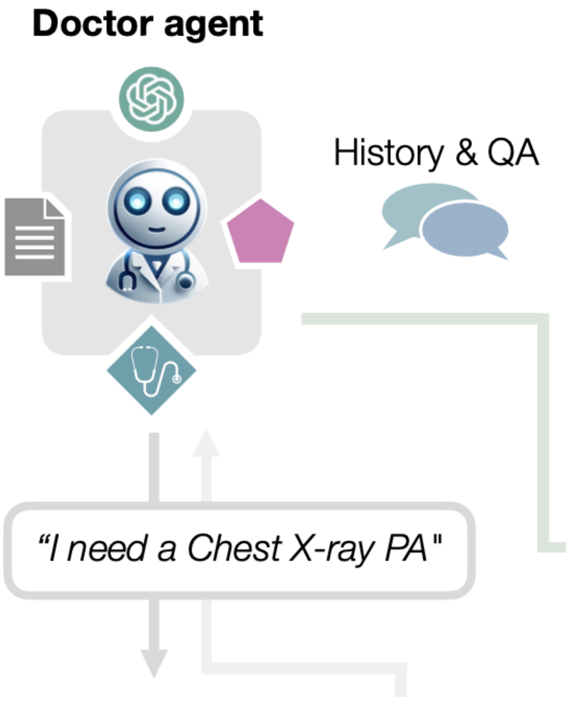
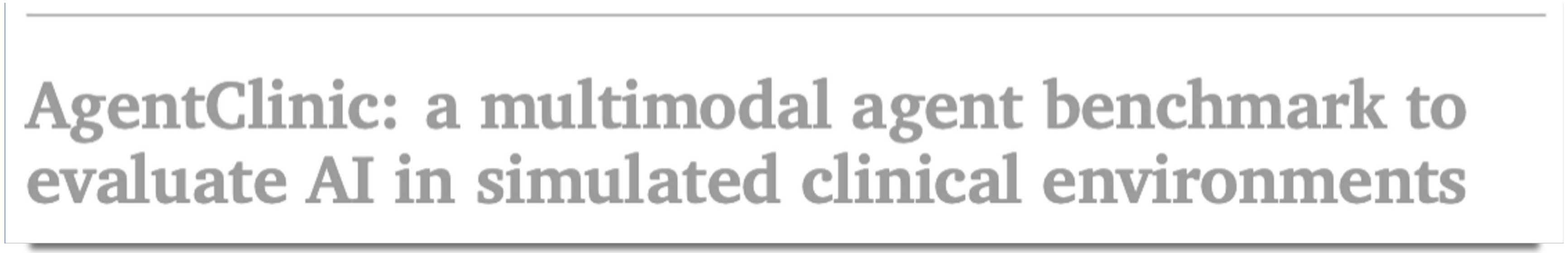
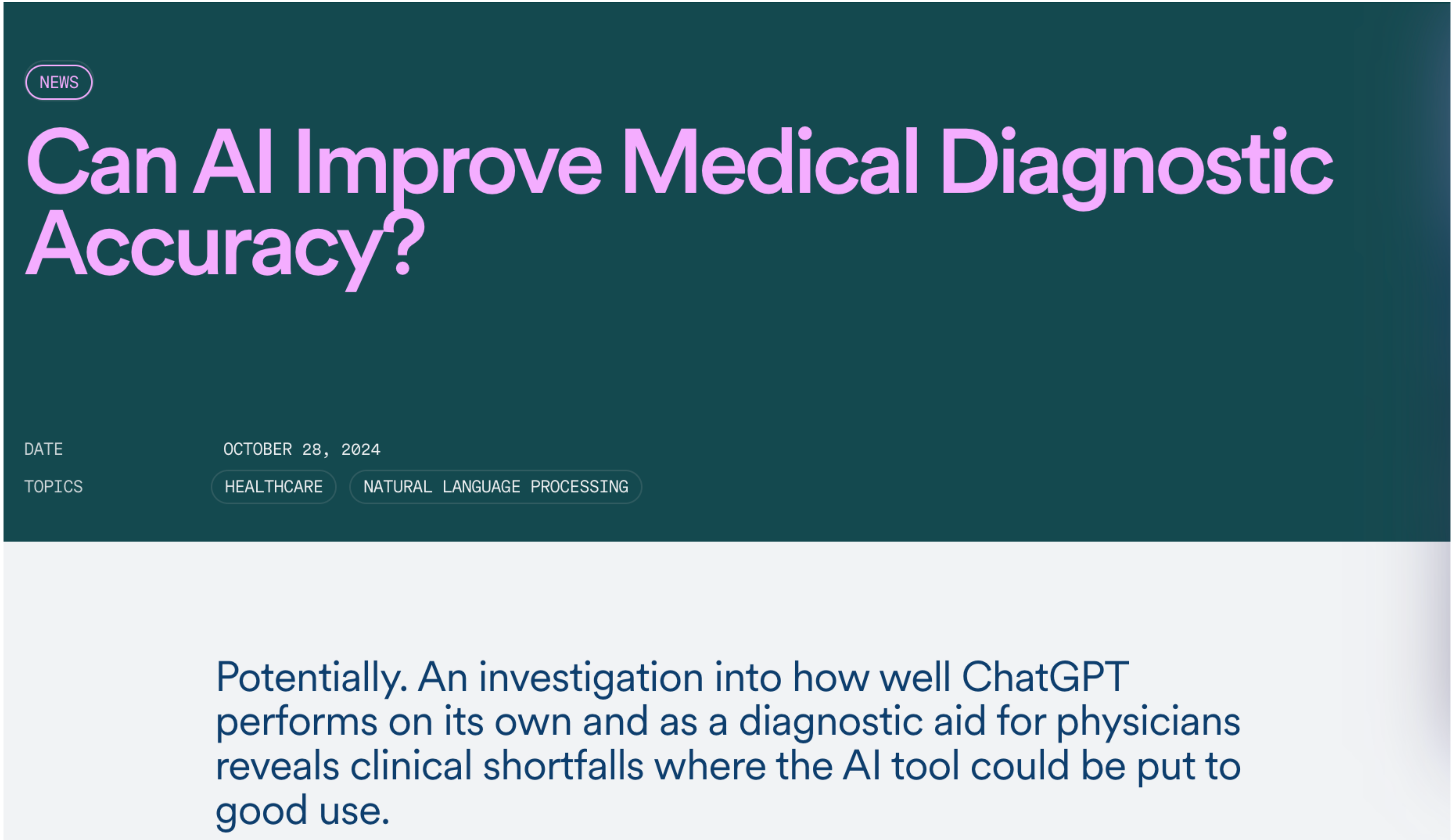


# Agentic AI Use Case: Clinical Diagnosis

Can AI improve clinical diagnosis accuracy?


AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments



NEWS

# Can AI Improve Medical Diagnostic Accuracy?

DATE OCTOBER 28, 2024

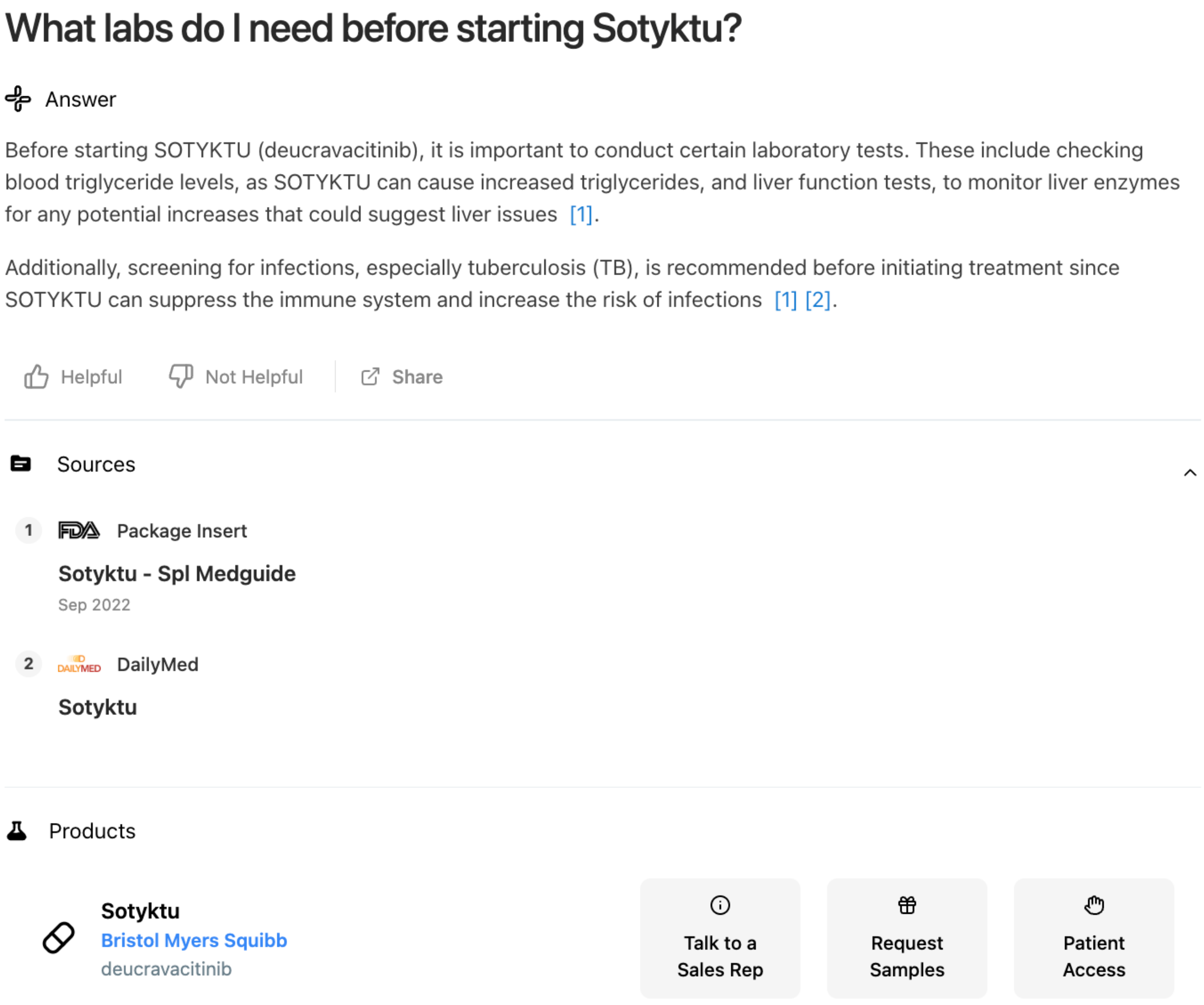
TOPICS HEALTHCARE NATURAL LANGUAGE PROCESSING

Potentially. An investigation into how well ChatGPT performs on its own and as a diagnostic aid for physicians reveals clinical shortfalls where the AI tool could be put to good use.

# Advancing medical AI with Med-Gemini, our family of research models fine-tuned for the medical domain

Our research on Med-Gemini holds the potential to support researchers, developers and specialists in fields like radiology, dermatology, and pathology, changing how complex medical information is synthesized and applied in real-world settings. Learn more

ReachRx



# What labs do I need before starting Sotyktu?

Answer

Before starting SOTYKTU (deucravacitinib), it is important to conduct certain laboratory tests. These include checking blood triglyceride levels, as SOTYKTU can cause increased triglycerides, and liver function tests, to monitor liver enzymes for any potential increases that could suggest liver issues [1].

Additionally, screening for infections, especially tuberculosis (TB), is recommended before initiating treatment since SOTYKTU can suppress the immune system and increase the risk of infections [1] [2].

