Clinical Reasoning in Medical AI: Evaluating, Enhancing, and Automating Assessment

Presenting Authors: Liam G. McCoy[1,2,3], Thomas A. Buckley[4], and Ethan Goh[5]

[1]University of Alberta [2]Massachusetts Institute of Technology [3]Beth Israel Deaconess Medical Center [4]Harvard Medical School [5]Stanford University

The Challenge:

- While large language models excel at knowledge-based assessments, their ability to reason through complex, ambiguous clinical scenarios—a cornerstone of medical practice—remains difficult to measure reliably and at scale.
- The gap between benchmark performance and real-world clinical reasoning capabilities presents a critical barrier to responsible implementation of AI in healthcare.
- New methodological approaches are urgently needed—approaches that can rigorously evaluate clinical reasoning, scale efficiently across diverse medical contexts, and provide meaningful comparisons to clinician performance.

Our Approach:

- We draw from long-validated traditions in clinical psychometrics to assess reasoning capabilities in medical AI systems, looking beyond knowledge retrieval to the cognitive processes that define expert clinical judgment.
- The future of clinical AI evaluation requires methodologies that balance rigorous assessment with scalability—enabling systematic testing across diverse medical contexts without prohibitive resource requirements.

Can Large Language Models Improve Physician Management Reasoning?

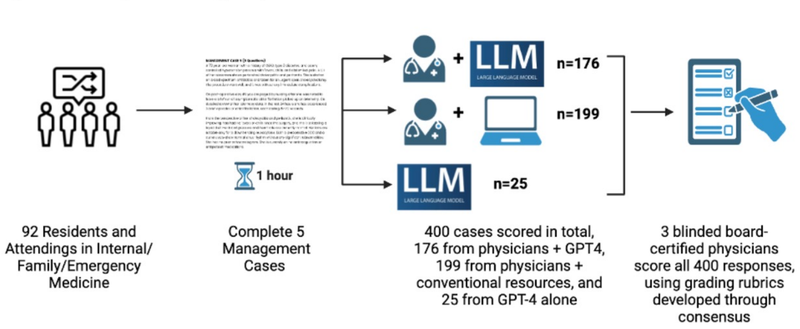
Ethan Goh, Robert J. Gallo, Eric Strong, Yingjie Weng, Hannah Kerman, Jason A. Freed, Joséphine A. Cool, Zahir Kanjee, Kathleen P. Lane, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Jason Hom, Jonathan H. Chen, Adam Rodman

Background

•Management reasoning involves navigating clinical uncertainty, integrating patient context, and weighing competing priorities

•While LLMs have shown high performance on diagnostic benchmarks, their impact on real-world management decisions is unknown

Figure 1: Study Flow Diagram



Scoring approach

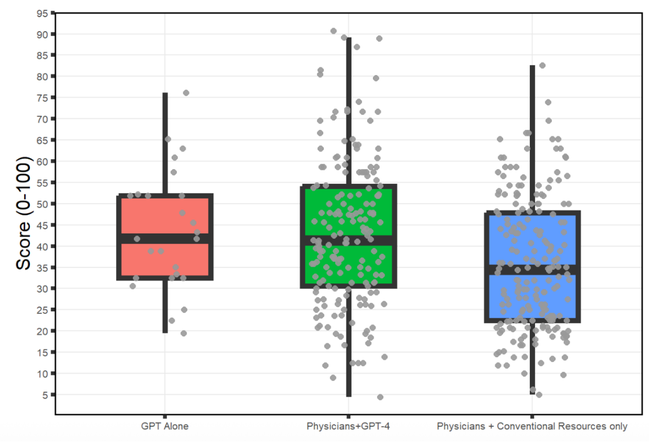
•Responses were evaluated using structured rubrics developed via a modified Delphi process

•Three blinded physician raters scored submissions

•The primary endpoint was total performance score

•Secondary endpoints: domain-specific scores, time spent

Figure 2: Comparisons of the Primary Outcome by GPT Alone vs Physicians with GPT-4 and with Conventional Resources Only (Total score standardized to 0-100)



Main findings

• +6.5 % absolute score gain with GPT-4 assistance

• Superior across management (+6.1 %), diagnosis (+12.1 %), and context (+6.2 %) domains

• GPT-4 alone performed statistically on par with GPT-4-augmented physicians (–0.9 %)

Implications:

• Augmenting clinicians with GPT-4 produced statistically meaningful performance gains

•Comparable scores for GPT-4 alone suggest potential for stand-alone LLM applications in certain clinical scenarios

Check out related Stanford ARISE studies:

Script Concordance Testing: A Novel Framework for Evaluating Reasoning in Medical AI Systems

Liam G. McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, James Cao, Stephen Bacchi, Nigel Fong, Nigel CK Tan, Kevin Tan, Thomas A. Buckley, Peter Brodeur, Leo Anthony Celi, Arjun Manrai, Aloysius Humbert, Adam Rodman

How to evaluate reasoning under uncertainty?

