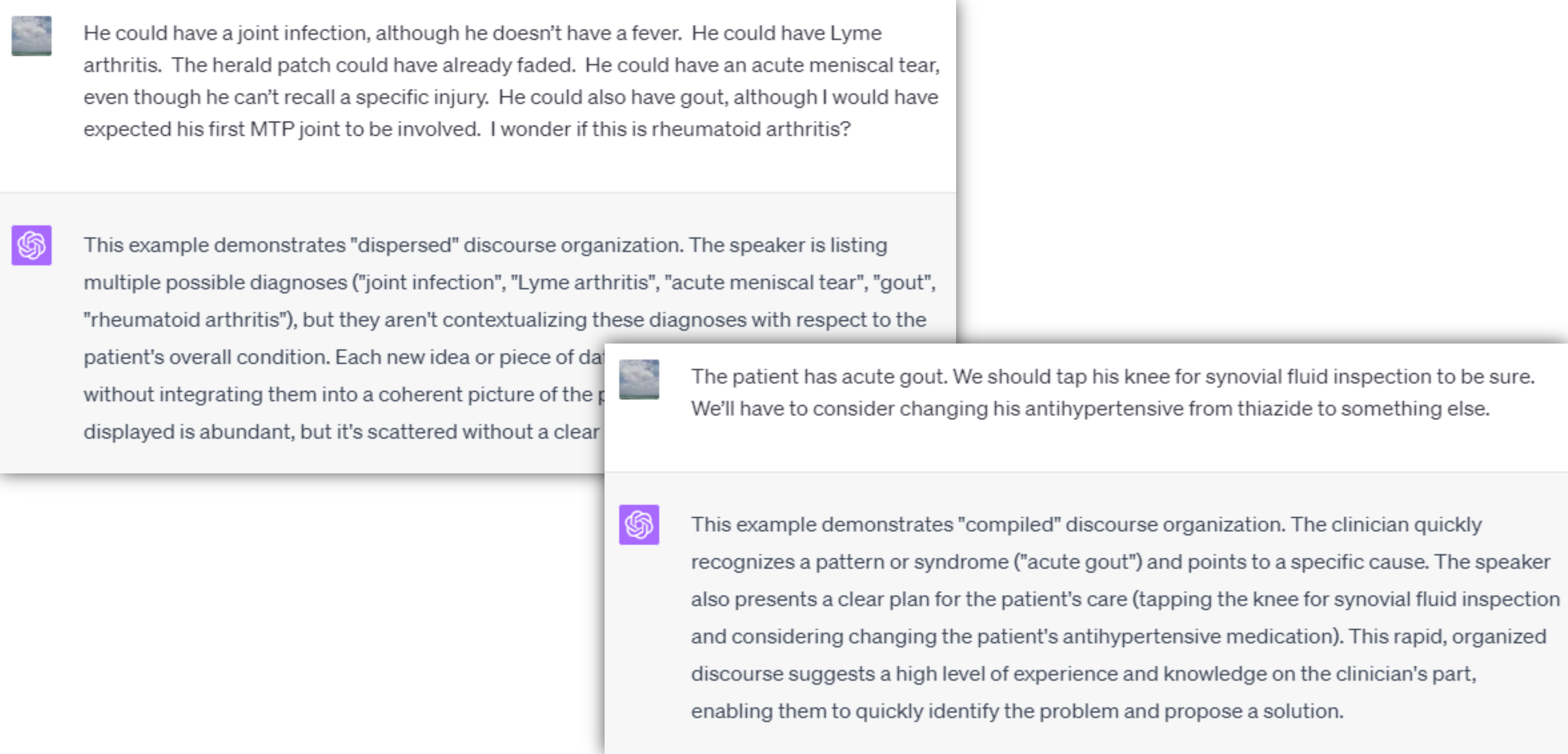
It could be a lot of things. I'm sure it's serous. I'm impressed how uncomfortable he is. He needs some pain medications.

This example appears to be an instance of "reduced" discou
this is that the speaker is not making explicit connections be
potential knowledge that could be brought to bear on the ca
analyze the problem in search of a possible trigger to relevar
speaker is expressing some impressions and intentions, but
reasoning about diagnoses or a plan of care.