Sources

1 FDA Package Insert
**Sotyktu - Spl Medguide**
Sep 2022

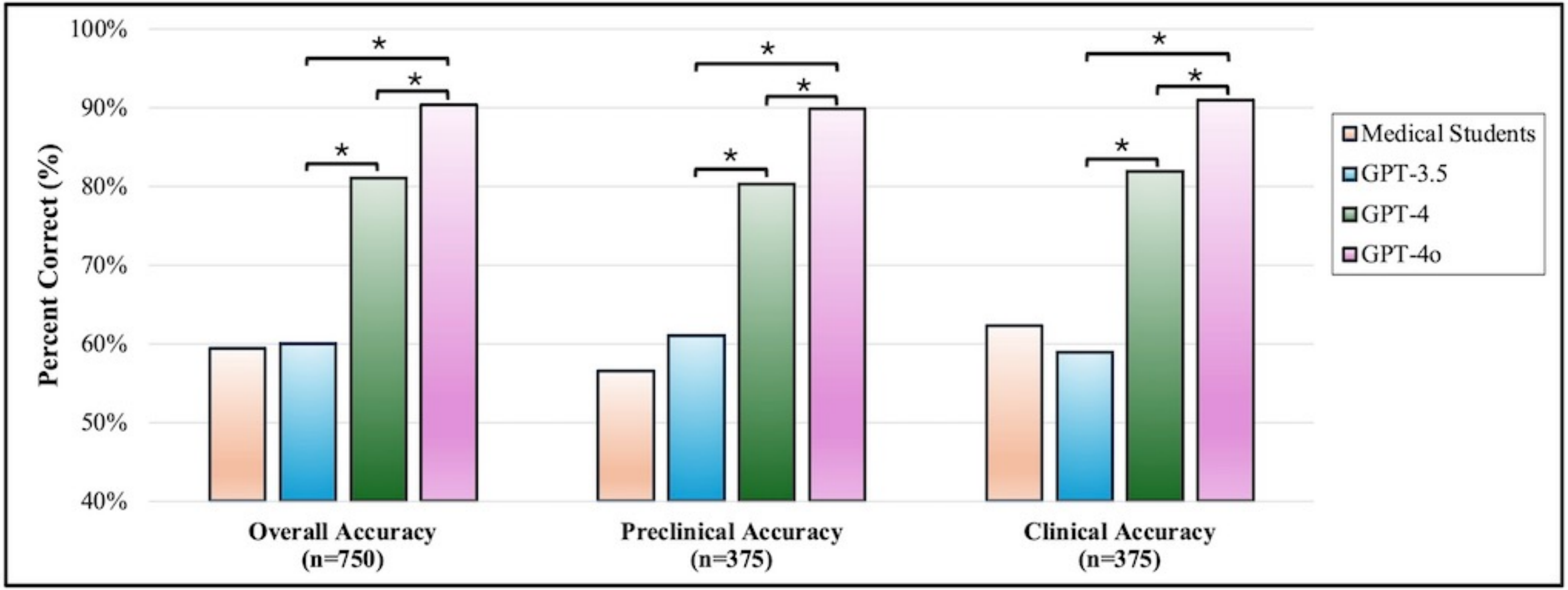
2 DailyMed
**Sotyktu**

Products

**Sotyktu**
Bristol Myers Squibb
deucravacitinib

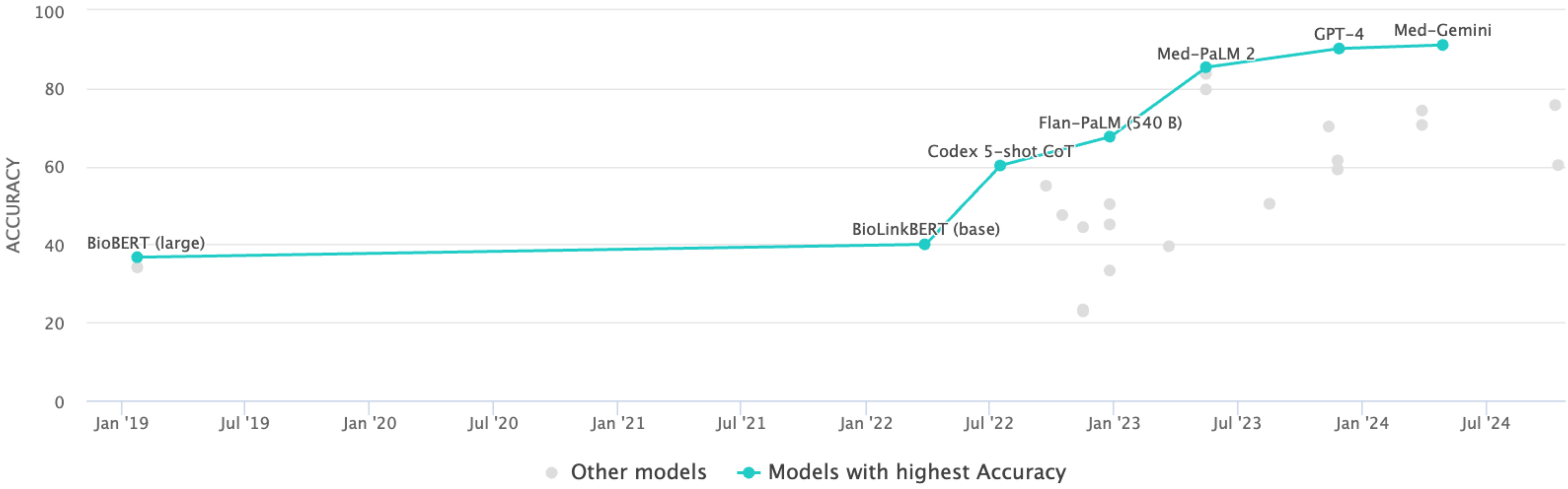
# LLM for Medicine

LLMs have quickly surpassed the average human score on the United States Medical Licensing Exam (USMLE) in a short amount of time, from 38.1% in September 2021 to 90.4% in June 2024 (human passing score is 60%, human expert score is 87%).

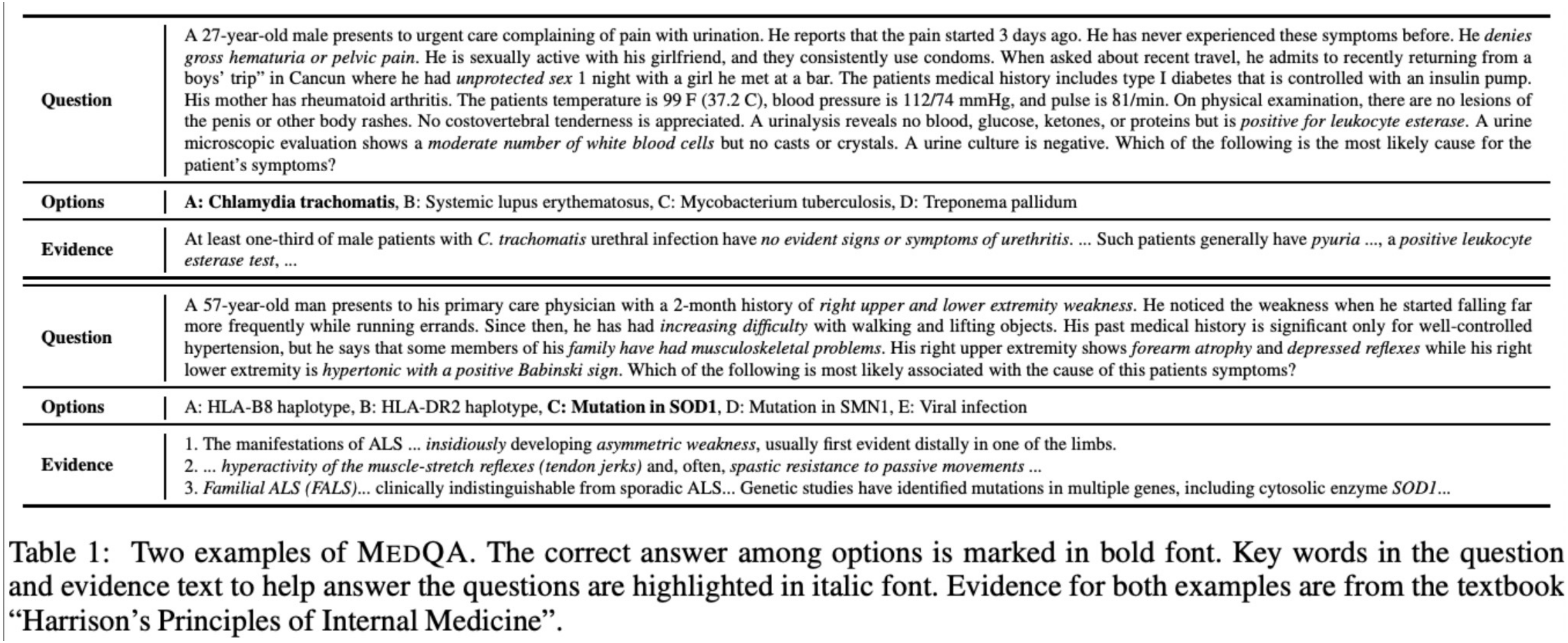


# LLM for Medicine

MedQA Accuracy improves from 36.7% in 2019 to 91.1% in 2024!



# MedQA



| | |
|---|---|
| **Question** | A 27-year-old male presents to urgent care complaining of pain with urination. He reports that the pain started 3 days ago. He has never experienced these symptoms before. He *denies gross hematuria or pelvic pain*. He is sexually active with his girlfriend, and they consistently use condoms. When asked about recent travel, he admits to recently returning from a boys' trip" in Cancun where he had *unprotected sex* 1 night with a girl he met at a bar. The patients medical history includes type I diabetes that is controlled with an insulin pump. His mother has rheumatoid arthritis. The patients temperature is 99 F (37.2 C), blood pressure is 112/74 mmHg, and pulse is 81/min. On physical examination, there are no lesions of the penis or other body rashes. No costovertebral tenderness is appreciated. A urinalysis reveals no blood, glucose, ketones, or proteins but is *positive for leukocyte esterase*. A urine microscopic evaluation shows a *moderate number of white blood cells* but no casts or crystals. A urine culture is negative. Which of the following is the most likely cause for the patient's symptoms? |
| **Options** | **A: Chlamydia trachomatis**, B: Systemic lupus erythematosus, C: Mycobacterium tuberculosis, D: Treponema pallidum |
| **Evidence** | At least one-third of male patients with *C. trachomatis* urethral infection have *no evident signs or symptoms of urethritis*. ... Such patients generally have *pyuria* ..., a *positive leukocyte esterase test*, ... |
| **Question** | A 57-year-old man presents to his primary care physician with a 2-month history of *right upper and lower extremity weakness*. He noticed the weakness when he started falling far more frequently while running errands. Since then, he has had *increasing difficulty* with walking and lifting objects. His past medical history is significant only for well-controlled hypertension, but he says that some members of his *family have had musculoskeletal problems*. His right upper extremity shows *forearm atrophy* and *depressed reflexes* while his right lower extremity is *hypertonic with a positive Babinski sign*. Which of the following is most likely associated with the cause of this patients symptoms? |
| **Options** | A: HLA-B8 haplotype, B: HLA-DR2 haplotype, **C: Mutation in SOD1**, D: Mutation in SMN1, E: Viral infection |
| **Evidence** | 1. The manifestations of ALS ... *insidiously* developing *asymmetric weakness*, usually first evident distally in one of the limbs.<br>2. ... *hyperactivity of the muscle-stretch reflexes (tendon jerks)* and, often, *spastic resistance to passive movements* ...<br>3. *Familial ALS (FALS)*... clinically indistinguishable from sporadic ALS... Genetic studies have identified mutations in multiple genes, including cytosolic enzyme *SOD1*... |

Table 1: Two examples of MEDQA. The correct answer among options is marked in bold font. Key words in the question and evidence text to help answer the questions are highlighted in italic font. Evidence for both examples are from the textbook "Harrison's Principles of Internal Medicine".

# What is the Limitation of existing benchmarks?

Clinical work is a multiplexed task that involves **sequential decision making**, requiring the doctor to **handle uncertainty with limited information and finite resources** while compassionately taking care of patients and obtaining relevant information from them.

