Brigham and Women's Hospital
Founding Member, Mass General Brigham

The Die is Cast: Two Competing Views of the Future of Physician-AI Collaboration and the Future of the Clinical Mind

Adam Rodman, MD, MPH, FACP
Director of AI Programs, Carl J Shapiro Institute for Education and Research
Division of General Internal Medicine
Beth Israel Deaconess Medical Center
Assistant Professor of Medicine
Harvard Medical School

Adam Rodman

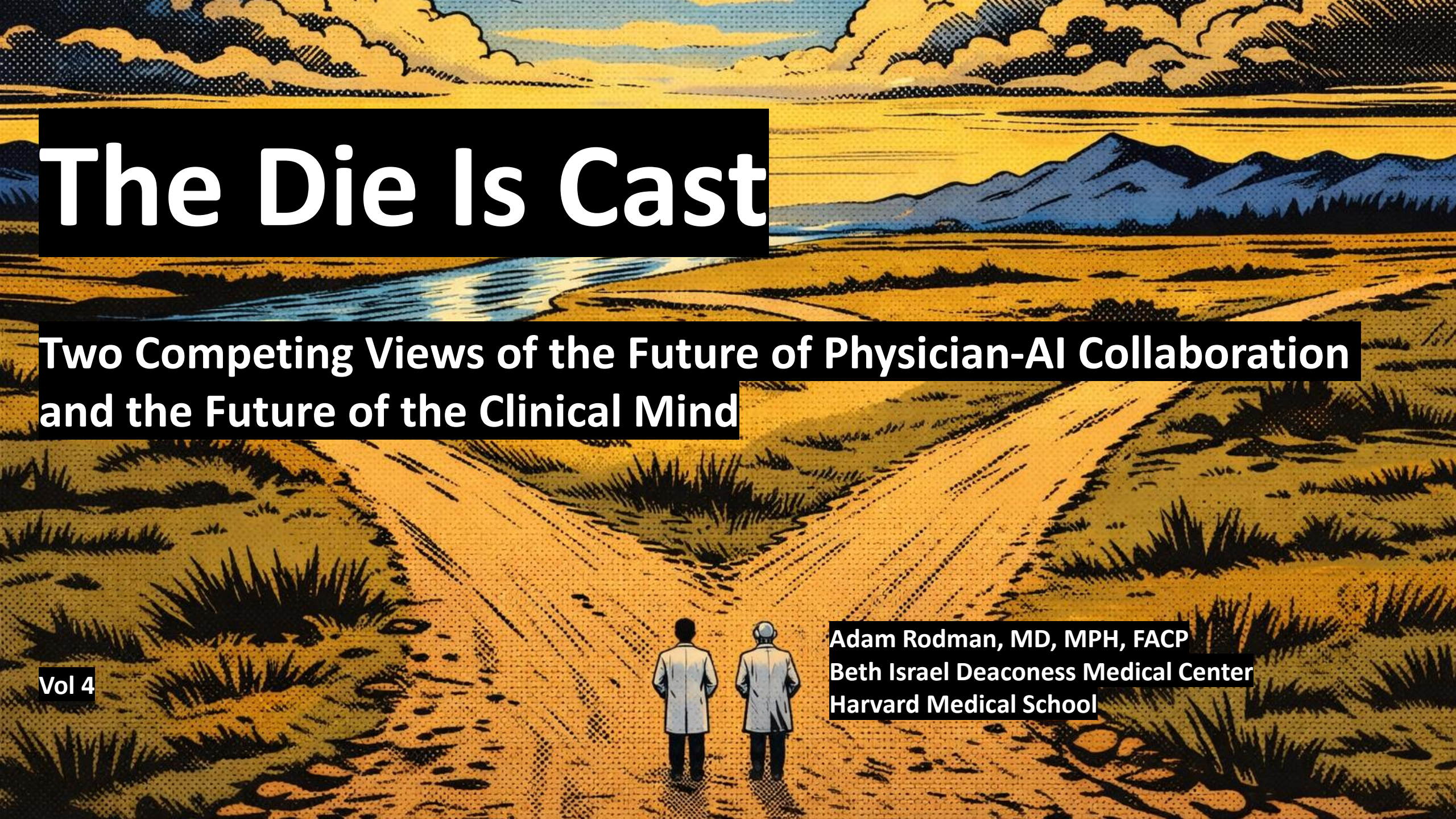
Tulane School of Medicine
Medicine Residency at Oregon Health and Science University
Global Health Fellowship at BIDMC in Botswana
Assistant Professor of Medicine at HMS
Clinical focus: human-computer interaction

