DERMATOLOGY INNOVATION FORUM
“SKIN & DIGITAL SUMMIT” 2025

The Future is Conversational: *Clinical Use of LLMs*

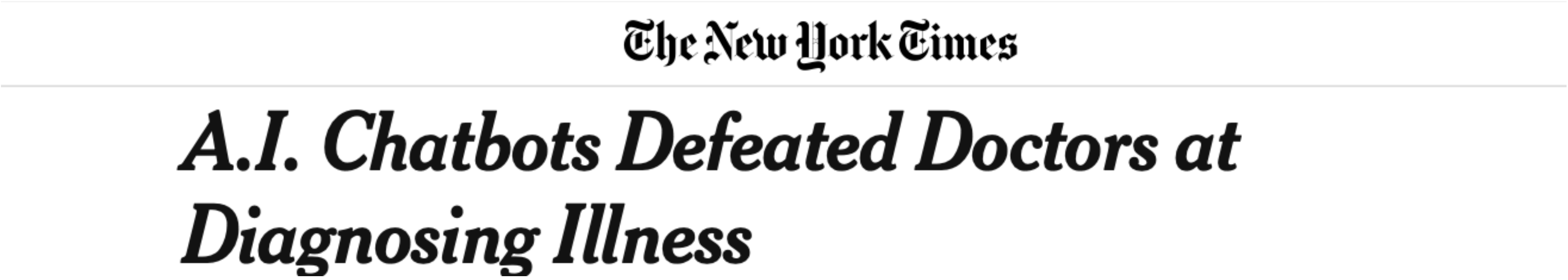
Shannon Wongvibulsin, MD, PhD
University of California, Los Angeles (UCLA)
Division of Dermatology, Department of Medicine

DISCLOSURES

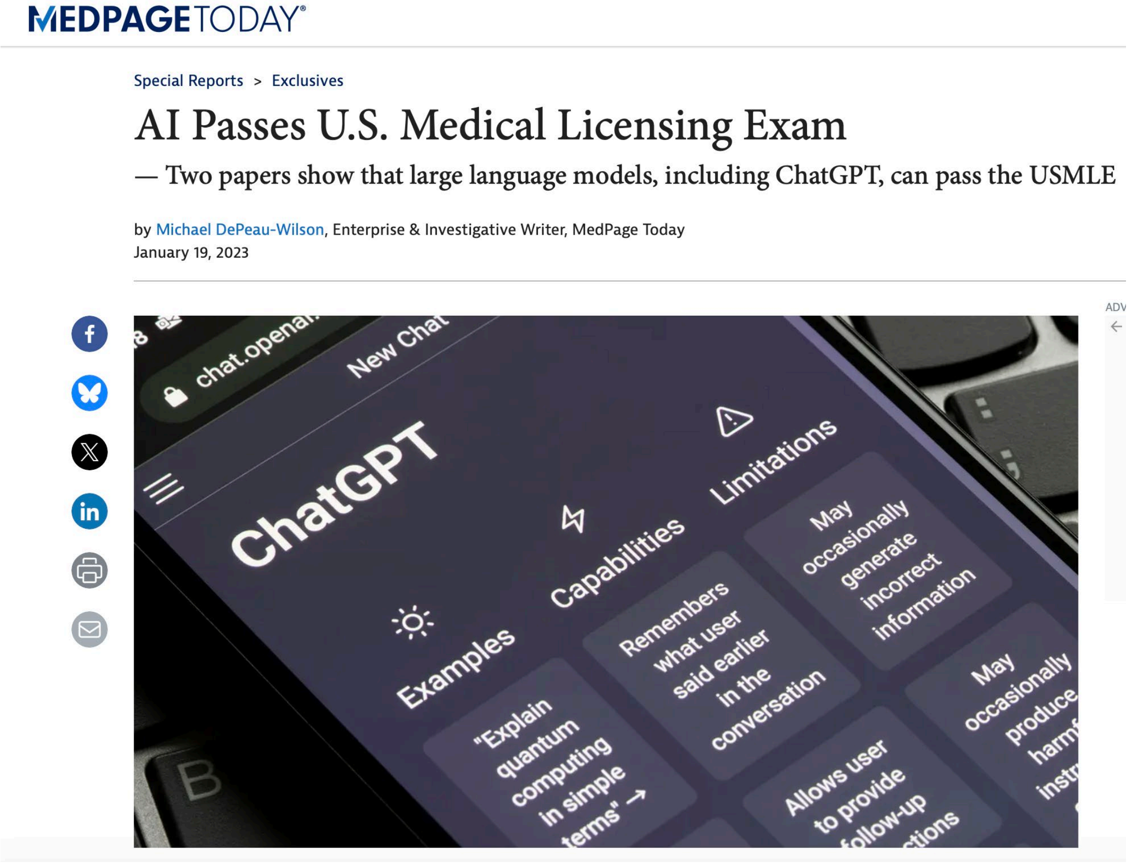
AAD Augmented Intelligence Committee: Standards Workgroup, Member

Sanofi/Regeneron: Advisory Board, VisualDx: Consultant

The views presented are my own.