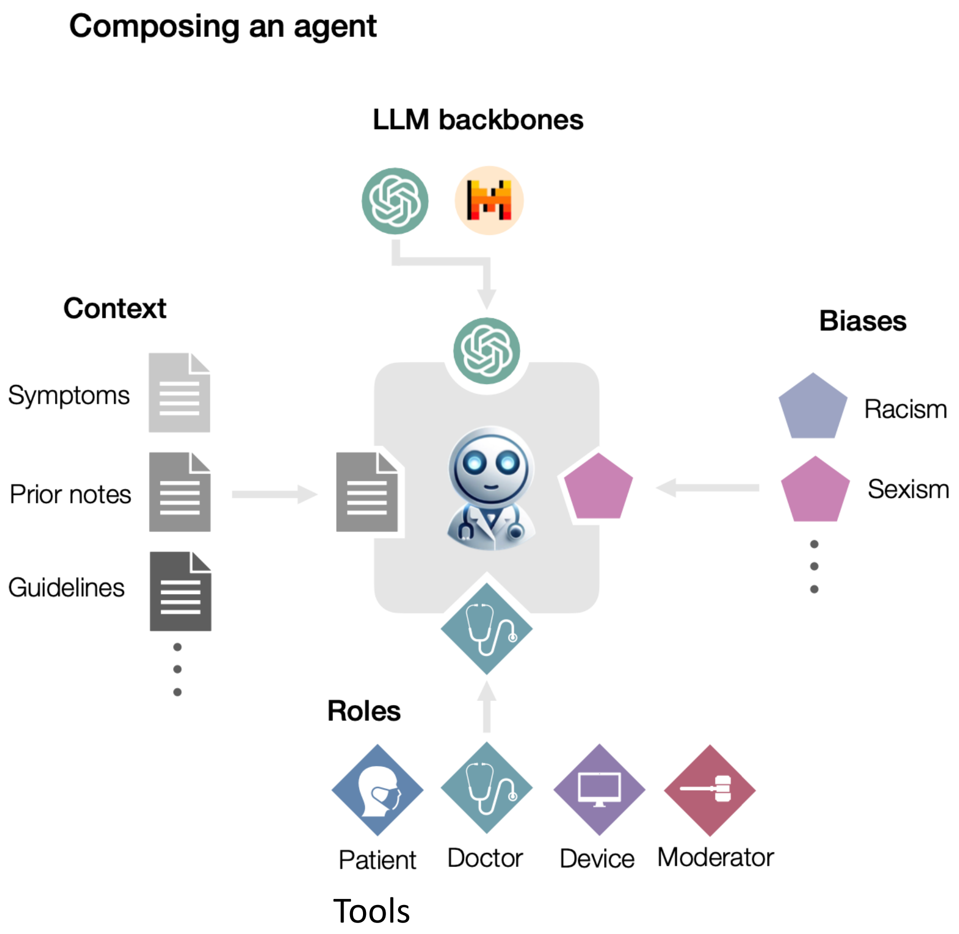
# Overcome limitation

**New benchmark for Medical diagnosis:**

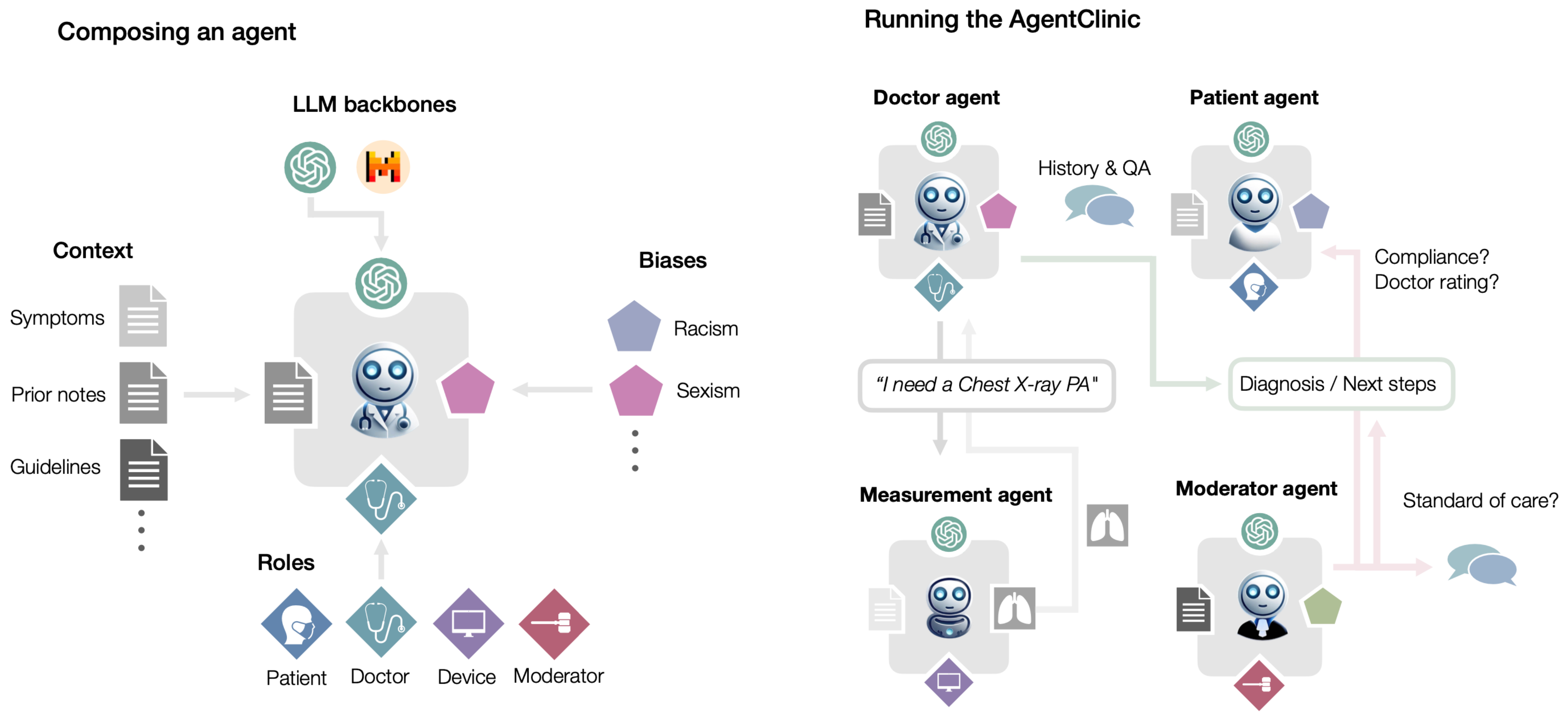
- interactive, dialogue-driven
- sequential decision making environments that require
  - patient data collection: symptoms, lifestyle, history
  - ordering appropriate medical exams
  - understanding medical images

AgentClinic: a multimodal agent benchmark for clinical decision making

# AgentClinic: a multimodal agent benchmark for clinical decision making



# AgentClinic: a multimodal agent benchmark for clinical decision making



**Composing an agent**

**LLM backbones**

**Context**

Symptoms

Prior notes

Guidelines

**Biases**

Racism

Sexism

**Roles**

Patient

Doctor

Device

Moderator

**Running the AgentClinic**

**Doctor agent**

**Patient agent**

History & QA

Compliance?
Doctor rating?

*"I need a Chest X-ray PA"*

Diagnosis / Next steps

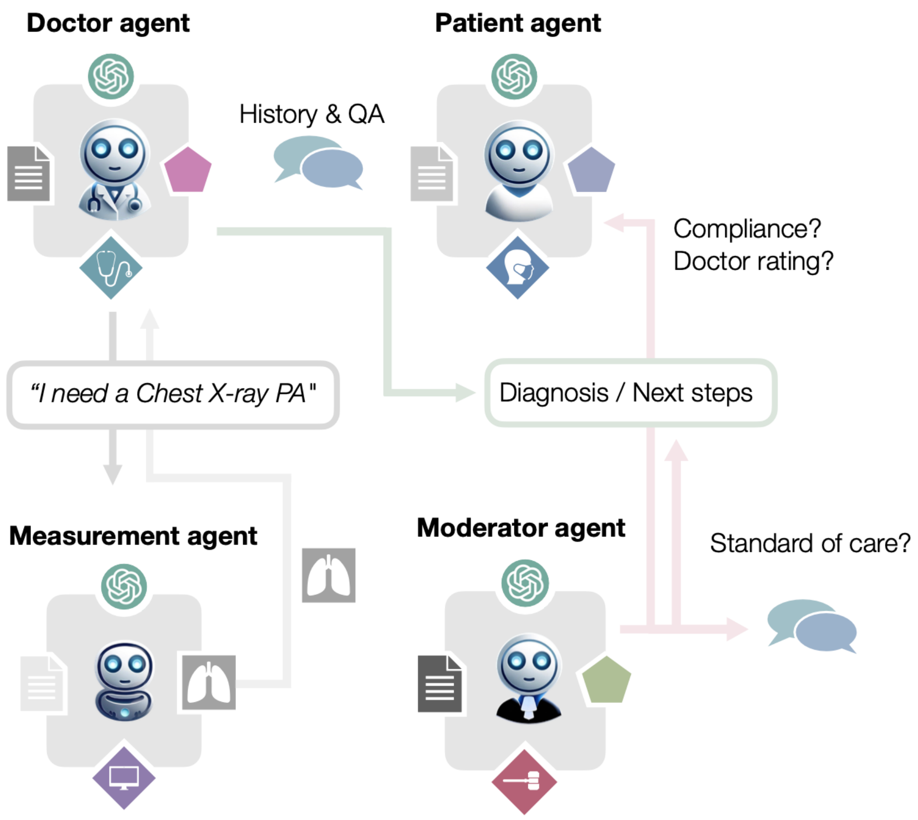
**Measurement agent**

**Moderator agent**

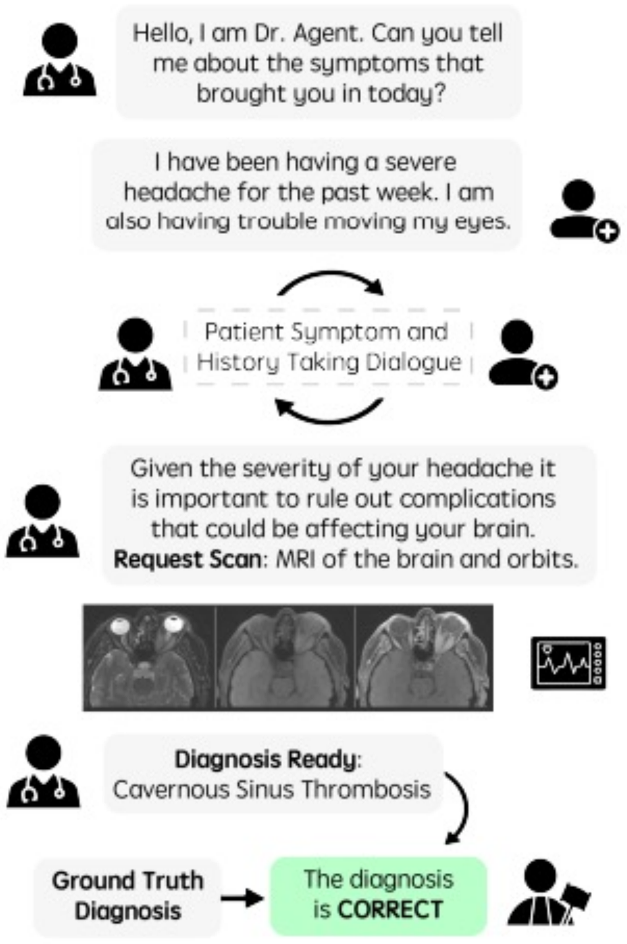
Standard of care?

# AgentClinic: a multimodal agent benchmark for clinical decision making

**Running the AgentClinic**



Example AgentClinic **Simulation**



Hello, I am Dr. Agent. Can you tell me about the symptoms that brought you in today?

I have been having a severe headache for the past week. I am also having trouble moving my eyes.

Patient Symptom and History Taking Dialogue

Given the severity of your headache it is important to rule out complications that could be affecting your brain. **Request Scan**: MRI of the brain and orbits.

**Diagnosis Ready**: Cavernous Sinus Thrombosis

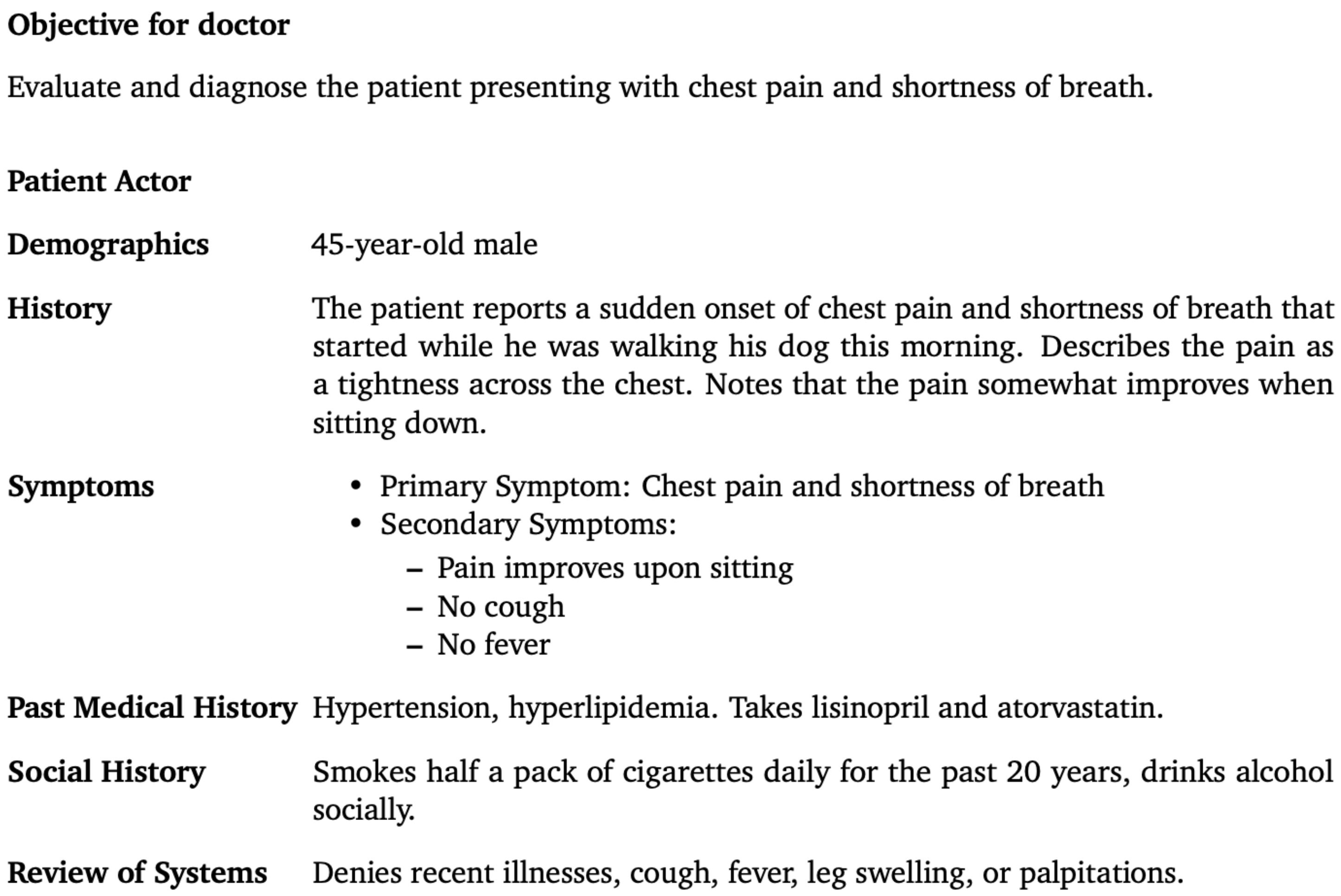
**Ground Truth Diagnosis** → The diagnosis is **CORRECT**

# Build agents for AgentClinic

- Agents built from real clinical cases sourced from:
  - US Medical Licensing Exam
  - Deidentified electronic health records (MIMIC-IV)
  - New England Journal of Medicine (NEJM) case challenges

- We separate information by what is provided to each agent, including
  - the objective for the doctor
  - patient history and symptoms for the patient
  - physical examination findings for the measurement
  - the correct diagnosis for the moderator.