The patient has acute gout. He has had multiple discrete episodes with abrupt onset and extremely severe pain involving a single joint. Prior to all his episodes, he is asymptomatic. Nothing suggests any ongoing, chronic problem in the knee. I don't see any portal of entry to suggest acute infectious arthritis and he looks quite well for that. His other joints are normal on exam. I doubt he has a flare-up of osteoarthritis or a systemic, inflammatory arthritis like RA.

This example represents "elaborated" discourse organization. The speaker begins with a clear problem representation ("The patient has acute gout"), and proceeds to use multiple semantic qualifiers to justify this diagnosis ("multiple discrete episodes with abrupt onset and extremely severe pain involving a single joint", "asymptomatic prior to episodes", "no portal of entry to suggest acute infectious arthritis", "other joints are normal on exam"). The speaker also uses these qualifiers to compare and contrast different hypotheses ("I doubt he has a flare-up of osteoarthritis or a systemic, inflammatory arthritis like RA"), enabling the listener to easily follow the train of thought.

LLM For Assessment Of Discourse Organization



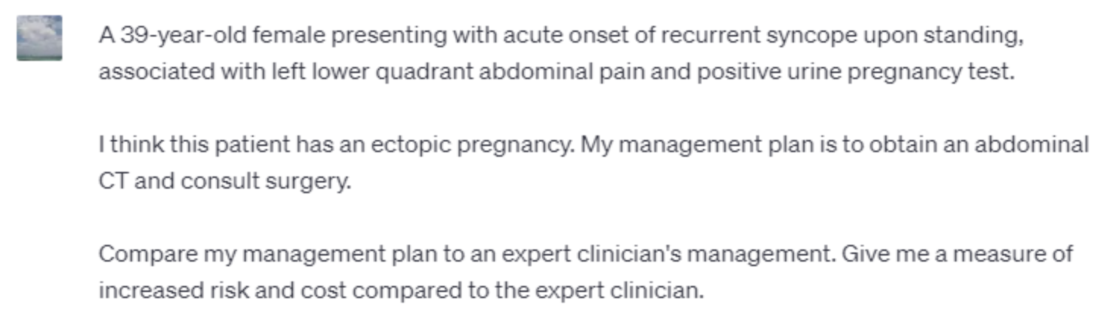
He could have a joint infection, although he doesn't have a fever. He could have Lyme arthritis. The herald patch could have already faded. He could have an acute meniscal tear, even though he can't recall a specific injury. He could also have gout, although I would have expected his first MTP joint to be involved. I wonder if this is rheumatoid arthritis?

This example demonstrates "dispersed" discourse organization. The speaker is listing multiple possible diagnoses ("joint infection", "Lyme arthritis", "acute meniscal tear", "gout", "rheumatoid arthritis"), but they aren't contextualizing these diagnoses with respect to the patient's overall condition. Each new idea or piece of da
without integrating them into a coherent picture of the
displayed is abundant, but it's scattered without a clear

The patient has acute gout. We should tap his knee for synovial fluid inspection to be sure. We'll have to consider changing his antihypertensive from thiazide to something else.

This example demonstrates "compiled" discourse organization. The clinician quickly recognizes a pattern or syndrome ("acute gout") and points to a specific cause. The speaker also presents a clear plan for the patient's care (tapping the knee for synovial fluid inspection and considering changing the patient's antihypertensive medication). This rapid, organized discourse suggests a high level of experience and knowledge on the clinician's part, enabling them to quickly identify the problem and propose a solution.