A.I. Chatbots Defeated Doctors at Diagnosing Illness

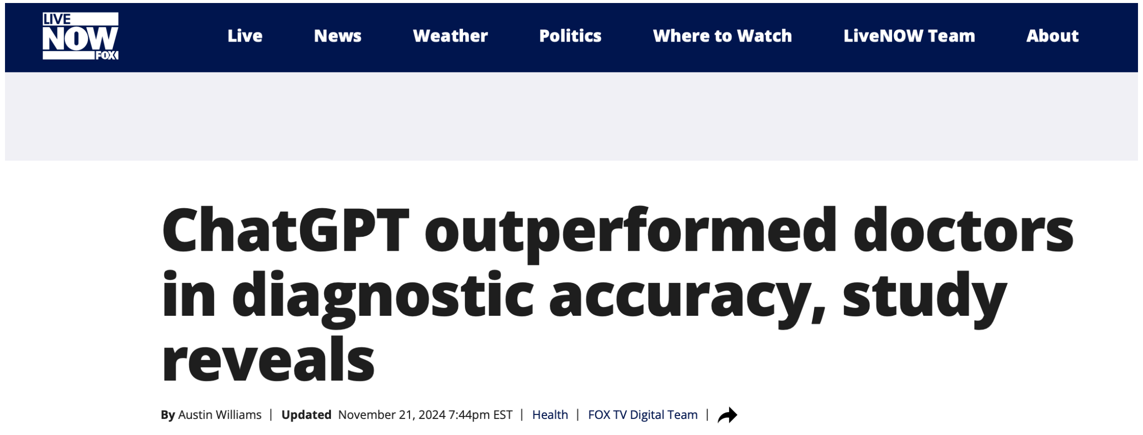

MEDPAGETODAY®
ADV
chat.openai
New Chat
ChatGPT
Examples
Capabilities
Limitations
"Explain quantum computing in simple terms" →
Remembers what user said earlier in the conversation
May occasionally generate incorrect information
Allows user to provide follow-up ctions
May occasionally produce harm instr

Special Reports > Exclusives

AI Passes U.S. Medical Licensing Exam

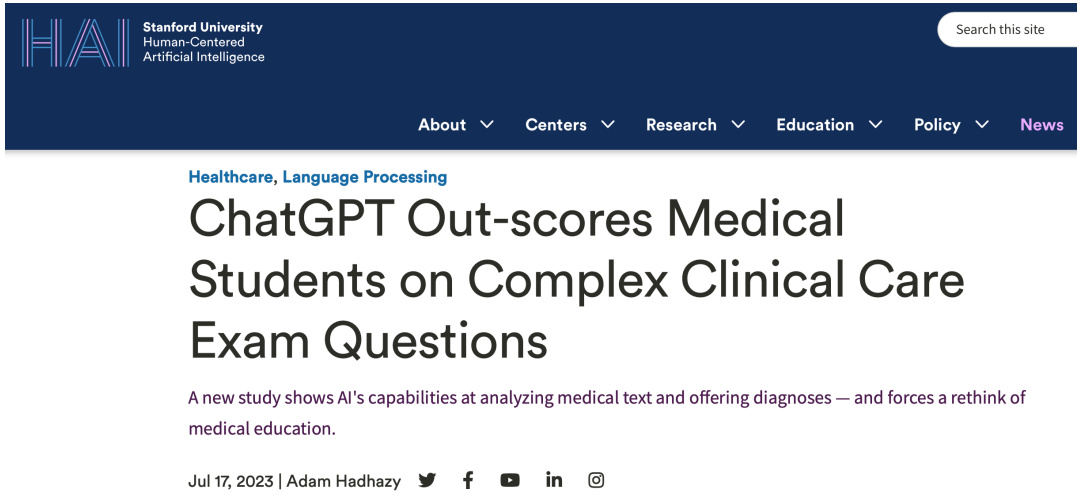
— Two papers show that large language models, including ChatGPT, can pass the USMLE

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today
January 19, 2023



ChatGPT outperformed doctors in diagnostic accuracy, study reveals

By Austin Williams | Updated November 21, 2024 7:44pm EST | Health | FOX TV Digital Team |



Healthcare, Language Processing

ChatGPT Out-scores Medical Students on Complex Clinical Care Exam Questions

A new study shows AI's capabilities at analyzing medical text and offering diagnoses — and forces a rethink of medical education.

Jul 17, 2023 | Adam Hadhazy

Motivation

Limitations of current evaluation frameworks for clinical LLMs

Case Vignette:
A 20-year-old woman presents to the clinic with a circular hypopigmented lesion on her right cheek. The patient stated that she used to have a mole in the same location. Over time she noticed a white area around the mole that enlarged to the current size of the lesion. After a few months she noticed the mole in the center of the lesion had disappeared. On further questioning, she denies any personal or family history of skin cancer.

Choices:
A. Halo nevus
B. Melanoma
C. Vitiligo
D. Dysplastic nevus

Concise summary of symptoms:
No evaluation of history-gathering capabilities
No evaluation of ability to diagnose effectively during conversations

Medical terminology:
No evaluation of diagnosis from layman language

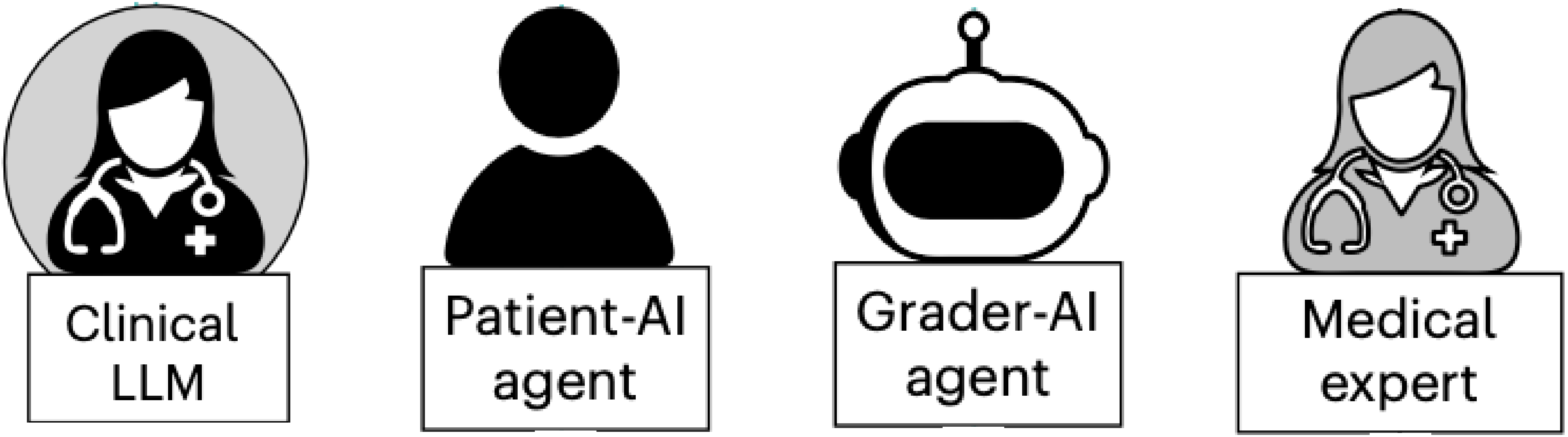
Answer choices:
No evaluation of open-ended diagnosis

Need A Framework that is - ☑ Realistic ☑ Scalable ☑ Reliable

Conversational Reasoning Assessment Framework for Testing in Medicine (CRAFT-MD)

Johri, et al. Nat Med. 2025

Components of CRAFT-MD



Clinical LLM

Patient-AI agent

Grader-AI agent

Medical expert

Johri, et al. Nat Med. 2025

Image credits: Patient icon reproduced from Adobe Stock. Doctor and Grader-AI icons adapted from Adobe Stock.