DISCLOSURES

- Dr. Rodman is a visiting research at Google DeepMind
- Dr. Rodman reports research funding from the Moore Foundation, Macy Foundation, Google Research, NIH, ARPA-H.

OBJECTIVES

1. Develop a mental model of the various psychological theories that explain clinical reasoning, including dual process theory, ecological psychology, distributed cognition, and collective intelligence.
2. Review the literature on the performance of LLMs in reasoning tasks and the "performance paradox"
3. Understand the concept of human-in-the-loop, -on-the-loop, and -out-of-the-loop collaboration systems and their relevance to clinical reasoning.


The Die Is Cast
Two Competing Views of the Future of Physician-AI Collaboration and the Future of the Clinical Mind
Vol 4
Adam Rodman, MD, MPH, FACP
Beth Israel Deaconess Medical Center
Harvard Medical School

Previous titles:

Vol 1: A Turing Test for Clinical Reasoning
Vol 2: Towards an AI Second Opinion
Vol 3: Towards a Medical Superintelligence
Vol 4: The Die Is Cast

POLITICO

EXCLUSIVE

Artificial intelligence begins prescribing medications in Utah

Pilot program will test how far patients and regulators are willing to trust AI in medicine.

Doctronic

By YASMIN KHORRAM and RUTH READER
01/06/2026 10:00 AM EST

Patient-facing AI company Doctronic launches trial in Utah for (semi)autonomous prescription refills for $4 each

IMAGE CREDITS: LOTUS HEALTH

VENTURE

Lotus Health nabs $35M for AI doctor that sees patients for free

Marina Temkin — 9:14 AM PST · February 3, 2026

EXCLUSIVE INDUSTRY NEWS

UpDoc's AI Gets FDA Nod to Act as 'Concierge Doctor' Between Visits

The venture-backed startup says its AI can adjust medication doses between doctor visits

By Brian Gormley

WSJ PRO June 25, 2026 6:00 am ET

Aa Listen (1 min)

UpDoc's FDA clearance was an important factor in the Cleveland Clinic's decision to test the technology. PHOTO: VALERIE PLESCH/BLOOMBERG NEWS

January 7, 2026 Product

Introducing ChatGPT Health

A dedicated experience in ChatGPT designed for health and wellness.

Join the waitlist ↗

TECH

OpenAI acquires health-care technology startup Torch for $60 million, source says

PUBLISHED MON, JAN 12 2026•4:35 PM EST | UPDATED 2 HOURS AGO

Ashley Capoot
@/IN/ASHLEY-CAPOOT/

SHARE

KEY POINTS

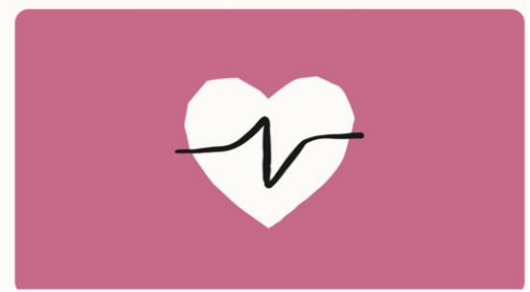
- OpenAI has acquired the health-care technology startup Torch.
- Torch was building a "unified medical memory" for AI that aimed to bring a patient's health data into one place.
- OpenAI and Torch did not disclose the terms of the acquisition.