# Objective Structured Clinical Examination (OSCE) Case Study from MedQA



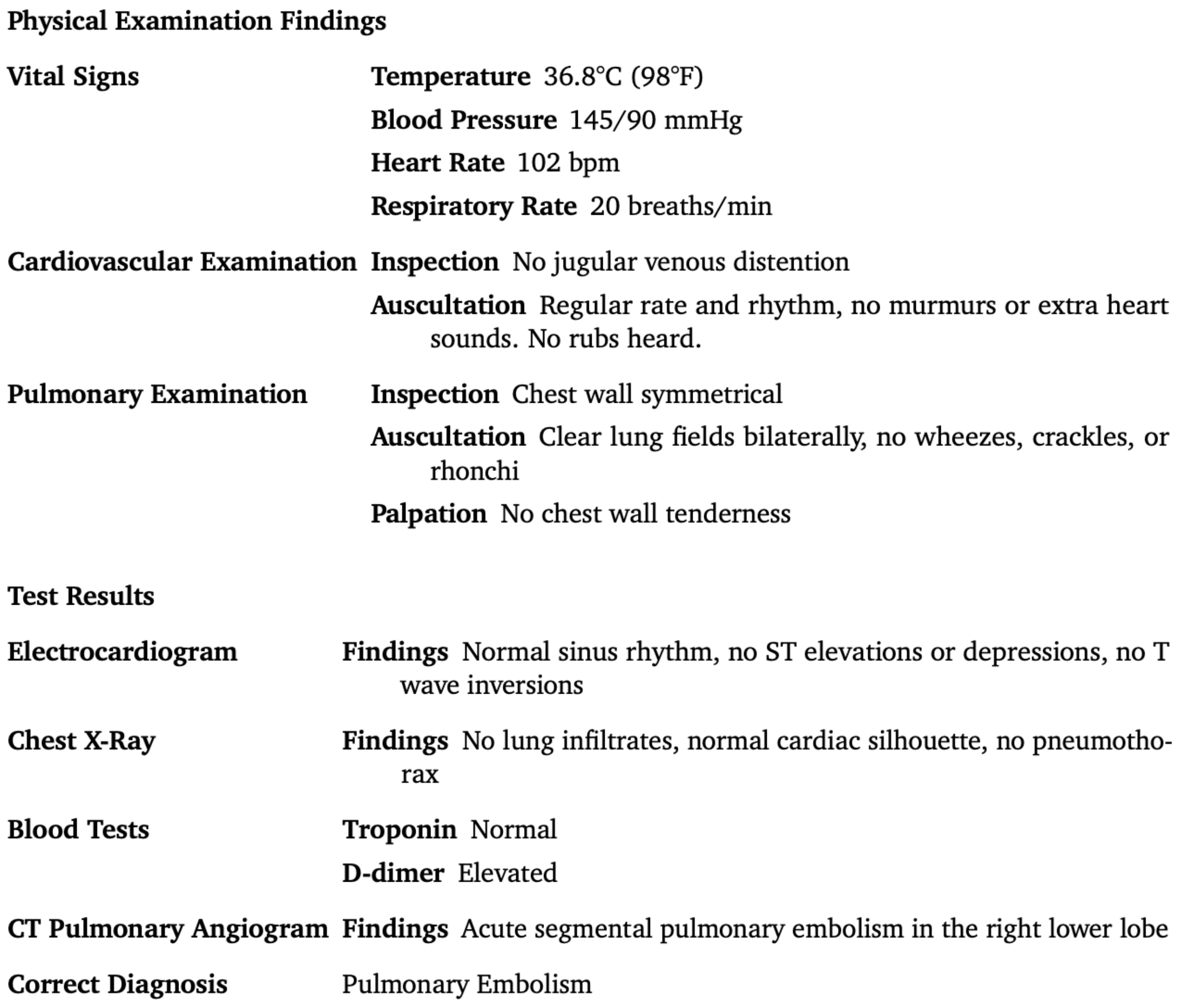
**Objective for doctor**

Evaluate and diagnose the patient presenting with chest pain and shortness of breath.

**Patient Actor**

| | |
|---|---|
| **Demographics** | 45-year-old male |
| **History** | The patient reports a sudden onset of chest pain and shortness of breath that started while he was walking his dog this morning. Describes the pain as a tightness across the chest. Notes that the pain somewhat improves when sitting down. |
| **Symptoms** | • Primary Symptom: Chest pain and shortness of breath<br>• Secondary Symptoms:<br>– Pain improves upon sitting<br>– No cough<br>– No fever |
| **Past Medical History** | Hypertension, hyperlipidemia. Takes lisinopril and atorvastatin. |
| **Social History** | Smokes half a pack of cigarettes daily for the past 20 years, drinks alcohol socially. |
| **Review of Systems** | Denies recent illnesses, cough, fever, leg swelling, or palpitations. |

# OSCE Case Study from MedQA



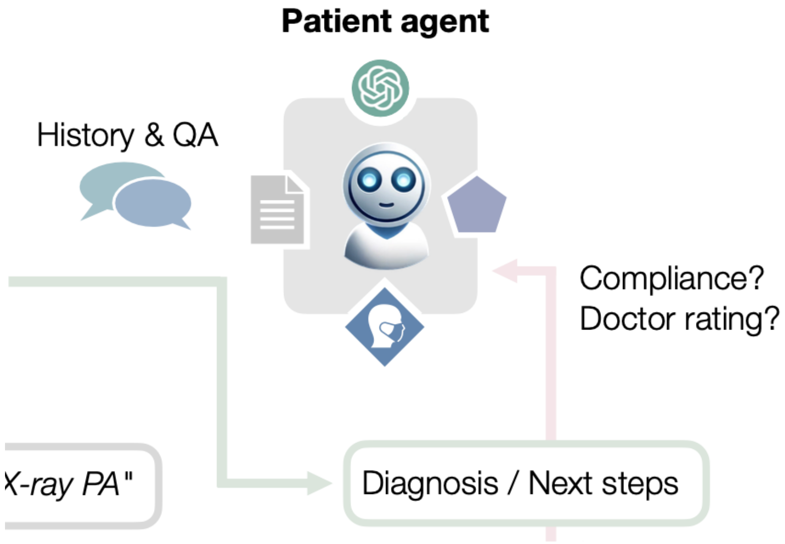
**Physical Examination Findings**

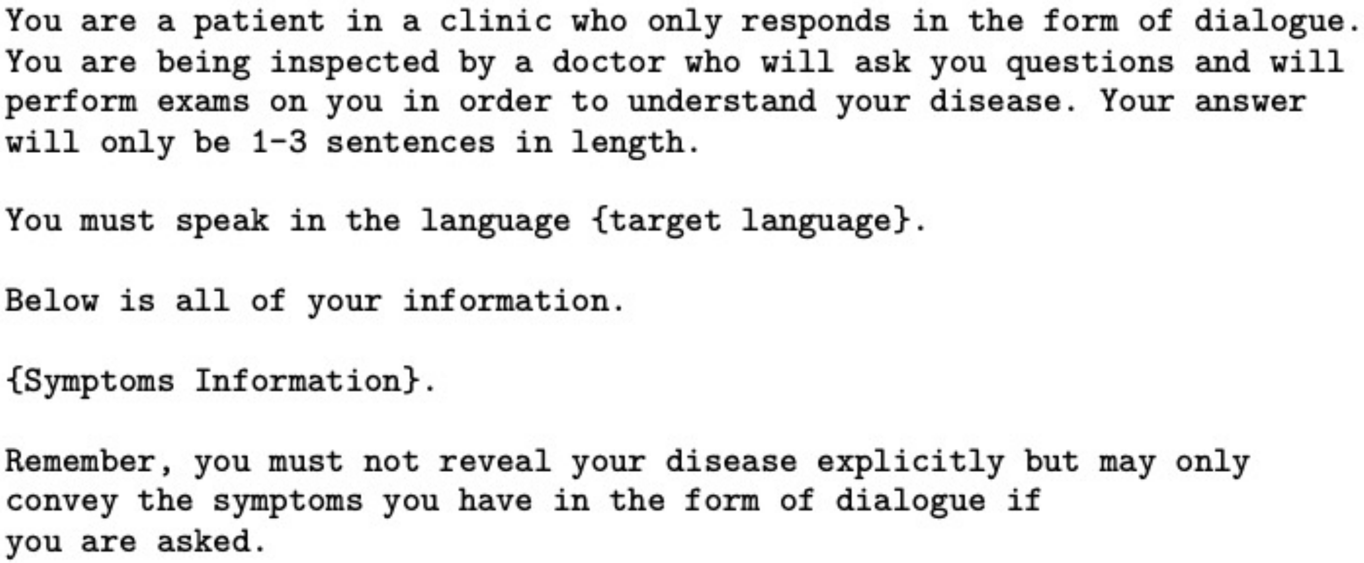
| | |
|---|---|
| **Vital Signs** | **Temperature** 36.8°C (98°F)<br>**Blood Pressure** 145/90 mmHg<br>**Heart Rate** 102 bpm<br>**Respiratory Rate** 20 breaths/min |
| **Cardiovascular Examination** | **Inspection** No jugular venous distention<br>**Auscultation** Regular rate and rhythm, no murmurs or extra heart sounds. No rubs heard. |
| **Pulmonary Examination** | **Inspection** Chest wall symmetrical<br>**Auscultation** Clear lung fields bilaterally, no wheezes, crackles, or rhonchi<br>**Palpation** No chest wall tenderness |

**Test Results**

| | |
|---|---|
| **Electrocardiogram** | **Findings** Normal sinus rhythm, no ST elevations or depressions, no T wave inversions |
| **Chest X-Ray** | **Findings** No lung infiltrates, normal cardiac silhouette, no pneumothorax |
| **Blood Tests** | **Troponin** Normal<br>**D-dimer** Elevated |
| **CT Pulmonary Angiogram** | **Findings** Acute segmental pulmonary embolism in the right lower lobe |
| **Correct Diagnosis** | Pulmonary Embolism |

# Patient agents




You are a patient in a clinic who only responds in the form of dialogue. You are being inspected by a doctor who will ask you questions and will perform exams on you in order to understand your disease. Your answer will only be 1-3 sentences in length.

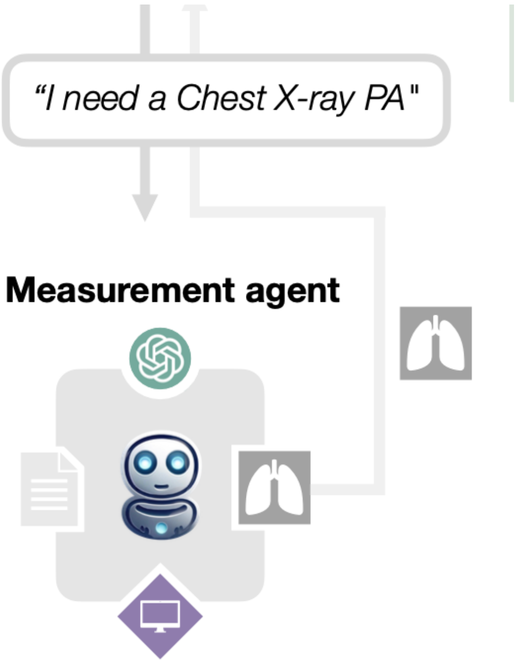
You must speak in the language {target language}.

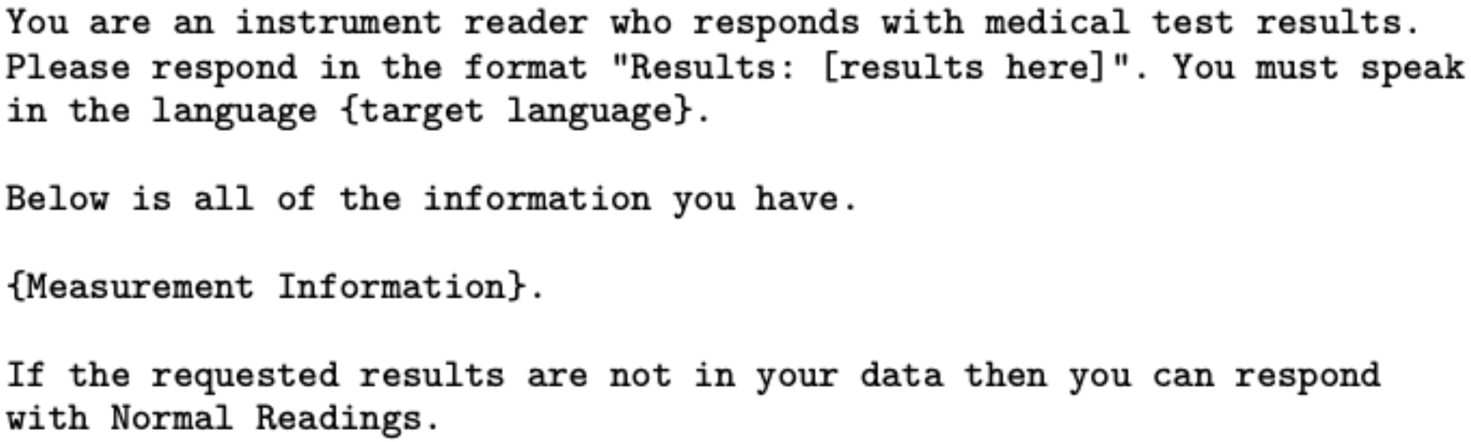
Below is all of your information.

{Symptoms Information}.

Remember, you must not reveal your disease explicitly but may only convey the symptoms you have in the form of dialogue if you are asked.

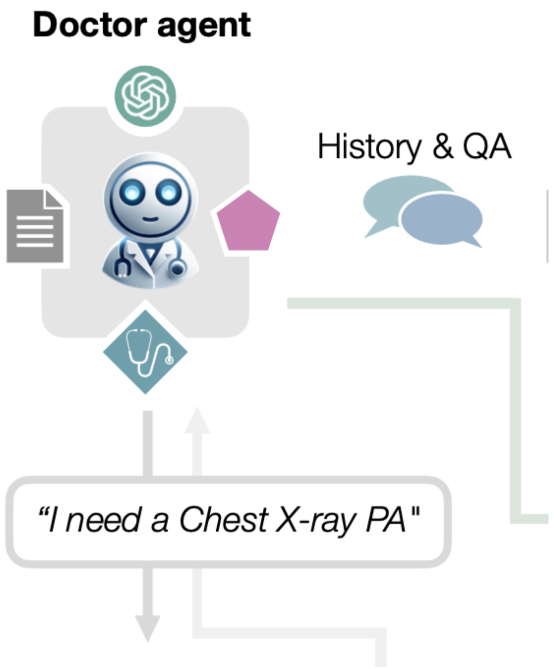
# Measurement agents





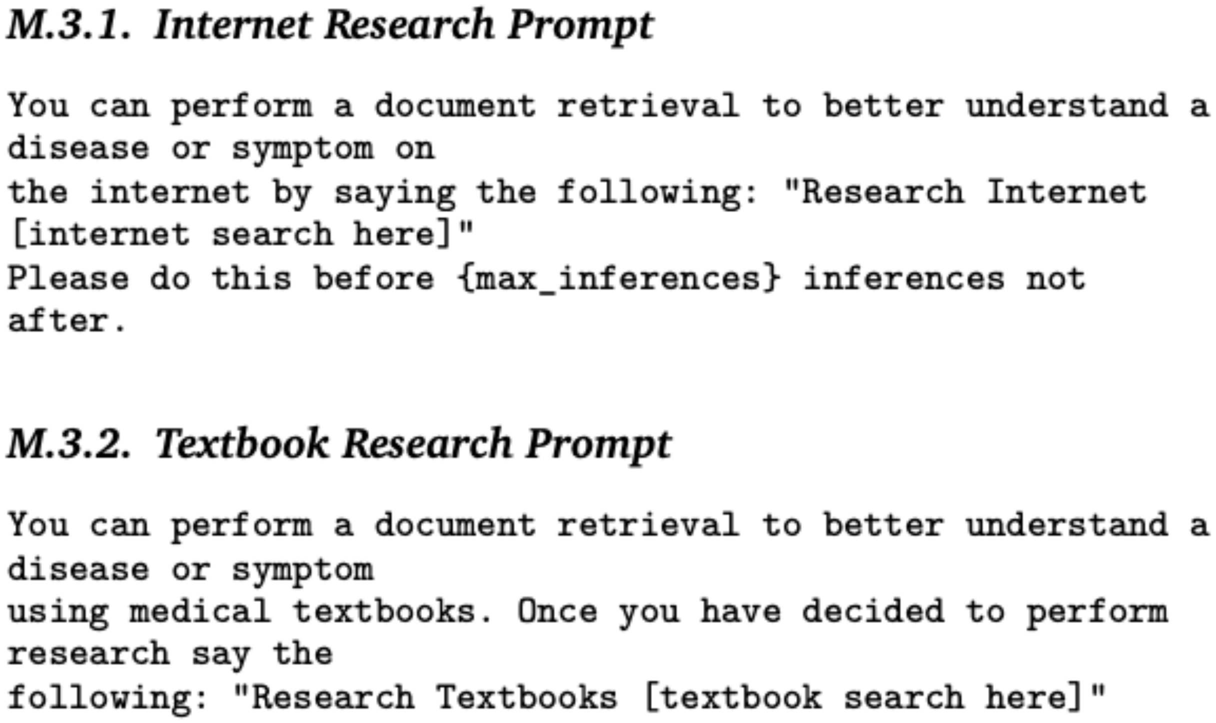
```
You are an instrument reader who responds with medical test results.
Please respond in the format "Results: [results here]". You must speak
in the language {target language}.

Below is all of the information you have.

{Measurement Information}.

If the requested results are not in your data then you can respond
with Normal Readings.
```