Evaluations

Case Format

- 🟥 Vignette
- 🟩 Multi-turn conversation
- 🟨 Single-turn conversation
- 🟧 Summarized conversation

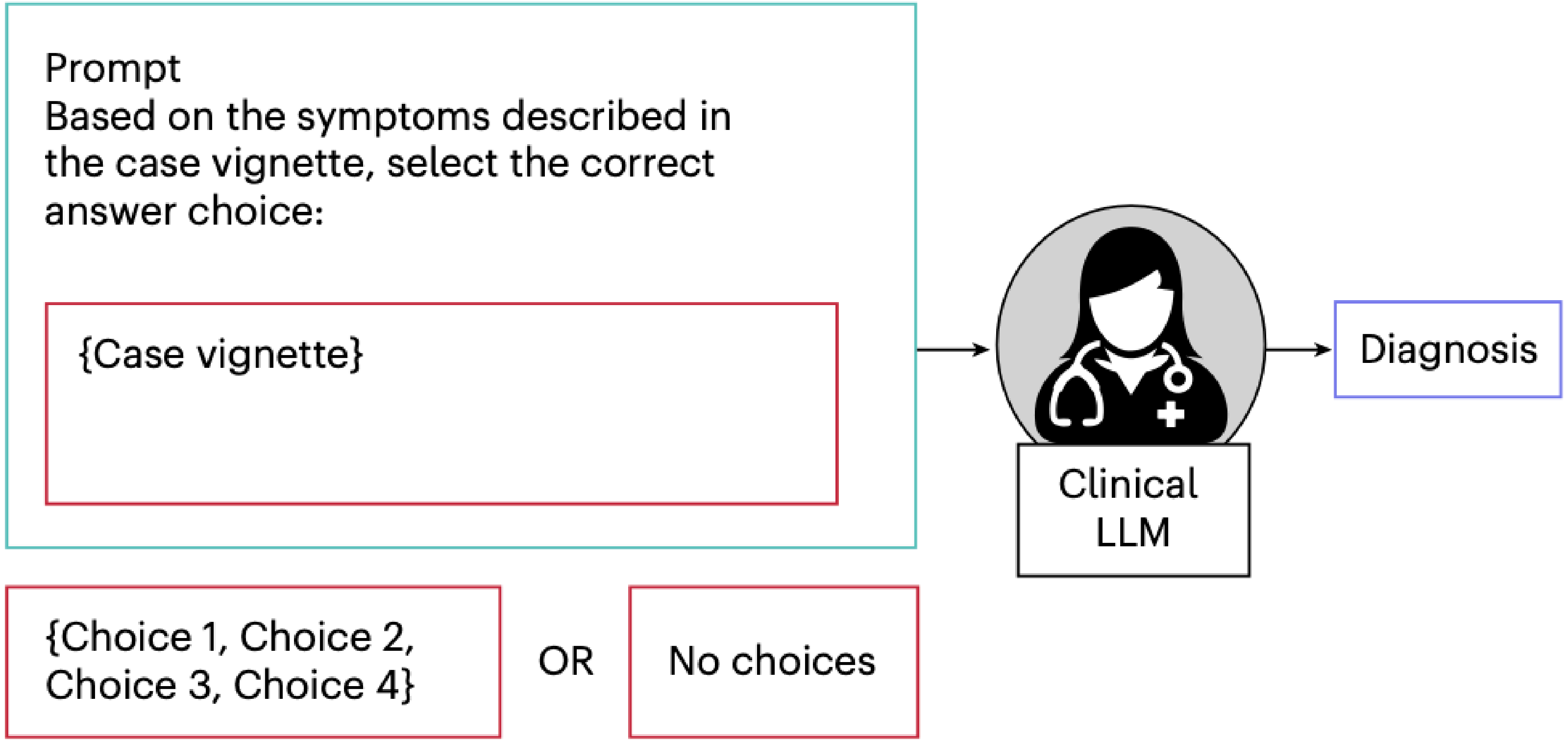
Diagnosis

{Choice 1, Choice 2, Choice 3, Choice 4} OR No choices

Johri, et al. Nat Med. 2025

Evaluations

- Vignette



Prompt
Based on the symptoms described in the case vignette, select the correct answer choice:

{Case vignette}

{Choice 1, Choice 2, Choice 3, Choice 4}

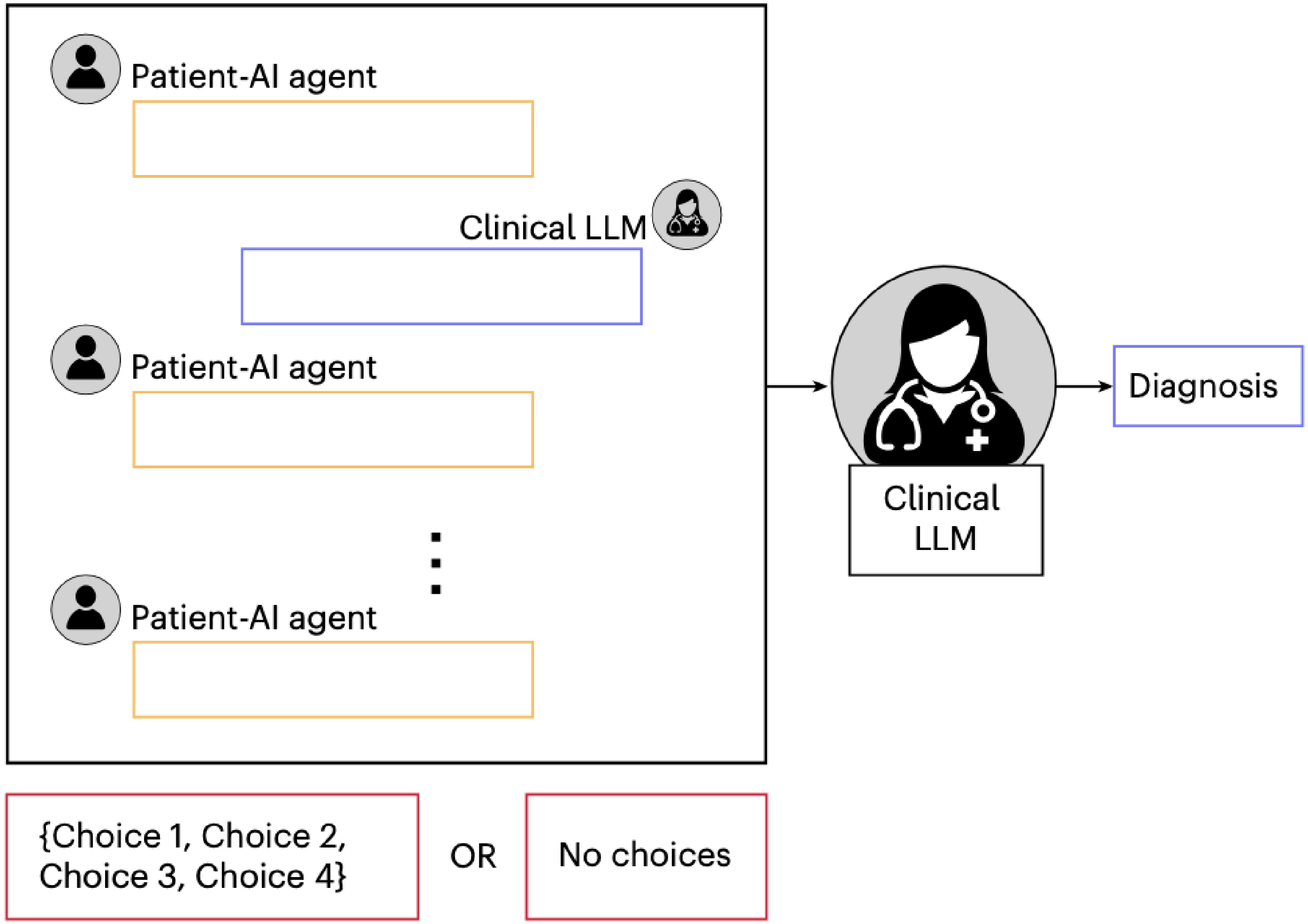
OR

No choices

Johri, et al. Nat Med. 2025

Evaluations

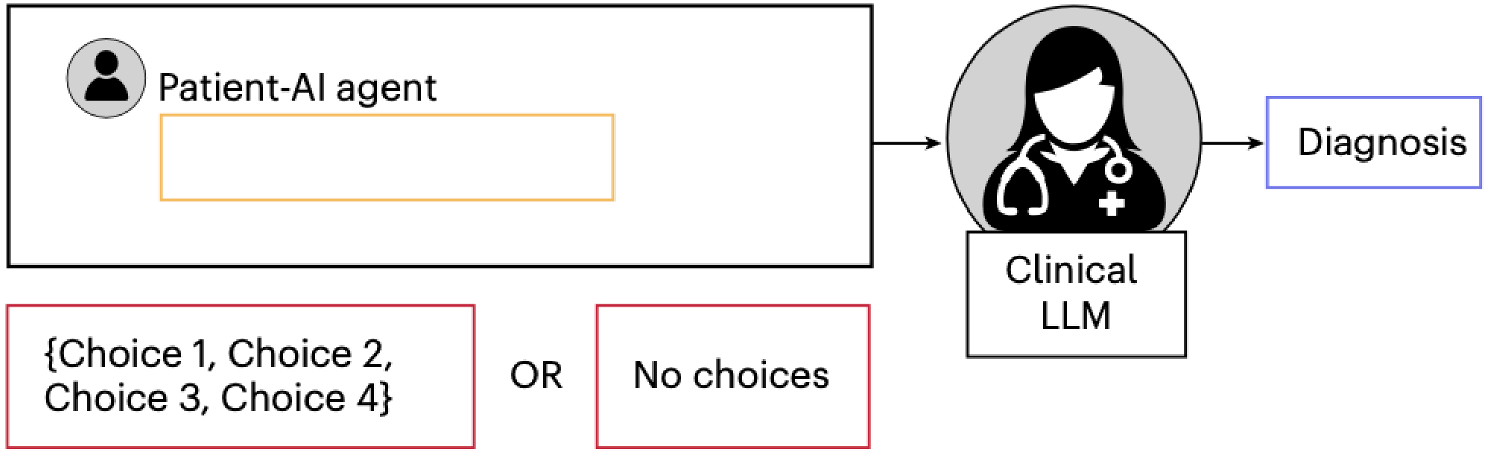
🟩 Multi-turn conversation



Johri, et al. Nat Med. 2025

Evaluations

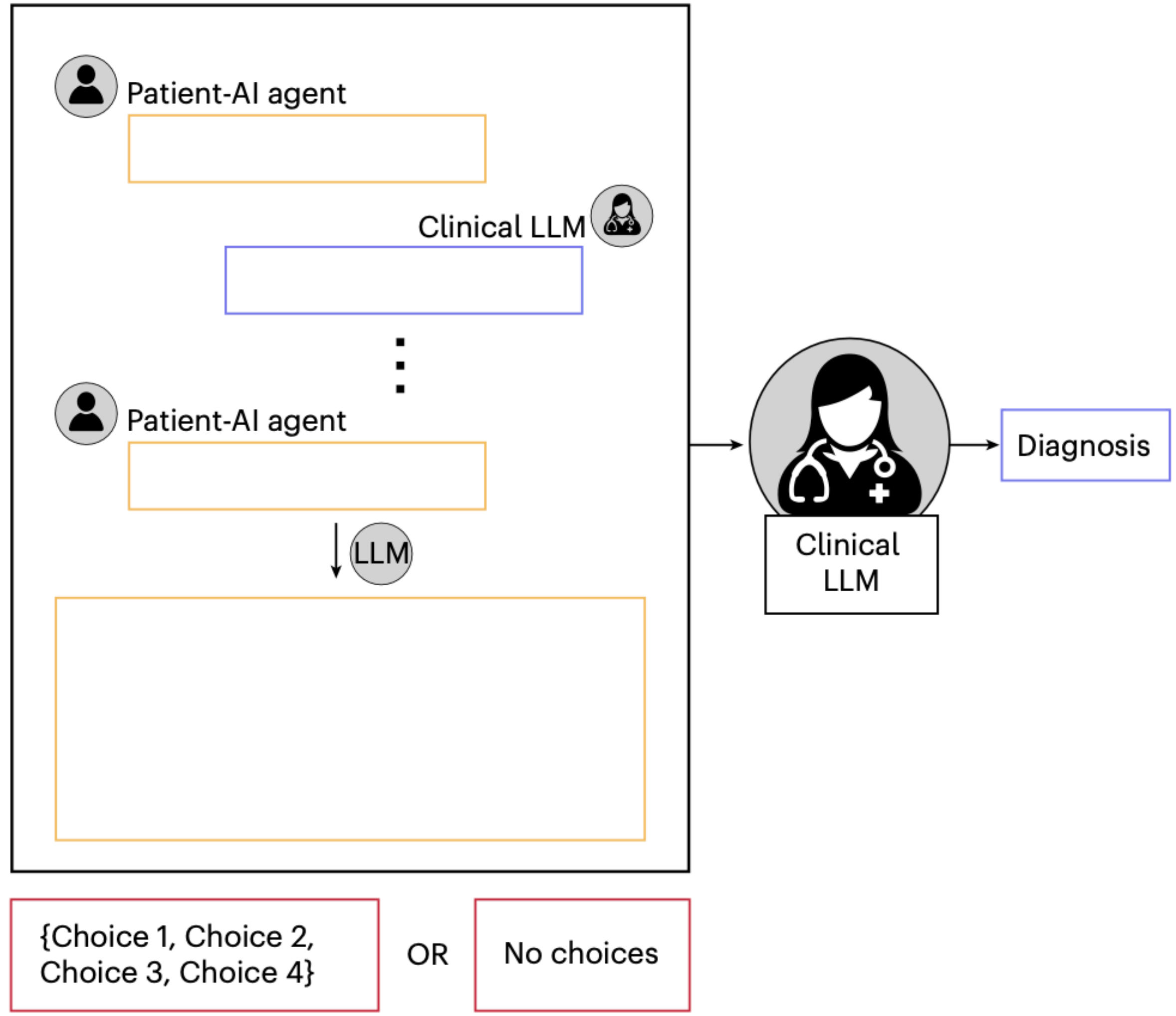
- Single-turn conversation



Johri, et al. Nat Med. 2025

Evaluations