•Script Concordance Tests (SCTs) assess how clinicians integrate new information, and update hypotheses

•Scored against expert baseline, acknowledging the range of clinical practice, as opposed to single answers

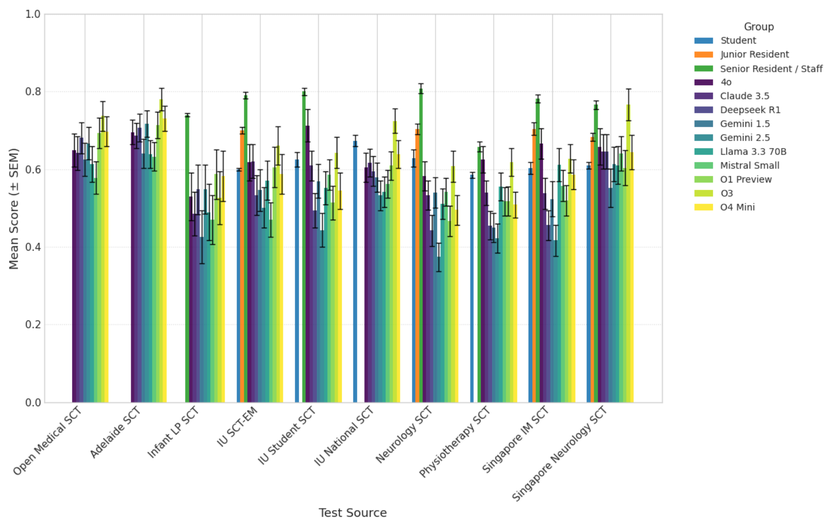
Background Info:	A 65 year old man comes into the emergency department complaining of severe chest pain and shortness of breath
Hypothesis:	If you were thinking of: *a diagnosis of myocardial infarction*
New Info:	And then you find: *that his troponin level is 0*
Question:	How does this impact the likelihood of your hypothesis?

	-2 Much less likely	-1 Somewhat less likely	0 No Change	+1 Somewhat more likely	+2 Much more likely
Expert Answers:	10	2	0	0	0
Points for Answer:	1	0.2	0	0	0

Figure 2-1: *Structure of a Script Concordance Test (SCT) Question*

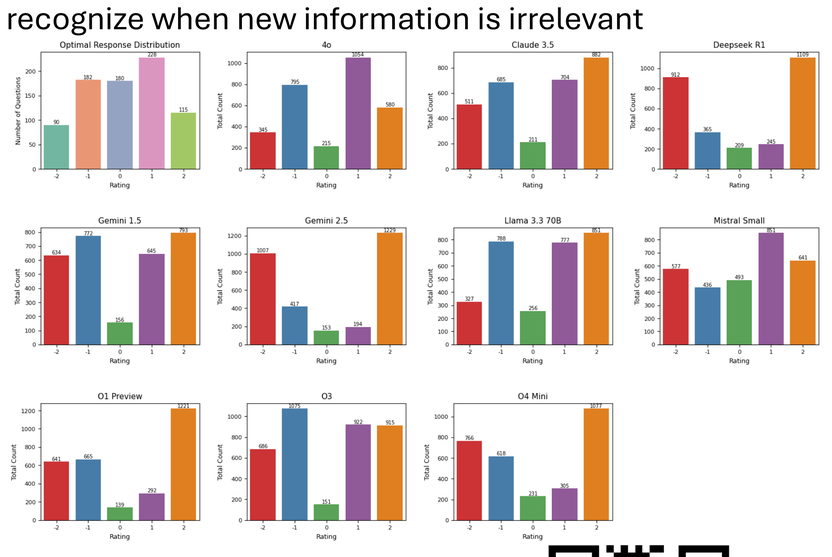
How do models perform?

•Unlike saturated multiple-choice benchmarks, the tested models fail to reach the level of staff doctor performance



Reasoning can lead to overconfidence

•Reasoning models performed particularly poorly, with an excess of extreme probability updates, and failure to recognize when new information is irrelevant



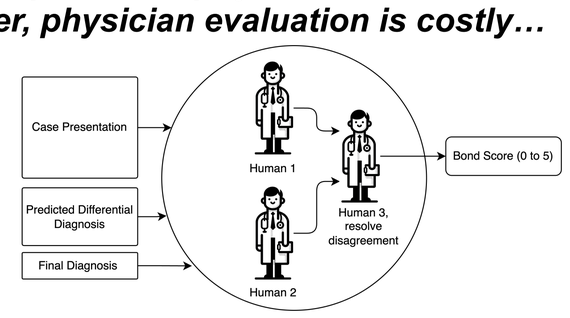
You can check out the benchmark at:

concor.dance

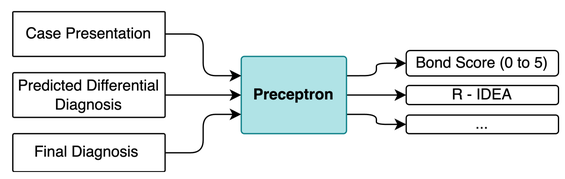
Preceptron: Automated Assessment of Large Language Models in Medicine