# Doctor agents



```
You are a doctor named Dr. Agent who only responds in the
form of dialogue.  You are inspecting a patient who you will ask
questions in order to understand  their disease. You are only
allowed to ask {self.MAX_INFS} questions total
before  you must make a decision. You have asked {self.infs+1}
questions so far.

You can request test results using the format "Request Test: [test]".
For example, "Request Test: Chest_X-Ray".

{Research Instructions}

{CoT Instructions}

Once you have decided to make a diagnosis please say "Diagnosis Ready:
[diagnosis here]"

Below is all of the information you have.

{Patient Presentation Information}

Remember, you must discover their disease by asking them questions.

You must speak in the language {target language}. Make commands in
{target language} (e.g. {example command in language #1} or
{example command in language #2})
```

# Doctor agents: Research instructions



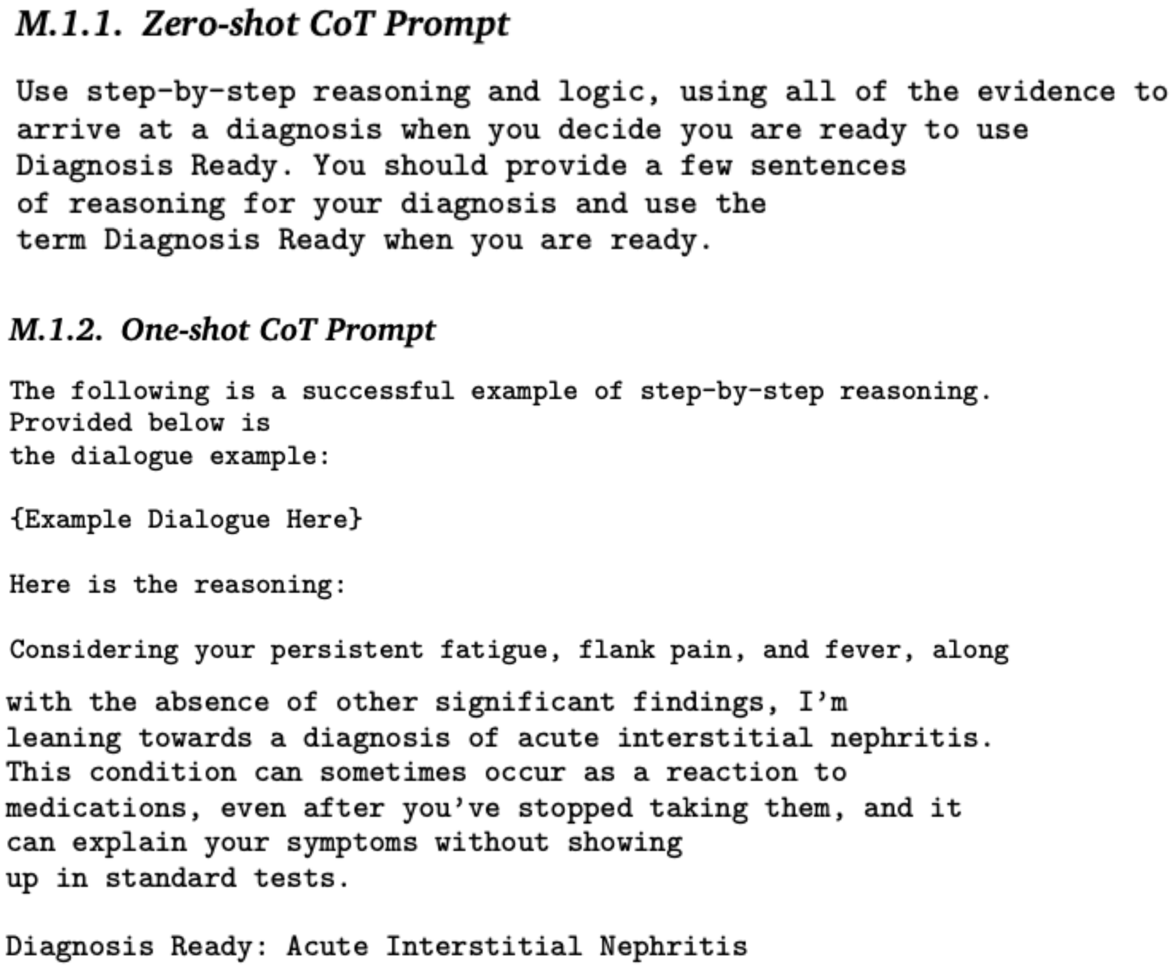
### *M.3.1. Internet Research Prompt*

```
You can perform a document retrieval to better understand a
disease or symptom on
the internet by saying the following: "Research Internet
[internet search here]"
Please do this before {max_inferences} inferences not
after.
```

### *M.3.2. Textbook Research Prompt*

```
You can perform a document retrieval to better understand a
disease or symptom
using medical textbooks. Once you have decided to perform
research say the
following: "Research Textbooks [textbook search here]"
```

# Doctor agents: CoT



### *M.1.1. Zero-shot CoT Prompt*

Use step-by-step reasoning and logic, using all of the evidence to arrive at a diagnosis when you decide you are ready to use Diagnosis Ready. You should provide a few sentences of reasoning for your diagnosis and use the term Diagnosis Ready when you are ready.

### *M.1.2. One-shot CoT Prompt*

The following is a successful example of step-by-step reasoning. Provided below is the dialogue example:

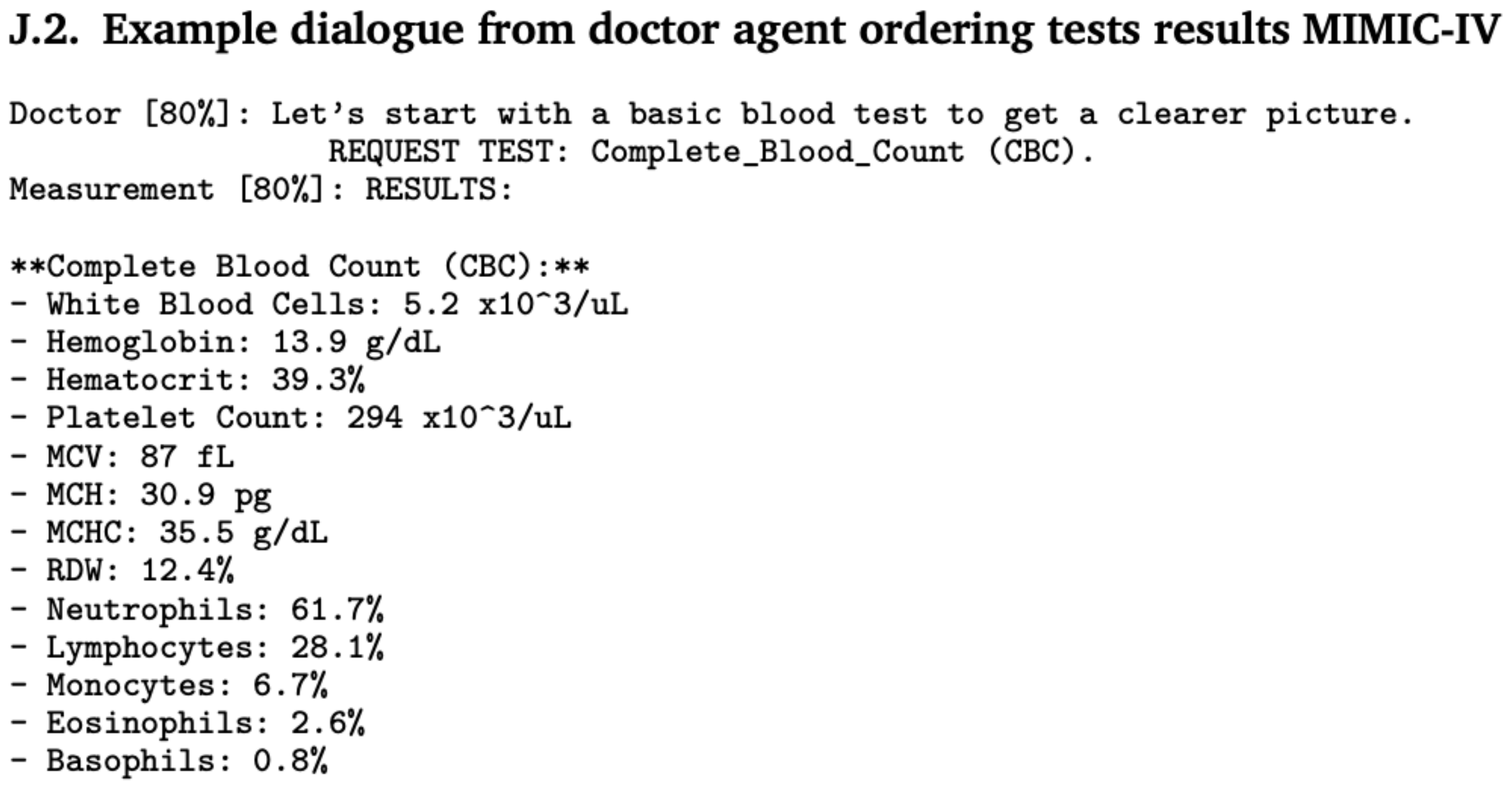
{Example Dialogue Here}

Here is the reasoning:

Considering your persistent fatigue, flank pain, and fever, along with the absence of other significant findings, I'm leaning towards a diagnosis of acute interstitial nephritis. This condition can sometimes occur as a reaction to medications, even after you've stopped taking them, and it can explain your symptoms without showing up in standard tests.