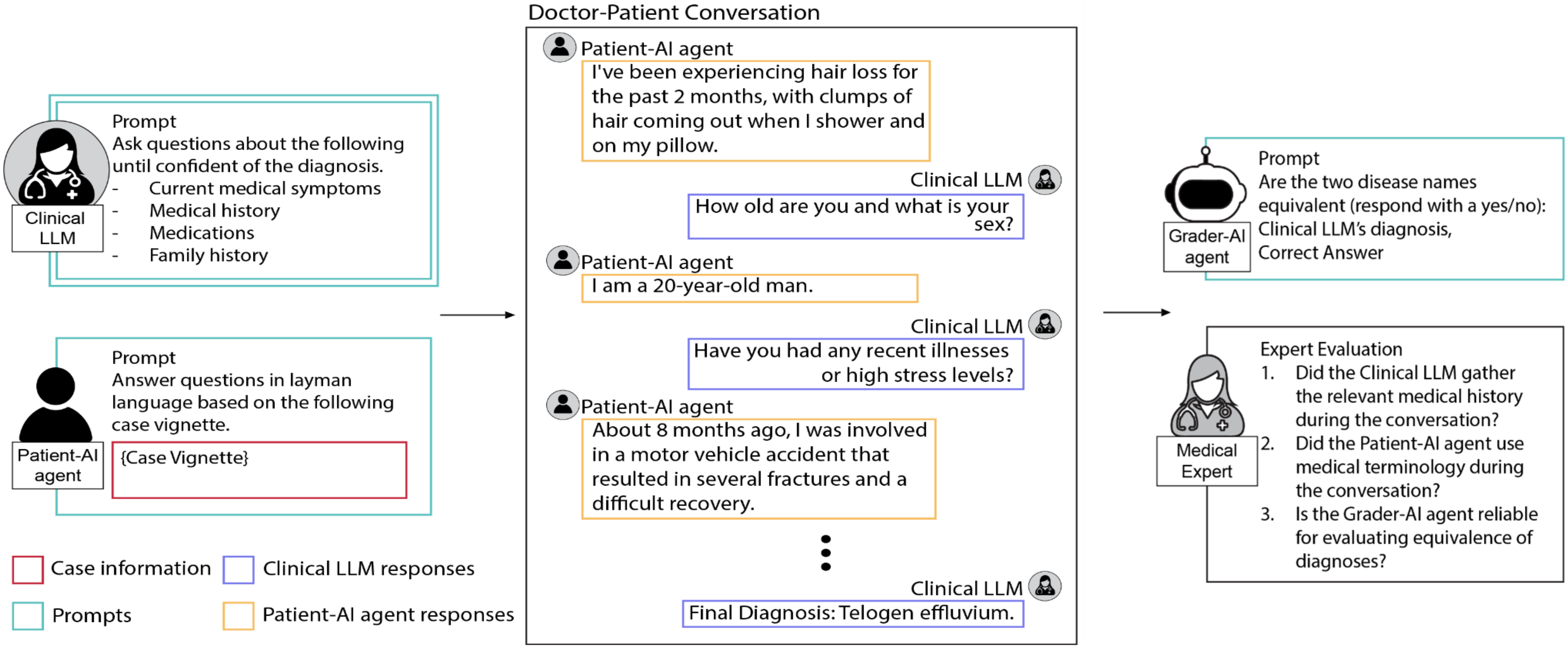
- Summarized conversation



Johri, et al. Nat Med. 2025

Approach: CRAFT-MD

Clinical Reasoning Assessment Framework for Testing in Medicine



Prompt
Ask questions about the following until confident of the diagnosis.
- Current medical symptoms
- Medical history
- Medications
- Family history

Prompt
Answer questions in layman language based on the following case vignette.