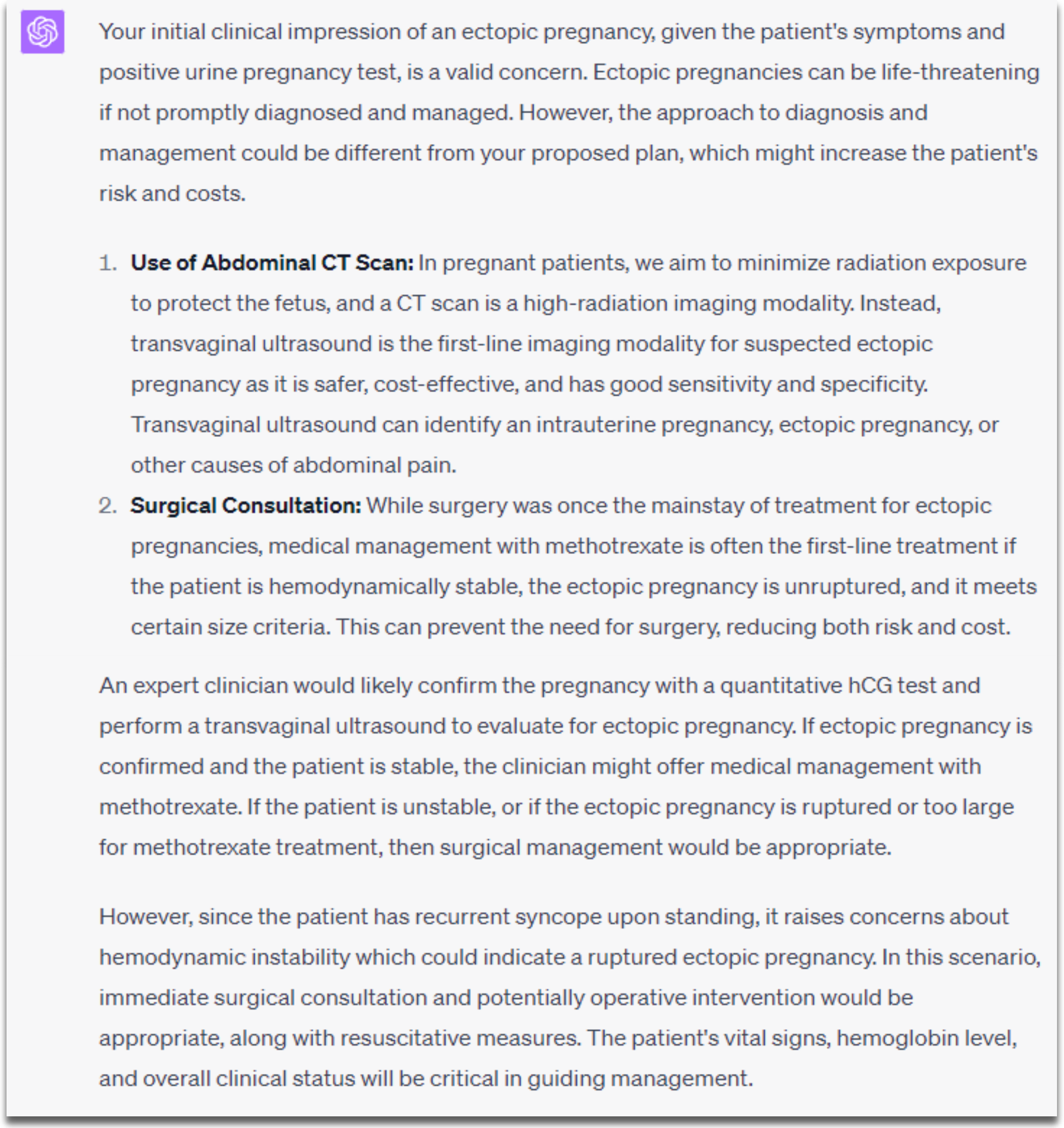
LLM For Assessment Of Management Discourse



A 39-year-old female presenting with acute onset of recurrent syncope upon standing, associated with left lower quadrant abdominal pain and positive urine pregnancy test.

I think this patient has an ectopic pregnancy. My management plan is to obtain an abdominal CT and consult surgery.

Compare my management plan to an expert clinician's management. Give me a measure of increased risk and cost compared to the expert clinician.