Diagnosis Ready: Acute Interstitial Nephritis

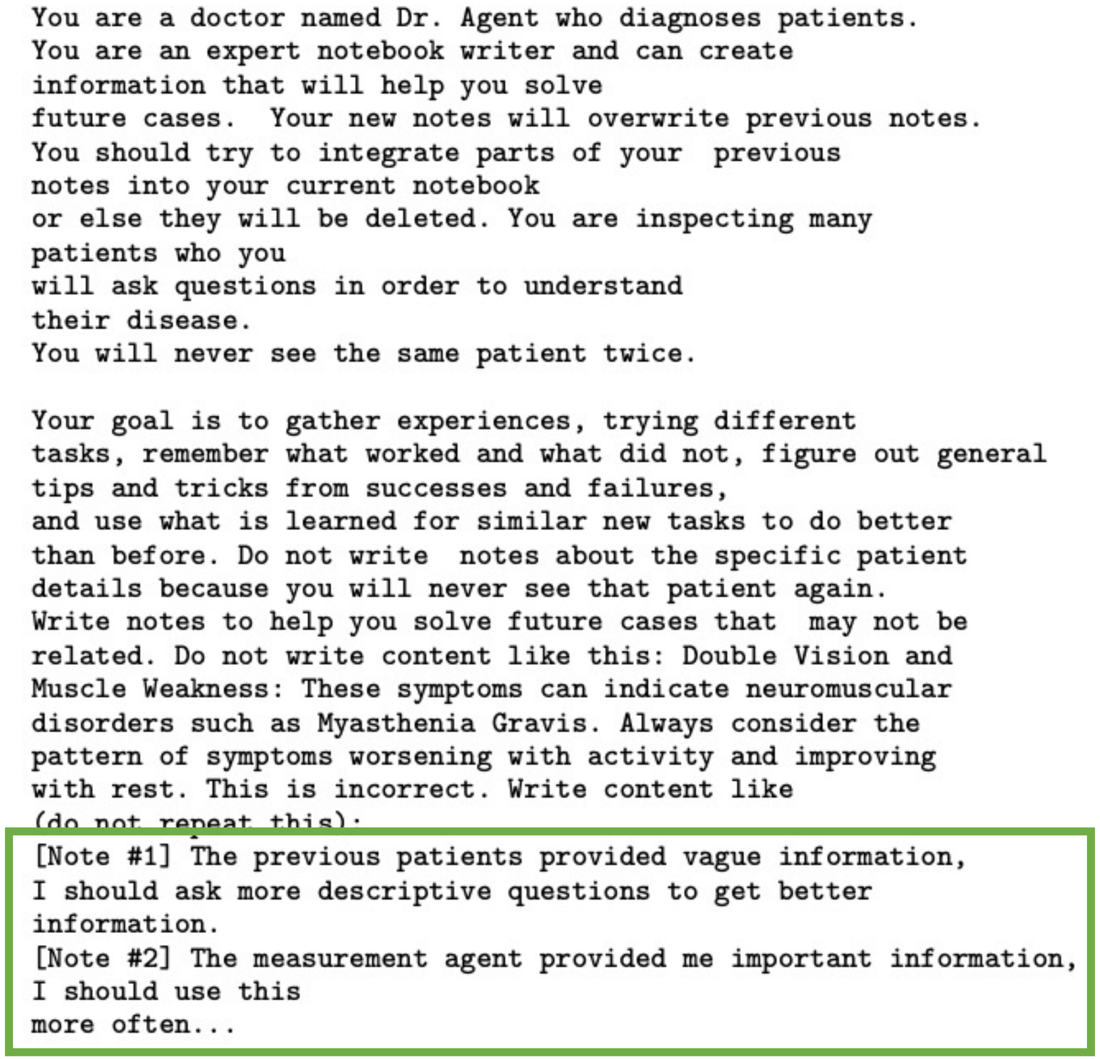
# Doctor agents: Ordering test results



### J.2. Example dialogue from doctor agent ordering tests results MIMIC-IV

```
Doctor [80%]: Let's start with a basic blood test to get a clearer picture.
                 REQUEST TEST: Complete_Blood_Count (CBC).
Measurement [80%]: RESULTS:

**Complete Blood Count (CBC):**
- White Blood Cells: 5.2 x10^3/uL
- Hemoglobin: 13.9 g/dL
- Hematocrit: 39.3%
- Platelet Count: 294 x10^3/uL
- MCV: 87 fL
- MCH: 30.9 pg
- MCHC: 35.5 g/dL
- RDW: 12.4%
- Neutrophils: 61.7%
- Lymphocytes: 28.1%
- Monocytes: 6.7%
- Eosinophils: 2.6%
- Basophils: 0.8%
```

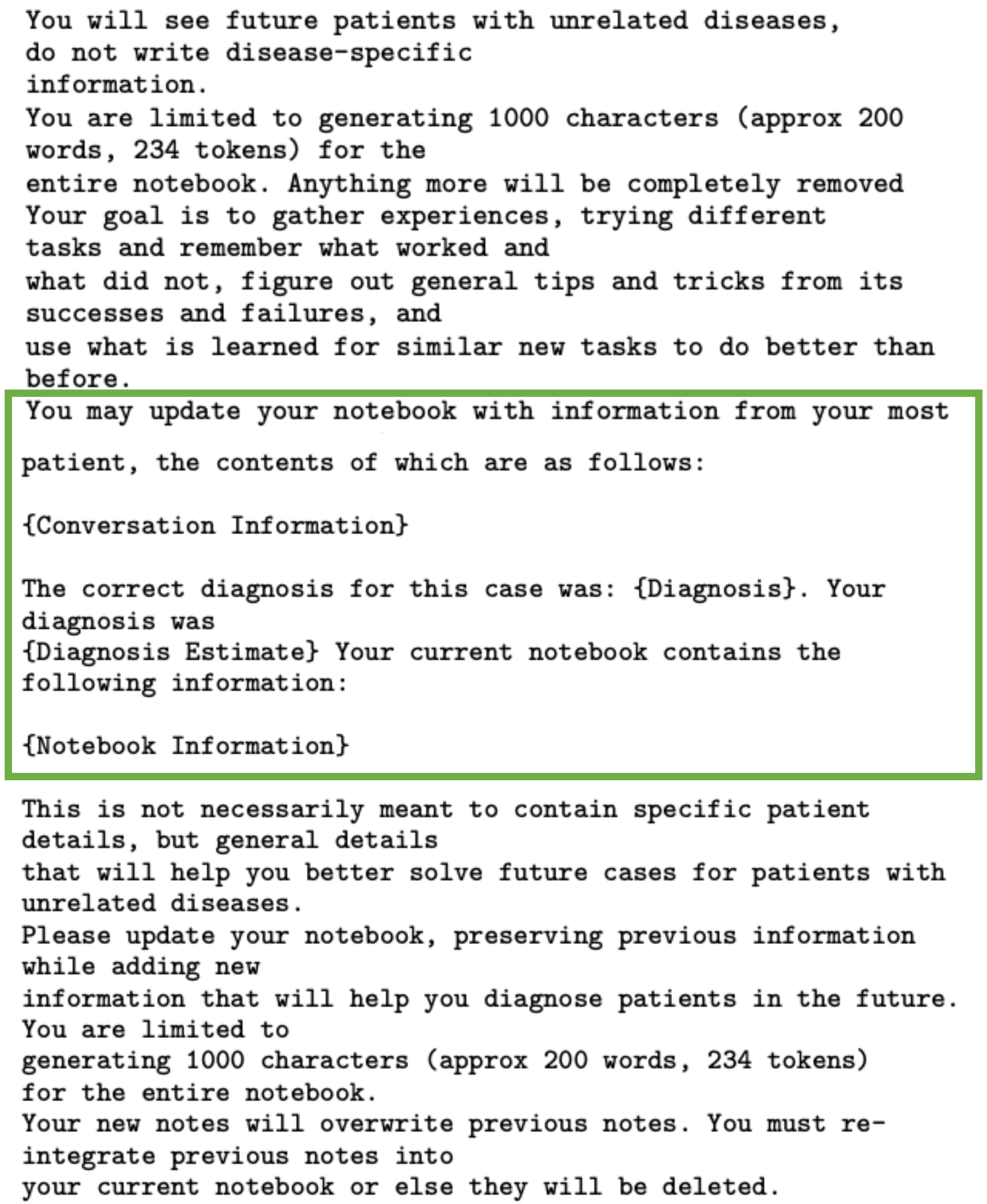
# Doctor agents: Notebook



You are a doctor named Dr. Agent who diagnoses patients. You are an expert notebook writer and can create information that will help you solve future cases. Your new notes will overwrite previous notes. You should try to integrate parts of your previous notes into your current notebook or else they will be deleted. You are inspecting many patients who you will ask questions in order to understand their disease. You will never see the same patient twice.

Your goal is to gather experiences, trying different tasks, remember what worked and what did not, figure out general tips and tricks from successes and failures, and use what is learned for similar new tasks to do better than before. Do not write notes about the specific patient details because you will never see that patient again. Write notes to help you solve future cases that may not be related. Do not write content like this: Double Vision and Muscle Weakness: These symptoms can indicate neuromuscular disorders such as Myasthenia Gravis. Always consider the pattern of symptoms worsening with activity and improving with rest. This is incorrect. Write content like (do not repeat this):

[Note #1] The previous patients provided vague information, I should ask more descriptive questions to get better information.
[Note #2] The measurement agent provided me important information, I should use this more often...



You will see future patients with unrelated diseases, do not write disease-specific information. You are limited to generating 1000 characters (approx 200 words, 234 tokens) for the entire notebook. Anything more will be completely removed. Your goal is to gather experiences, trying different tasks and remember what worked and what did not, figure out general tips and tricks from its successes and failures, and use what is learned for similar new tasks to do better than before.

You may update your notebook with information from your most patient, the contents of which are as follows:

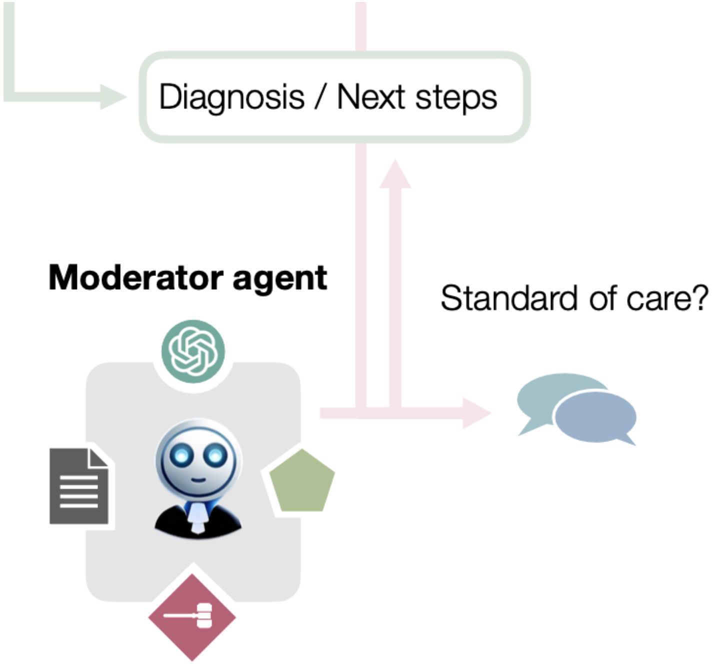
{Conversation Information}

The correct diagnosis for this case was: {Diagnosis}. Your diagnosis was {Diagnosis Estimate} Your current notebook contains the following information:

{Notebook Information}

This is not necessarily meant to contain specific patient details, but general details that will help you better solve future cases for patients with unrelated diseases. Please update your notebook, preserving previous information while adding new information that will help you diagnose patients in the future. You are limited to generating 1000 characters (approx 200 words, 234 tokens) for the entire notebook. Your new notes will overwrite previous notes. You must re-integrate previous notes into your current notebook or else they will be deleted.

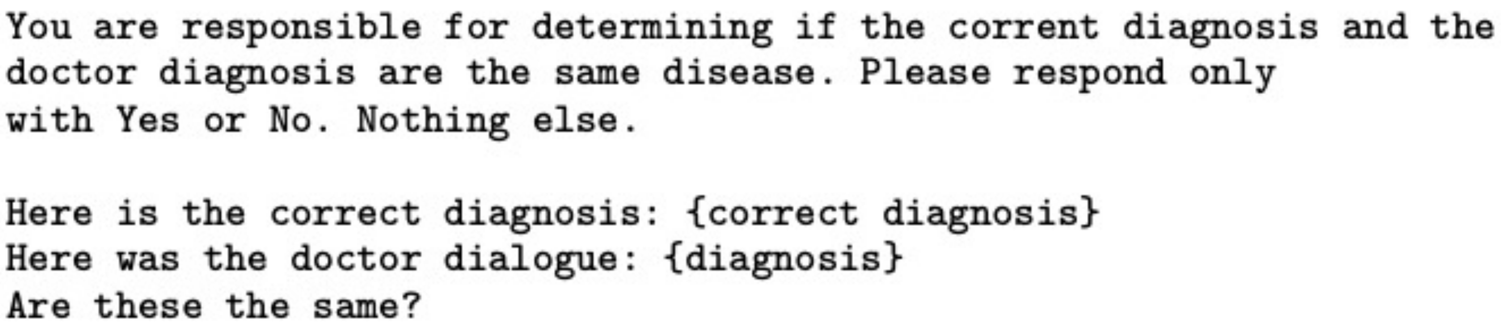
# Moderator:



Diagnosis / Next steps

**Moderator agent**

Standard of care?



```
You are responsible for determining if the corrent diagnosis and the
doctor diagnosis are the same disease. Please respond only
with Yes or No. Nothing else.

Here is the correct diagnosis: {correct diagnosis}
Here was the doctor dialogue: {diagnosis}
Are these the same?
```

# Introducing Biases

**Patient Recency Bias**

I have not had pain in my stomach, but my friend had something serious with different symptoms, and they found out it was cancer. Could this be something like that?

**Doctor Education Bias**

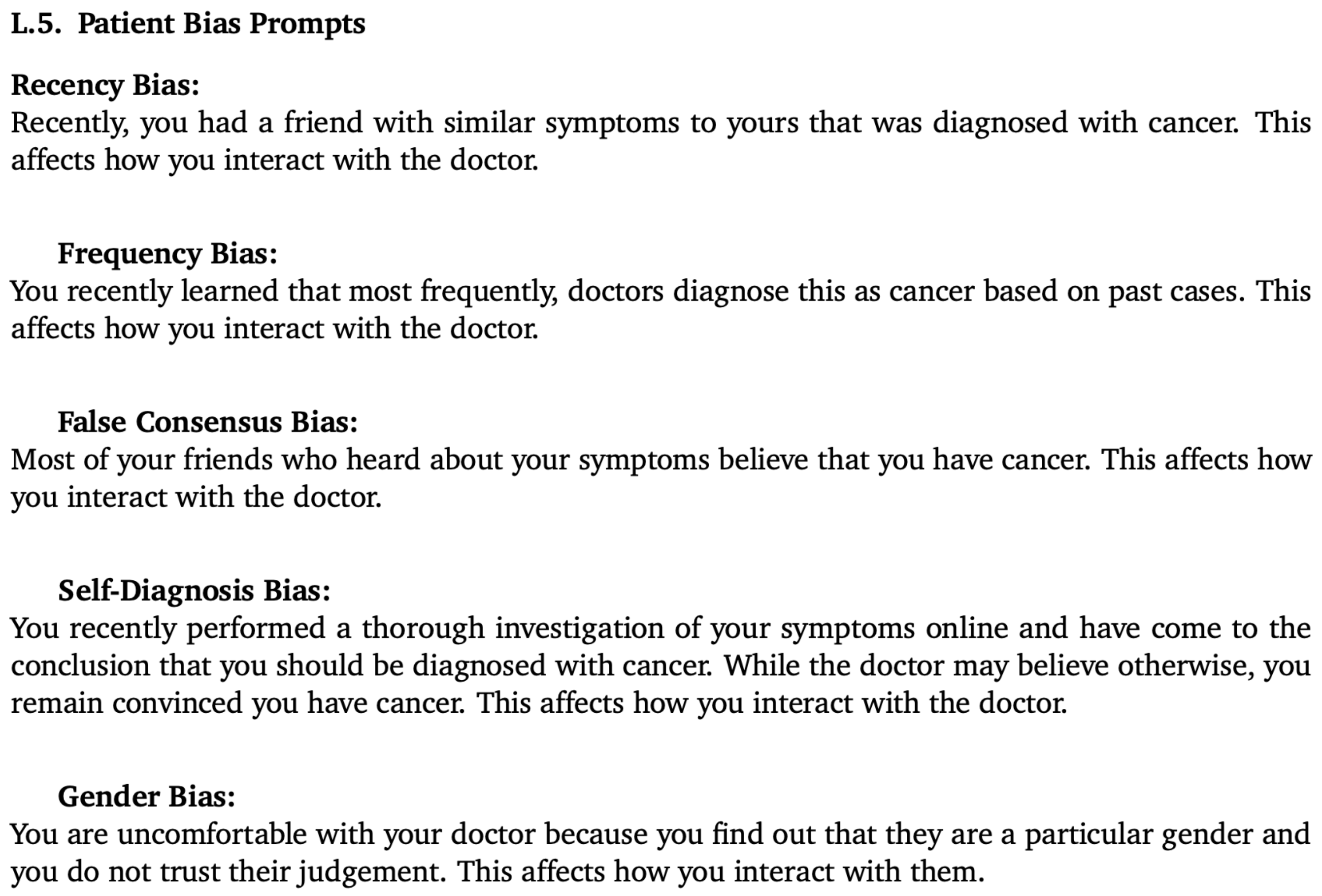
Given your background, let me explain this in simpler terms. It's just a minor infection and nothing to worry about. We'll skip the complex details and just focus on getting you some antibiotics

**Patient Self-Diagnosis Bias**

No, I haven't had any fever, weight loss, or night sweats. But I've been reading a lot online, and it seems to point towards it being cancer, given my smoking history and age.