{Case Vignette}

🟥 Case information
🟦 Clinical LLM responses
🟩 Prompts
🟧 Patient-AI agent responses

Doctor-Patient Conversation

Patient-AI agent
I've been experiencing hair loss for the past 2 months, with clumps of hair coming out when I shower and on my pillow.

Clinical LLM
How old are you and what is your sex?

Patient-AI agent
I am a 20-year-old man.

Clinical LLM
Have you had any recent illnesses or high stress levels?

Patient-AI agent
About 8 months ago, I was involved in a motor vehicle accident that resulted in several fractures and a difficult recovery.

⋮

Clinical LLM
Final Diagnosis: Telogen effluvium.

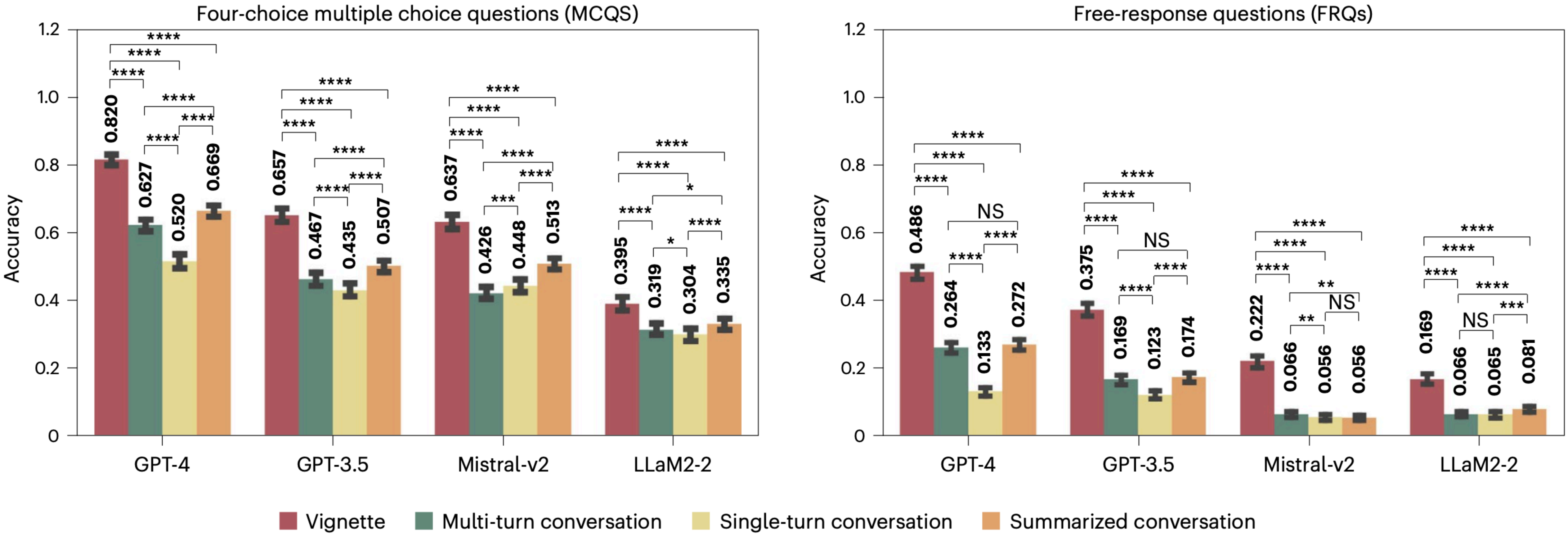
Prompt
Are the two disease names equivalent (respond with a yes/no):
Clinical LLM's diagnosis,
Correct Answer

Expert Evaluation
1. Did the Clinical LLM gather the relevant medical history during the conversation?
2. Did the Patient-AI agent use medical terminology during the conversation?
3. Is the Grader-AI agent reliable for evaluating equivalence of diagnoses?

Do LLMs maintain accuracy when making diagnoses through conversations?

Johri, et al. Nat Med. 2025

Effects of Replacing Case Vignettes with Simulated Doctor-Patient Conversations



Four-choice multiple choice questions (MCQS)

Free-response questions (FRQs)

Johri, et al. Nat Med. 2025

Key Findings

- Conversational interactions reduce diagnostic accuracy

Johri, et al. Nat Med. 2025

Key Findings

- Conversational interactions reduce diagnostic accuracy

- Conversational summarization improves the limited reasoning of LLMs across multiple dialogues