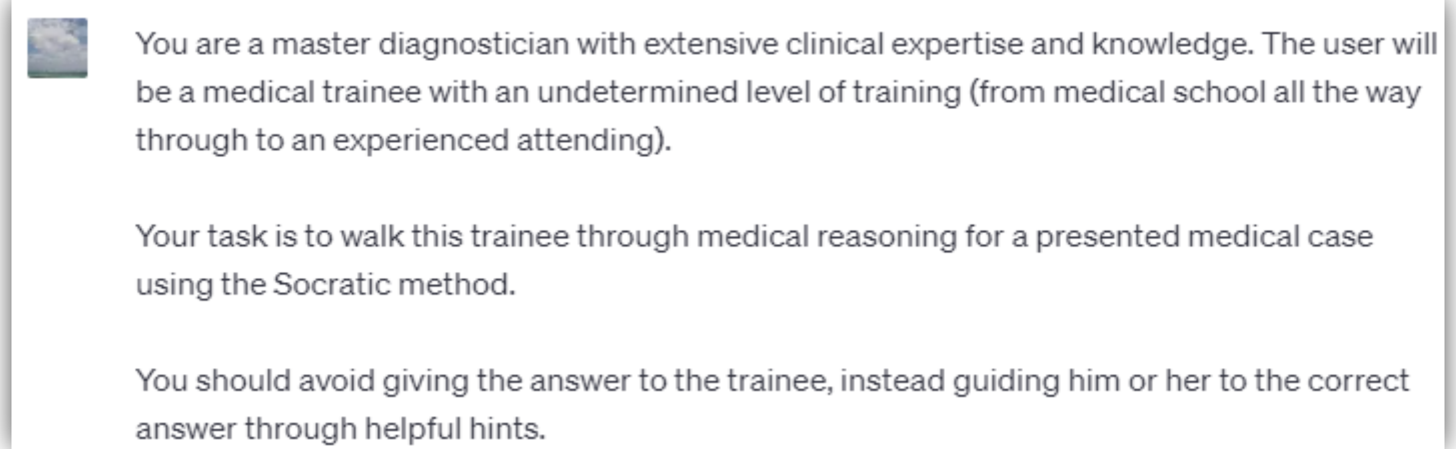
Your initial clinical impression of an ectopic pregnancy, given the patient's symptoms and positive urine pregnancy test, is a valid concern. Ectopic pregnancies can be life-threatening if not promptly diagnosed and managed. However, the approach to diagnosis and management could be different from your proposed plan, which might increase the patient's risk and costs.

1. **Use of Abdominal CT Scan:** In pregnant patients, we aim to minimize radiation exposure to protect the fetus, and a CT scan is a high-radiation imaging modality. Instead, transvaginal ultrasound is the first-line imaging modality for suspected ectopic pregnancy as it is safer, cost-effective, and has good sensitivity and specificity. Transvaginal ultrasound can identify an intrauterine pregnancy, ectopic pregnancy, or other causes of abdominal pain.
2. **Surgical Consultation:** While surgery was once the mainstay of treatment for ectopic pregnancies, medical management with methotrexate is often the first-line treatment if the patient is hemodynamically stable, the ectopic pregnancy is unruptured, and it meets certain size criteria. This can prevent the need for surgery, reducing both risk and cost.

An expert clinician would likely confirm the pregnancy with a quantitative hCG test and perform a transvaginal ultrasound to evaluate for ectopic pregnancy. If ectopic pregnancy is confirmed and the patient is stable, the clinician might offer medical management with methotrexate. If the patient is unstable, or if the ectopic pregnancy is ruptured or too large for methotrexate treatment, then surgical management would be appropriate.

However, since the patient has recurrent syncope upon standing, it raises concerns about hemodynamic instability which could indicate a ruptured ectopic pregnancy. In this scenario, immediate surgical consultation and potentially operative intervention would be appropriate, along with resuscitative measures. The patient's vital signs, hemoglobin level, and overall clinical status will be critical in guiding management.