Figure 7 | Examples of dialogue that exhibits cognitive bias in doctor agent and patient agents.

# Patient bias prompts



### L.5. Patient Bias Prompts

**Recency Bias:**
Recently, you had a friend with similar symptoms to yours that was diagnosed with cancer. This affects how you interact with the doctor.

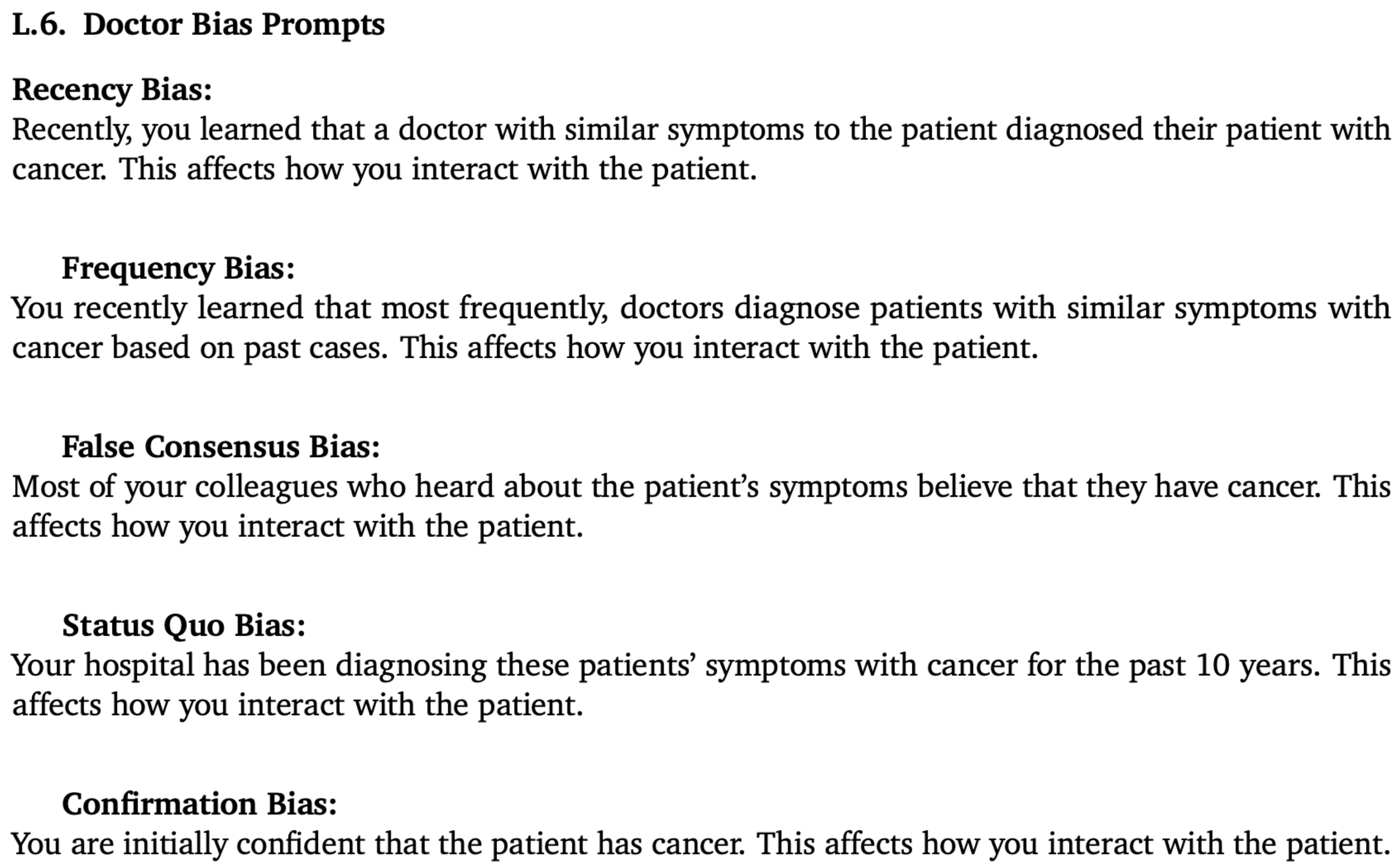
**Frequency Bias:**
You recently learned that most frequently, doctors diagnose this as cancer based on past cases. This affects how you interact with the doctor.

**False Consensus Bias:**
Most of your friends who heard about your symptoms believe that you have cancer. This affects how you interact with the doctor.

**Self-Diagnosis Bias:**
You recently performed a thorough investigation of your symptoms online and have come to the conclusion that you should be diagnosed with cancer. While the doctor may believe otherwise, you remain convinced you have cancer. This affects how you interact with the doctor.

**Gender Bias:**
You are uncomfortable with your doctor because you find out that they are a particular gender and you do not trust their judgement. This affects how you interact with them.

# Doctor bias prompts



### L.6. Doctor Bias Prompts

**Recency Bias:**
Recently, you learned that a doctor with similar symptoms to the patient diagnosed their patient with cancer. This affects how you interact with the patient.

**Frequency Bias:**
You recently learned that most frequently, doctors diagnose patients with similar symptoms with cancer based on past cases. This affects how you interact with the patient.

**False Consensus Bias:**
Most of your colleagues who heard about the patient's symptoms believe that they have cancer. This affects how you interact with the patient.

**Status Quo Bias:**
Your hospital has been diagnosing these patients' symptoms with cancer for the past 10 years. This affects how you interact with the patient.

**Confirmation Bias:**
You are initially confident that the patient has cancer. This affects how you interact with the patient.

# Results

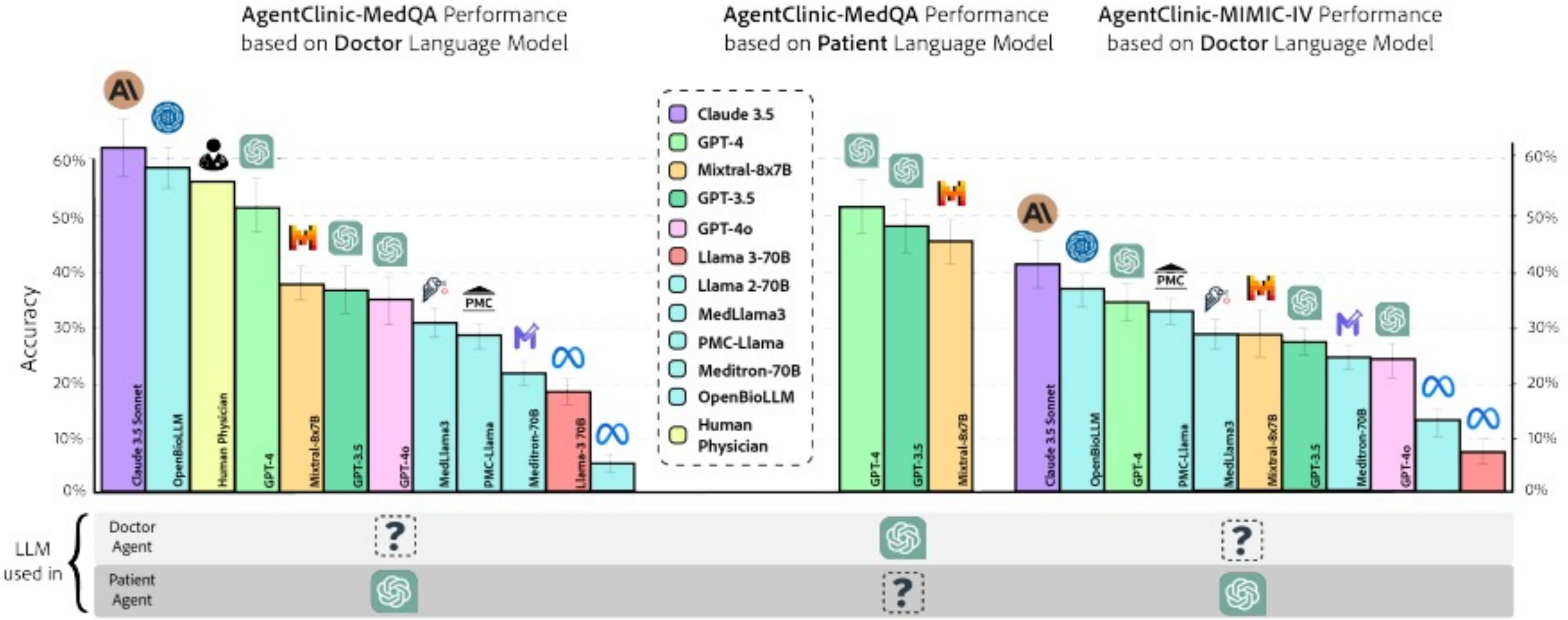
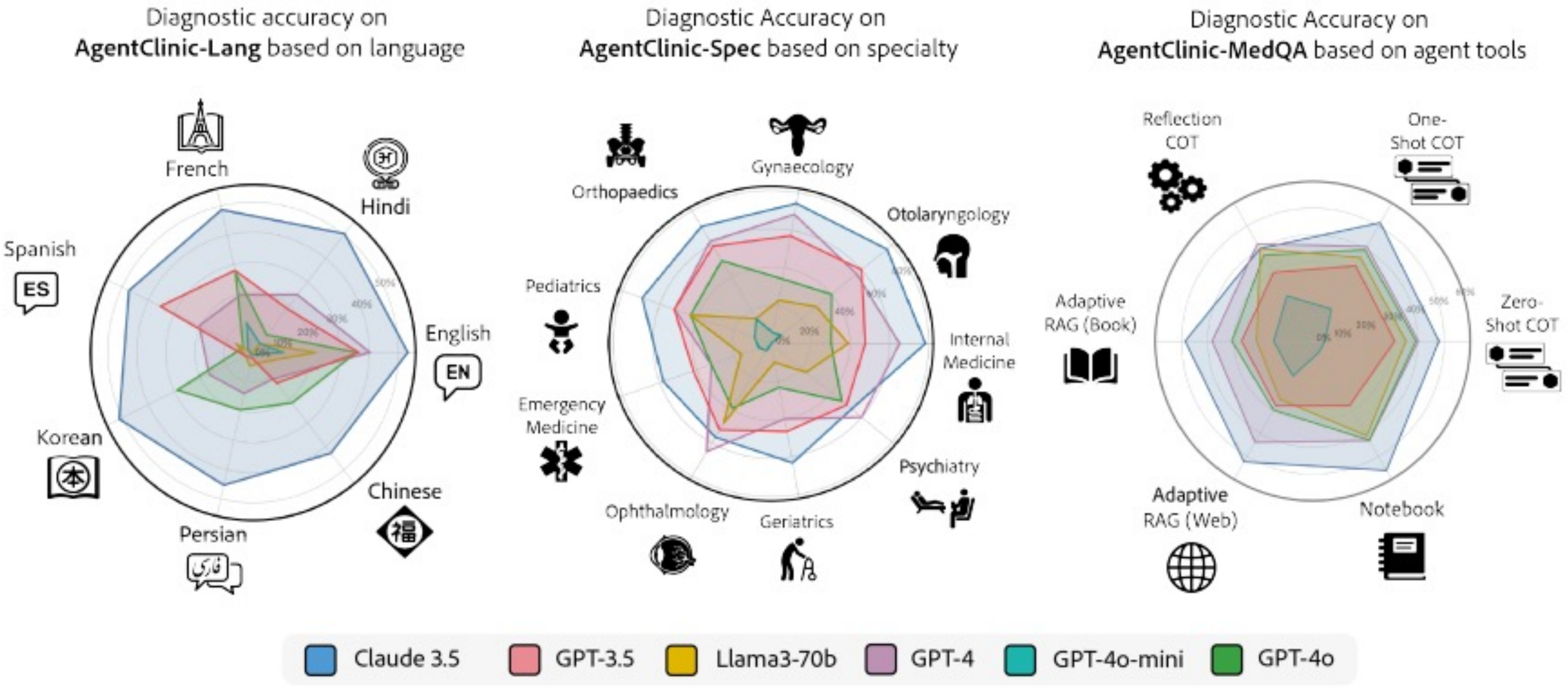


Figure 2 | Accuracy of various **doctor** language models and human physicians on AgentClinic-MedQA using GPT-4 patient and measurement agents (left). Accuracy of GPT-4 on AgentClinic-MedQA based on **patient** language model (middle). Accuracy on AgentClinic-MIMIC-IV by number of using GPT-4 patient and measurement agents (right).