Thomas A. Buckley, Zahir Kanjee, Byron Crowe, Anthony M. Pettinato, Aashna P. Shah, Liam G. McCoy, Daniel Restrepo, Ethan Goh, Jonathan H. Chen, Adrian D. Haimovich, Katherine E. Goodman, Daniel J. Morgan, Raja-Elie E. Abdulnour, Laura Zwaan, Adam Rodman, Arjun K. Manrai

LLMs shown to be capable in clinical challenges through qualitative physician evaluation. However, physician evaluation is costly…

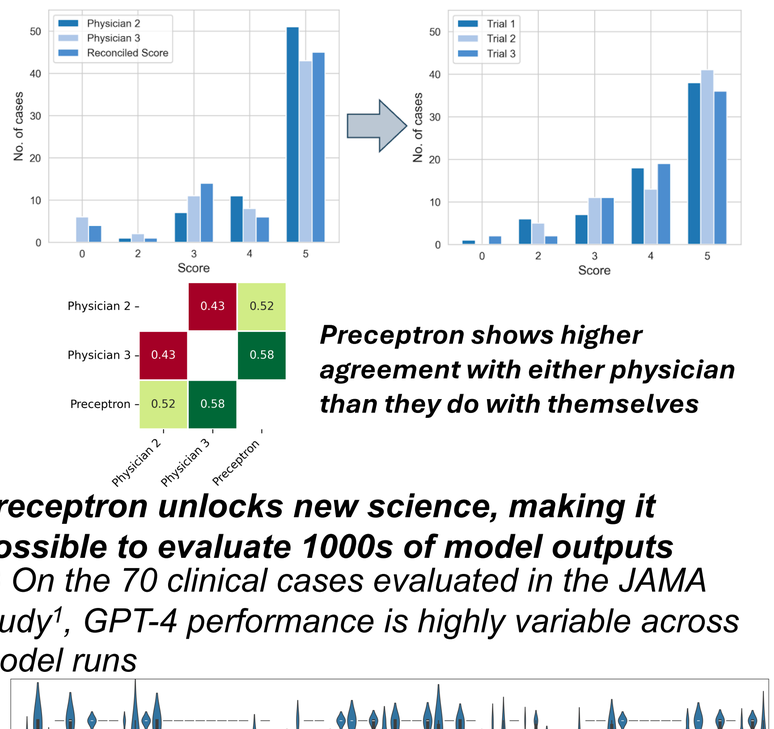


Introducing Preceptron…



Preceptron reproduces an influential JAMA paper[1]

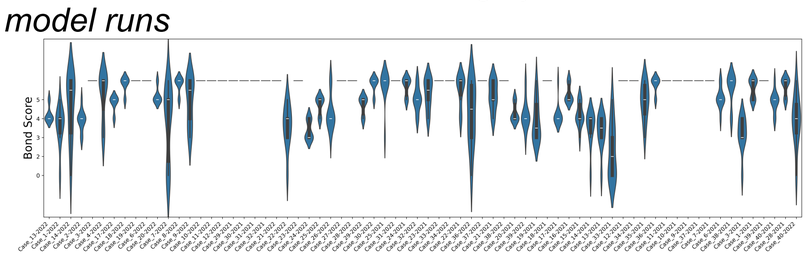
A. JAMA Paper **B. Reproduced Using Preceptron**



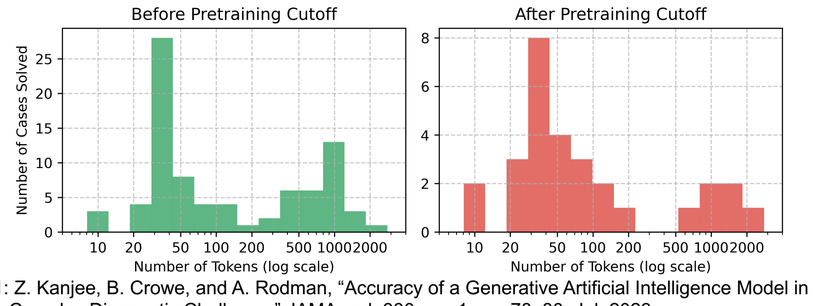
Preceptron shows higher agreement with either physician than they do with themselves

Preceptron unlocks new science, making it possible to evaluate 1000s of model outputs

1) On the 70 clinical cases evaluated in the JAMA study[1], GPT-4 performance is highly variable across model runs



2) Across 143 clinical vignettes[2], GPT-4 can identify the correct diagnosis with just a few sentences of the case presentation



1: Z. Kanjee, B. Crowe, and A. Rodman, "Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge," JAMA, vol. 330, no. 1, pp. 78–80, Jul. 2023.

2: Datasets used: 143 *NEJM* Clinicopathologic Cases (CPCS), published between 2020 and 2024