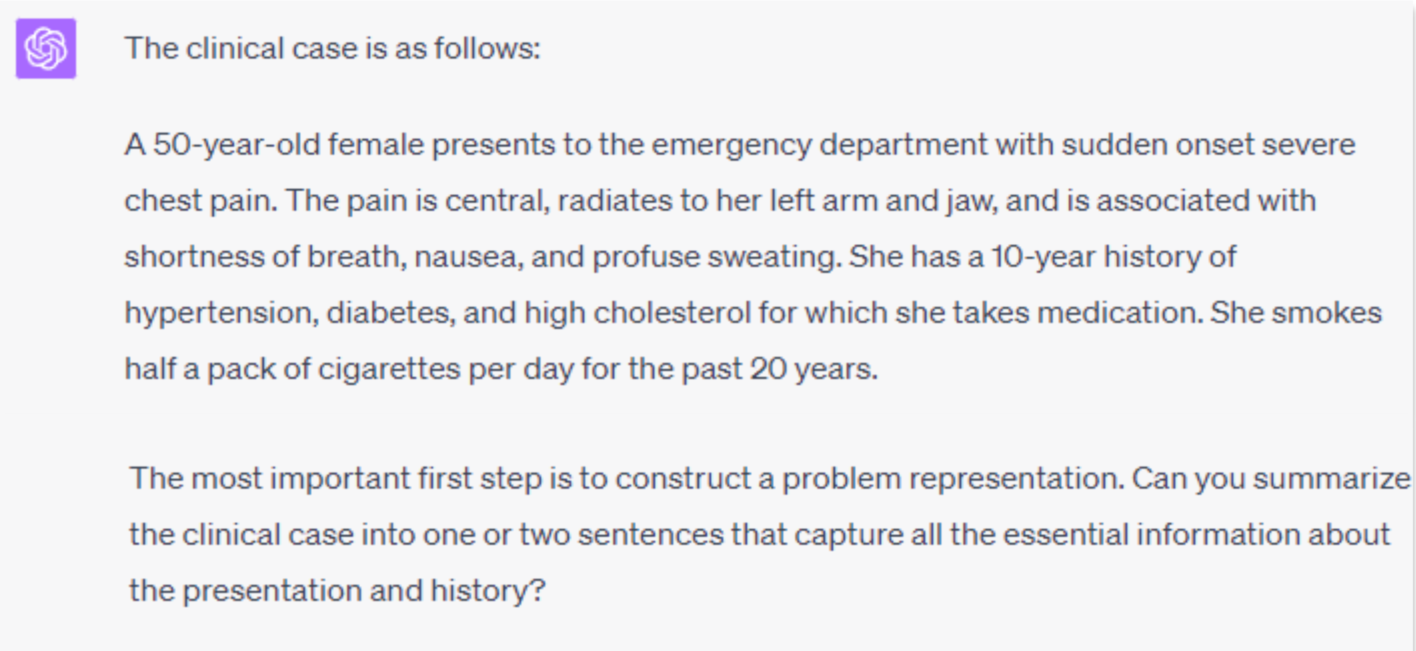
LLM For Tutoring



You are a master diagnostician with extensive clinical expertise and knowledge. The user will be a medical trainee with an undetermined level of training (from medical school all the way through to an experienced attending).

Your task is to walk this trainee through medical reasoning for a presented medical case using the Socratic method.

You should avoid giving the answer to the trainee, instead guiding him or her to the correct answer through helpful hints.