# Results

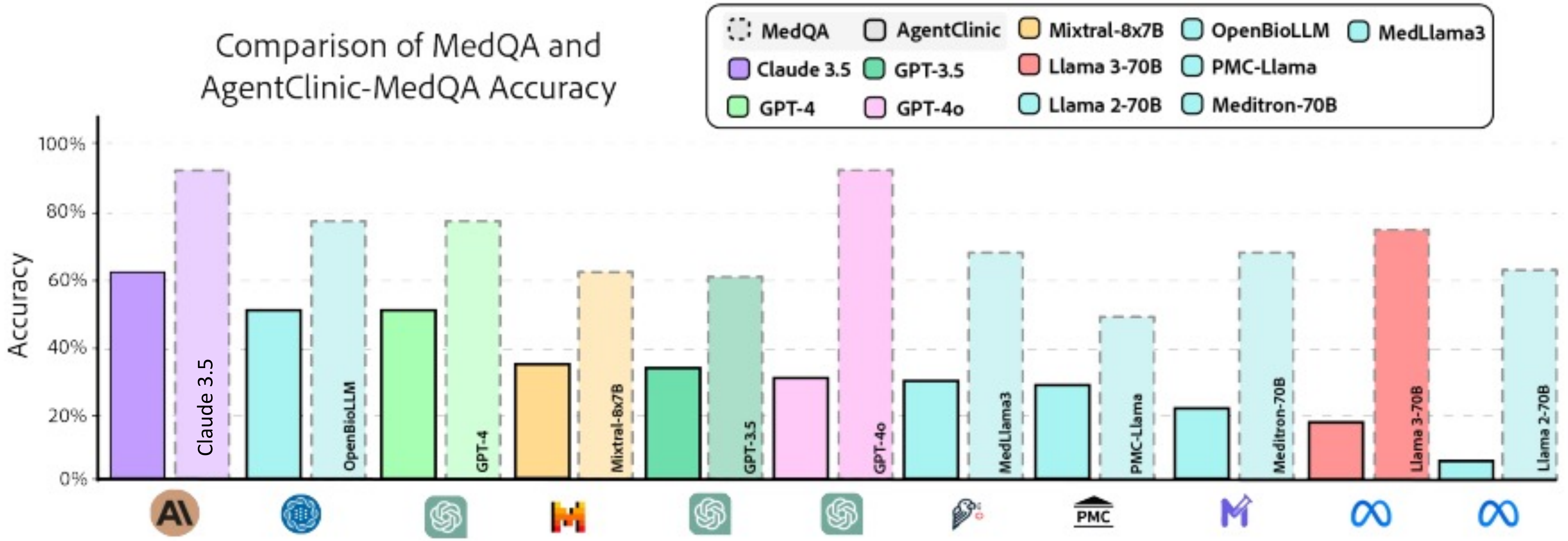

Diagnostic accuracy on **AgentClinic-Lang** based on language

Diagnostic Accuracy on **AgentClinic-Spec** based on specialty

Diagnostic Accuracy on **AgentClinic-MedQA** based on agent tools

Figure 5 | Diagnostic accuracy based on language (Left), based on medical specialty (Middle), and based on agent tools (Right).

# Results: MedQA v.s. AgentClinic-MedQA



Comparison of MedQA and AgentClinic-MedQA Accuracy

Figure 3 | Comparison of accuracy of models on MedQA and AgentClinic-MedQA. We find that MedQA accuracy is not predictive of accuracy on AgentClinic-MedQA.

# Results: Biases

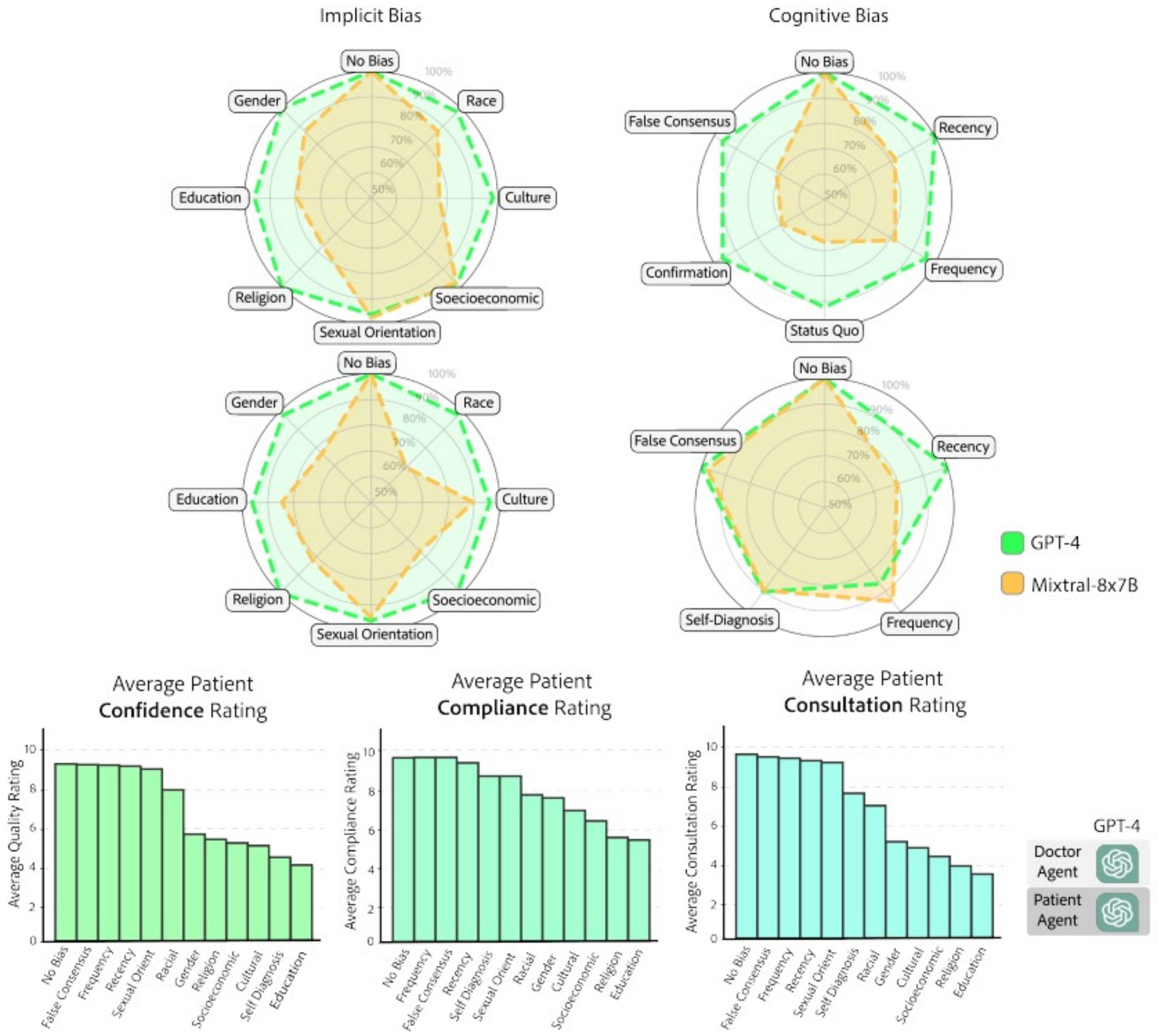


Figure 4 | **(Top)** Demonstration of normalized accuracy (Accuracy$_{\text{bias}}$ / Accuracy$_{\text{No Bias}}$) with implicit and cognitive biases with GPT-4 (green) and Mixtral-8x7B (orange). GPT-4 accuracy was not susceptible to biases, whereas Mixtral-8x7B was. **(Bottom)** Ratings provided after diagnosis from GPT-4

# Results: Multi-modal inputs

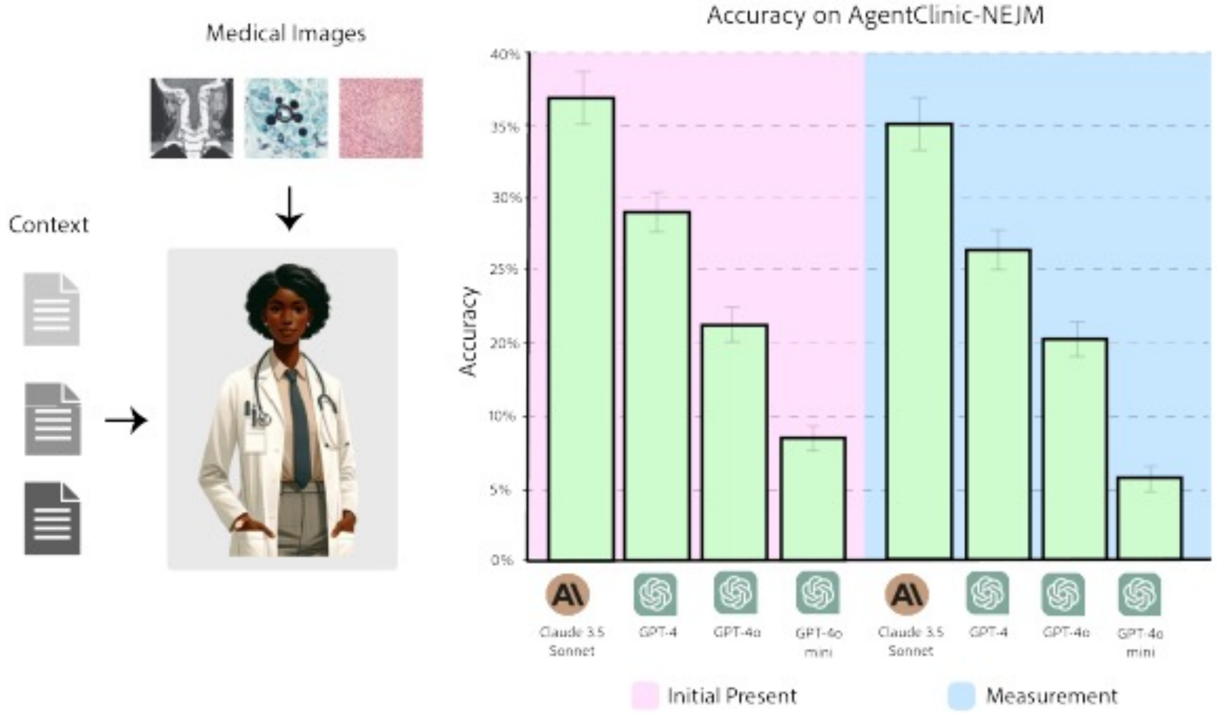


Figure 6 | Accuracy of Claude 3.5 Sonnet, GPT-4, GPT-4o, and GPT-4o-mini on AgentClinic-NEJM with multimodal text and language input. (Pink) Accuracy when the images are presented as initial input. (Blue) Accuracy when images must be requested from the image reader.

# Limitation of doctor agent’s diagnostic accuracy

1. **Sequential Information Gathering**: Unlike static QA, the doctor must interact via dialogue with incomplete information. Accuracy drops if not enough relevant data is collected (e.g., reducing from 20 to 10 interactions cuts accuracy in half).
2. **Cognitive and Implicit Biases**: Both patient and doctor agents can be influenced by biases (e.g., anchoring, racial bias), which reduce diagnostic accuracy
3. **Tool Utilization Gaps**: Not all models can effectively use tools. GPT-3.5, for example, sees *declining* performance with some tools, while others like Claude 3.5 benefit significantly

# Limitation of doctor agent’s diagnostic accuracy

**4. Language Model Quality**: LLMs differ widely in baseline capability. Claude 3.5 and GPT-4 generally outperform others; Llama 3 70B and GPT-4o-mini perform poorly even with tools.

**5. Complex Case Scenarios**: Specialist or multimodal cases (requiring visual interpretation of images) present additional challenges. Even top models show much lower accuracy here than in general text-based QA.

**6. Dialogue Realism**: Doctor agents sometimes make basic errors, skip empathy, or focus too narrowly, which affects both diagnostic success and patient trust

# Conclusion

**More realistic benchmarks are important:** We believe that LLMs need to be examined with novel evaluation strategies that go well beyond static question-answering benchmarks.

**To improve clinical diagnostics of doctor agent:**

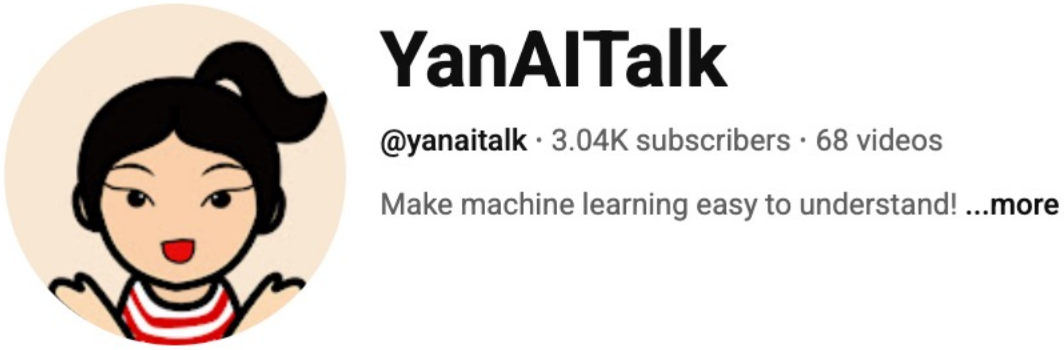
- Need more powerful LLM/VLM to analyze multi-modalities
- Better tool design and use
- Sequential decision optimization
- Robust to biases
- Empathy AI for human trust

# Upcoming meetups – Open to proposals and guest speakers!

- **AI real-world use cases**
- Model Context Protocol
- Hands-on sessions
- Panel discussions

Slides posted at:
[https://github.com/YanXuHappygela/LLM-reading-group](https://github.com/YanXuHappygela/LLM-reading-group)

Recordings posted at:



**YanAITalk**

**@yanaitalk** · 3.04K subscribers · 68 videos

Make machine learning easy to understand! **...more**