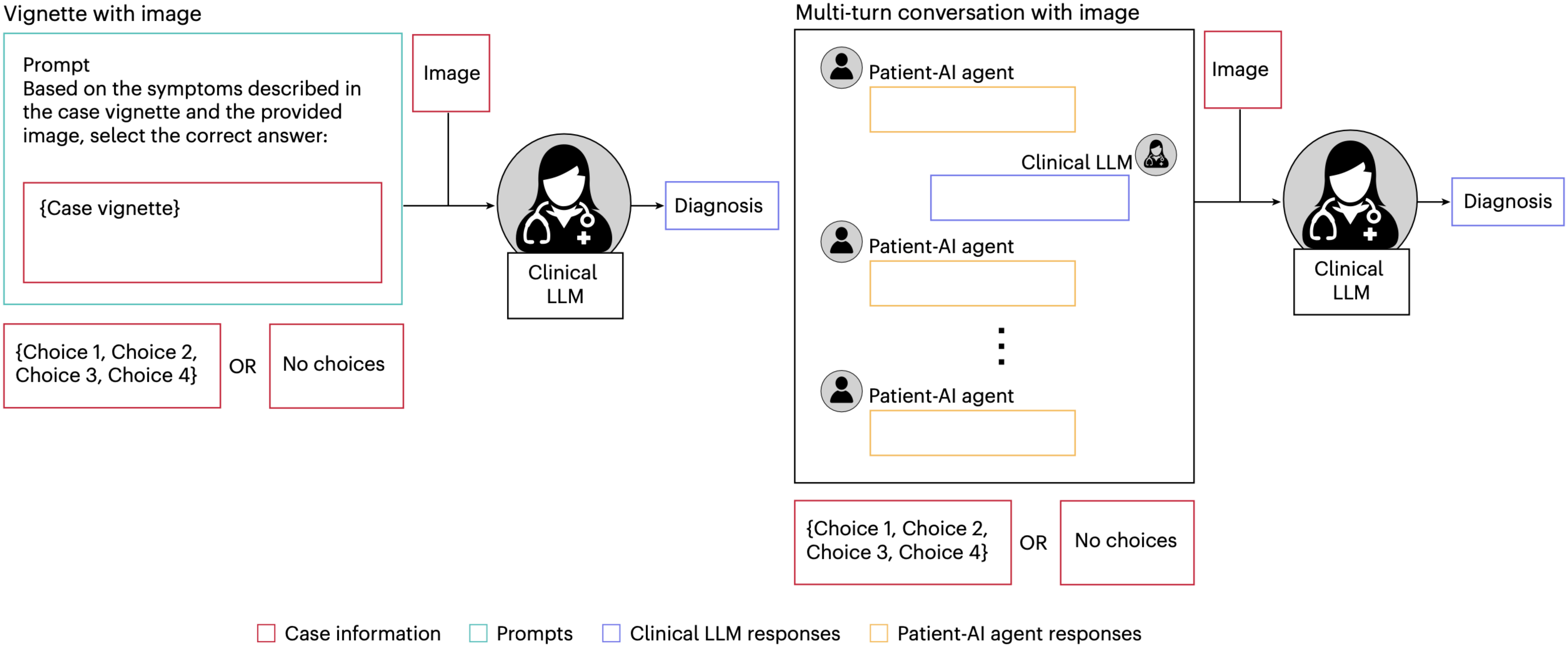
Johri, et al. Nat Med. 2025

Key Findings

- Conversational interactions reduce diagnostic accuracy
- Conversational summarization improves the limited reasoning of LLMs across multiple dialogues
- Trends persist in open-ended diagnoses and across specialties

Johri, et al. Nat Med. 2025

Evaluation of Image Comprehension



Vignette with image

Prompt
Based on the symptoms described in the case vignette and the provided image, select the correct answer:

{Case vignette}

Image

Clinical LLM

Diagnosis

{Choice 1, Choice 2, Choice 3, Choice 4} OR No choices

Multi-turn conversation with image

Patient-AI agent

Clinical LLM

Patient-AI agent

⋮

Patient-AI agent

Image

Clinical LLM

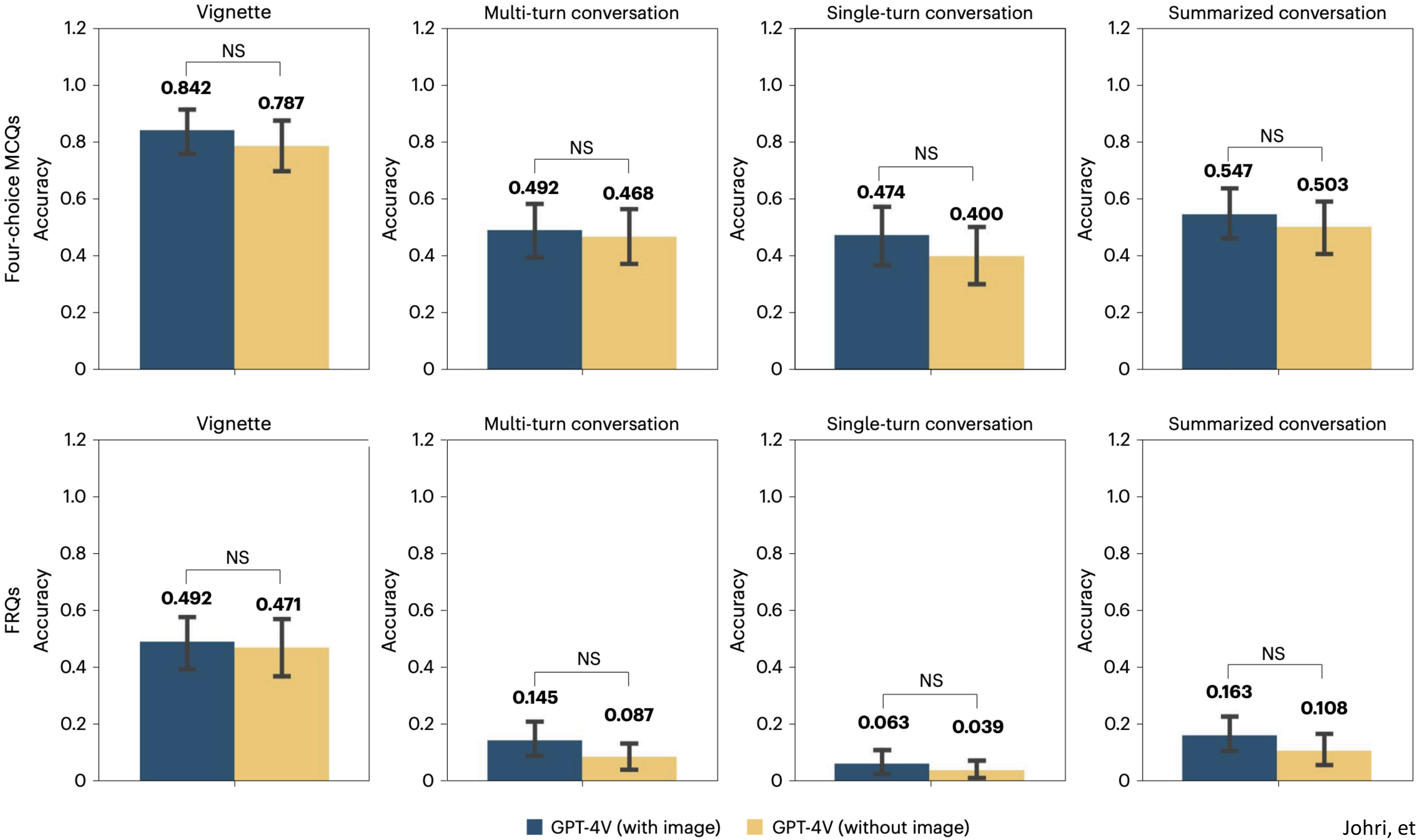
Diagnosis

{Choice 1, Choice 2, Choice 3, Choice 4} OR No choices

□ Case information □ Prompts □ Clinical LLM responses □ Patient-AI agent responses

Johri, et al. Nat Med. 2025

Multimodal Models are Limited in Image Comprehension



Johri, et al. Nat Med. 2025

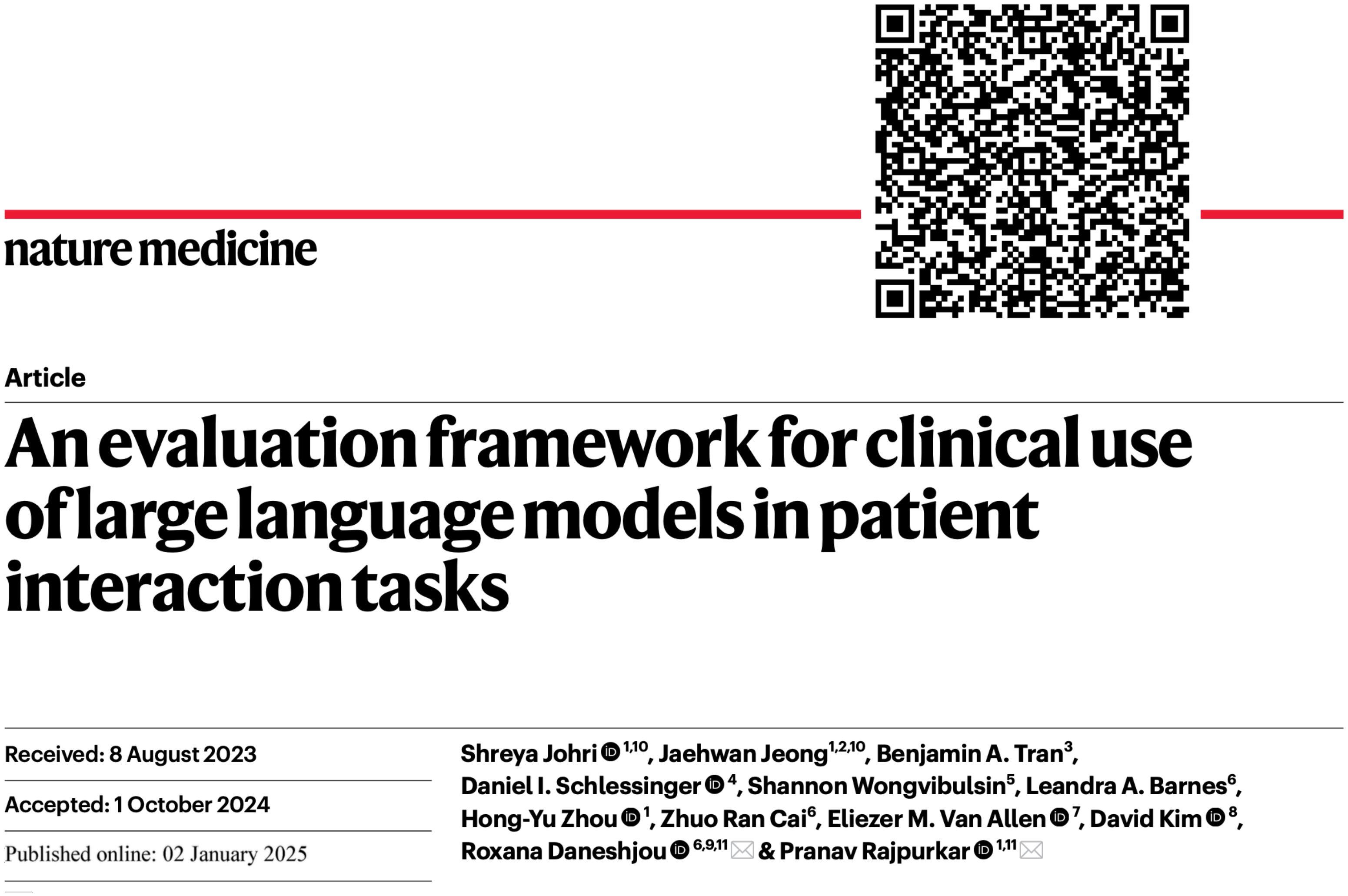
Recommendations for Evaluation of Clinical LLMs

- Evaluate diagnostic accuracy through realistic doctor–patient **conversations**
- Employ **open-ended questions** for evaluating diagnostic reasoning
- Assess comprehensive **history taking** skills
- Evaluate LLMs on the **synthesis** of information from conversations

Johri, et al. Nat Med. 2025

Thank You!

Email:
swongvibulsin@mednet.ucla.edu



nature medicine

Article

An evaluation framework for clinical use of large language models in patient interaction tasks

Received: 8 August 2023

Accepted: 1 October 2024

Published online: 02 January 2025

Shreya Johri[1,10], Jaehwan Jeong[1,2,10], Benjamin A. Tran[3], Daniel I. Schlessinger[4], Shannon Wongvibulsin[5], Leandra A. Barnes[6], Hong-Yu Zhou[1], Zhuo Ran Cai[6], Eliezer M. Van Allen[7], David Kim[8], Roxana Daneshjou[6,9,11] ✉ & Pranav Rajpurkar[1,11] ✉