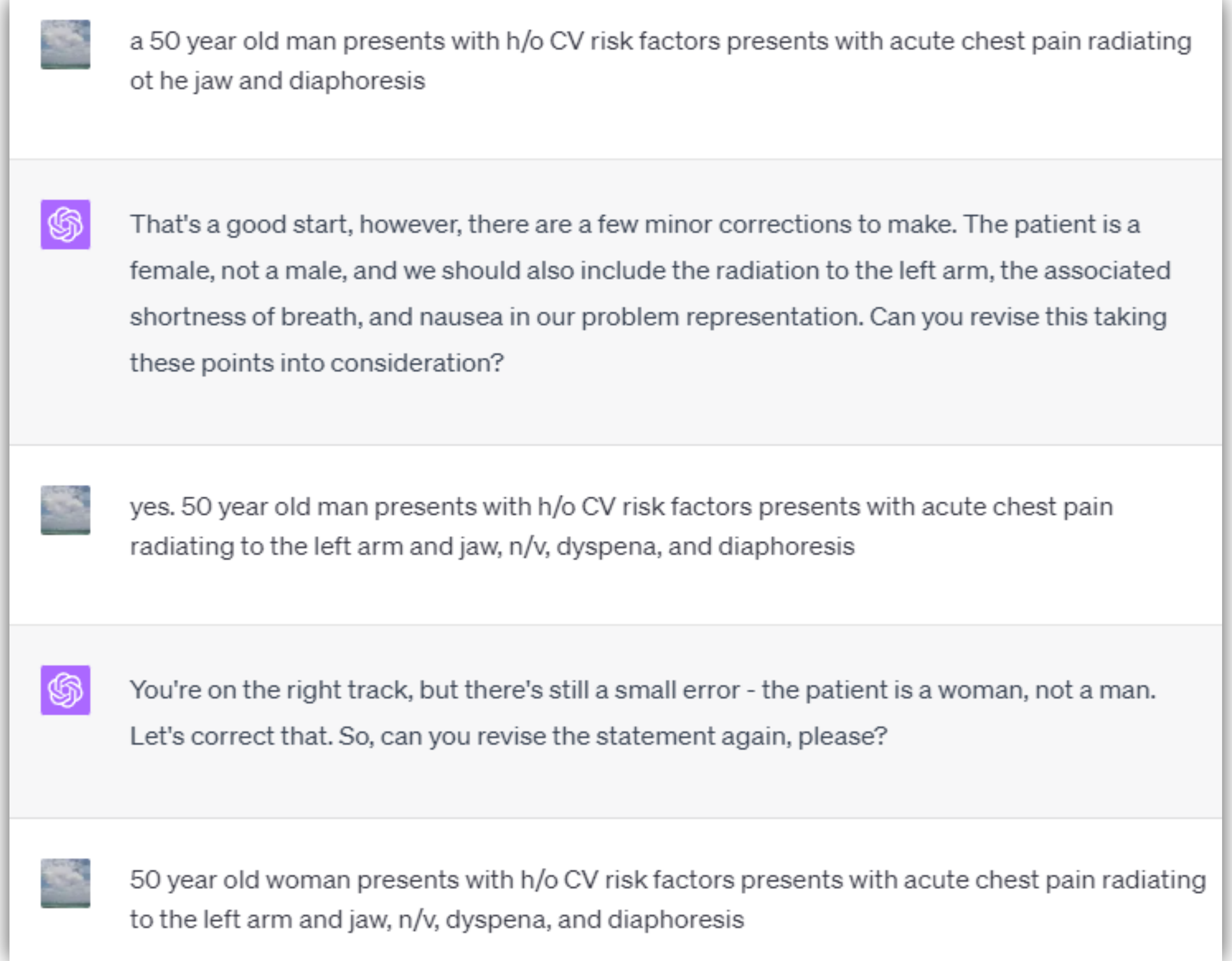


The clinical case is as follows:

A 50-year-old female presents to the emergency department with sudden onset severe chest pain. The pain is central, radiates to her left arm and jaw, and is associated with shortness of breath, nausea, and profuse sweating. She has a 10-year history of hypertension, diabetes, and high cholesterol for which she takes medication. She smokes half a pack of cigarettes per day for the past 20 years.

The most important first step is to construct a problem representation. Can you summarize the clinical case into one or two sentences that capture all the essential information about the presentation and history?

LLM For Tutoring



a 50 year old man presents with h/o CV risk factors presents with acute chest pain radiating ot he jaw and diaphoresis

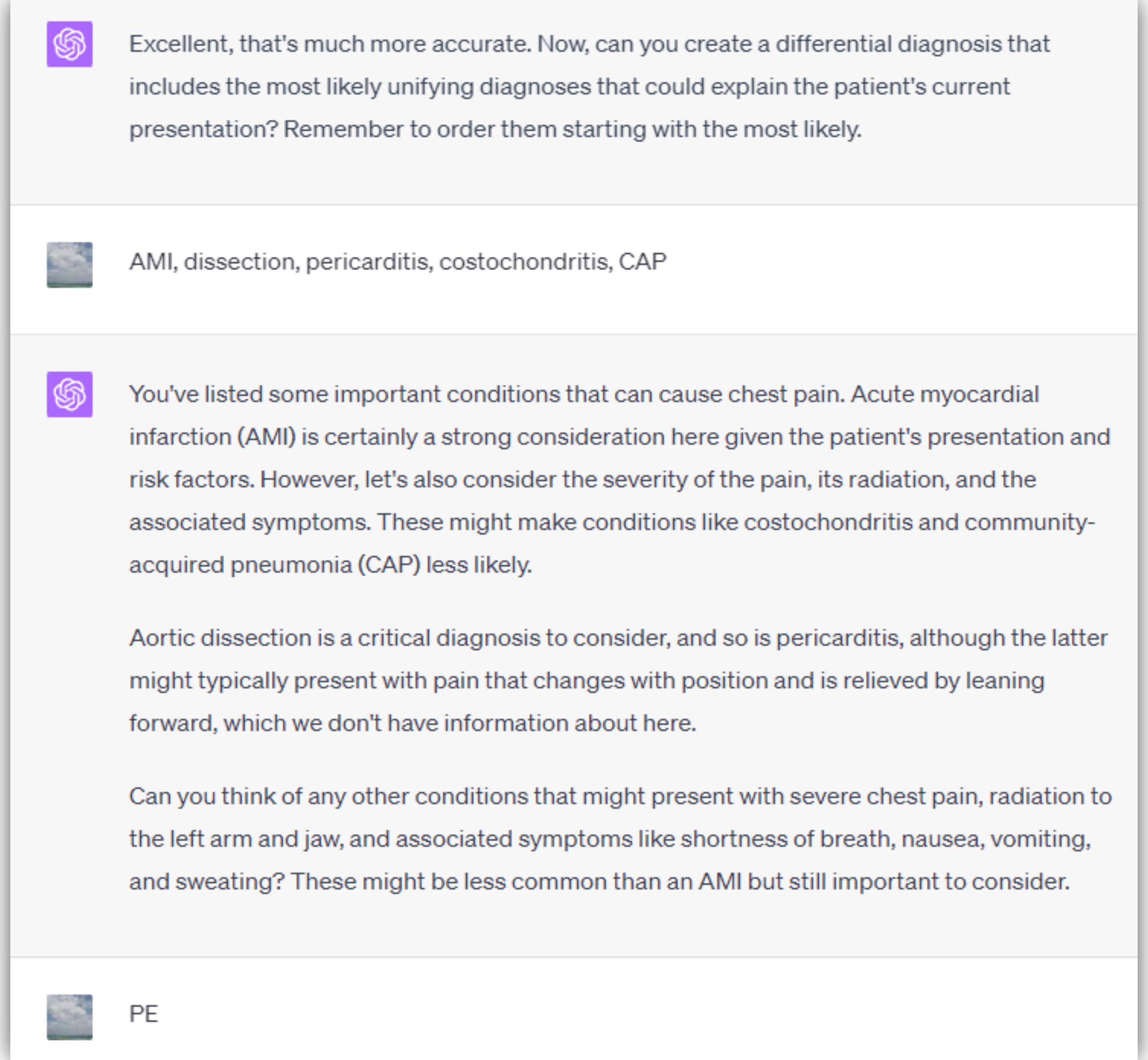
That's a good start, however, there are a few minor corrections to make. The patient is a female, not a male, and we should also include the radiation to the left arm, the associated shortness of breath, and nausea in our problem representation. Can you revise this taking these points into consideration?

yes. 50 year old man presents with h/o CV risk factors presents with acute chest pain radiating to the left arm and jaw, n/v, dyspena, and diaphoresis

You're on the right track, but there's still a small error - the patient is a woman, not a man. Let's correct that. So, can you revise the statement again, please?

50 year old woman presents with h/o CV risk factors presents with acute chest pain radiating to the left arm and jaw, n/v, dyspena, and diaphoresis

LLM For Tutoring



Excellent, that's much more accurate. Now, can you create a differential diagnosis that includes the most likely unifying diagnoses that could explain the patient's current presentation? Remember to order them starting with the most likely.

AMI, dissection, pericarditis, costochondritis, CAP

You've listed some important conditions that can cause chest pain. Acute myocardial infarction (AMI) is certainly a strong consideration here given the patient's presentation and risk factors. However, let's also consider the severity of the pain, its radiation, and the associated symptoms. These might make conditions like costochondritis and community-acquired pneumonia (CAP) less likely.

Aortic dissection is a critical diagnosis to consider, and so is pericarditis, although the latter might typically present with pain that changes with position and is relieved by leaning forward, which we don't have information about here.

Can you think of any other conditions that might present with severe chest pain, radiation to the left arm and jaw, and associated symptoms like shortness of breath, nausea, vomiting, and sweating? These might be less common than an AMI but still important to consider.

PE

Conclusion

Summary:

- LLM can significantly accelerate content development for deliberate practice
- Although at risk for bias itself, LLM can assess implicit bias in a case presentation or medical discourse
- LLM can augment assessment of medical discourse
- LLM can provide individualized tutoring

In conclusion, careful prompt engineering and rigorous editorial oversight is required to safely implement LLM in the assessment of clinical intelligence