TRENDI

Advancing Claude in healthcare and the life sciences

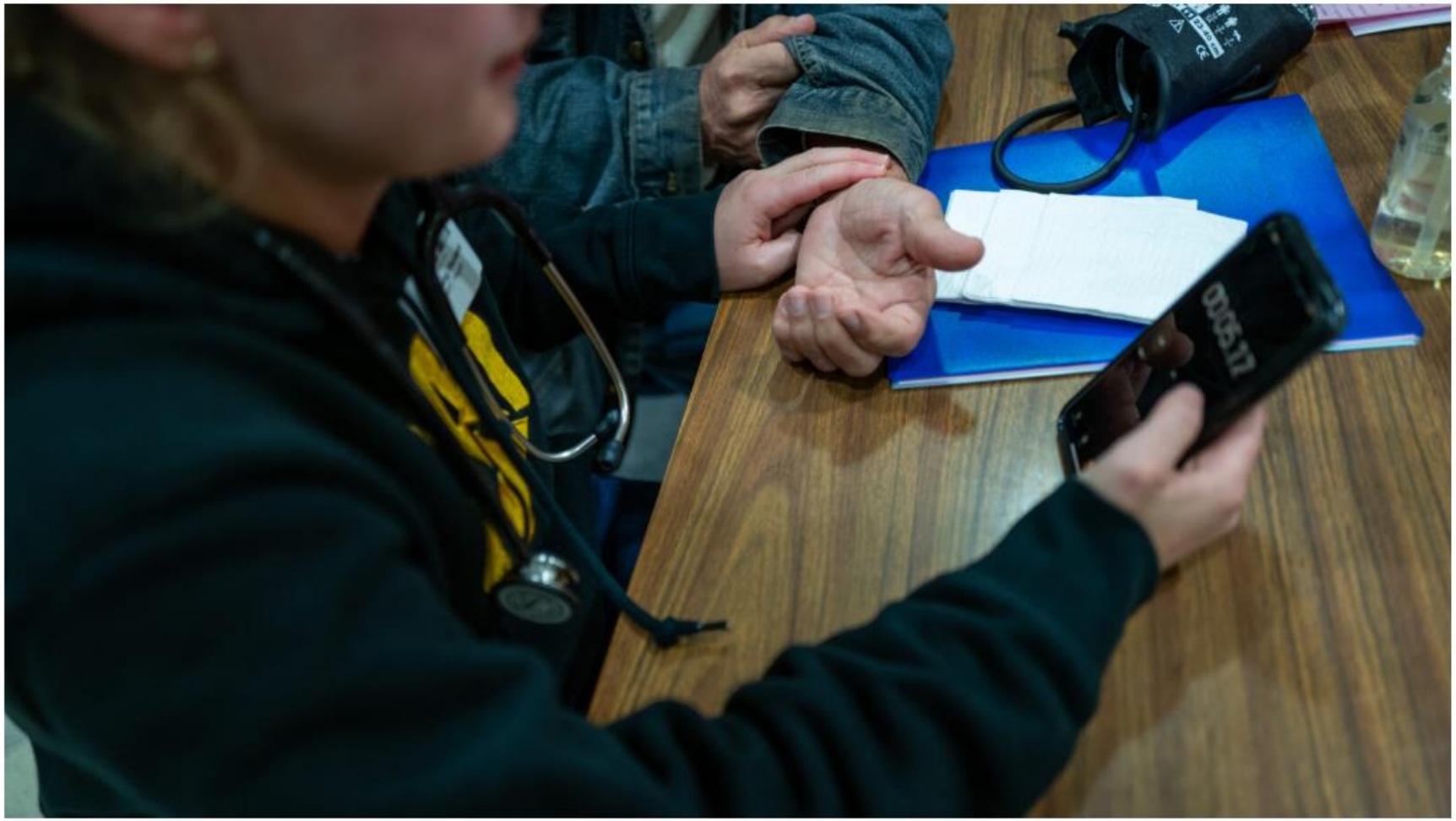
Jan 11, 2026

Join today's livestream

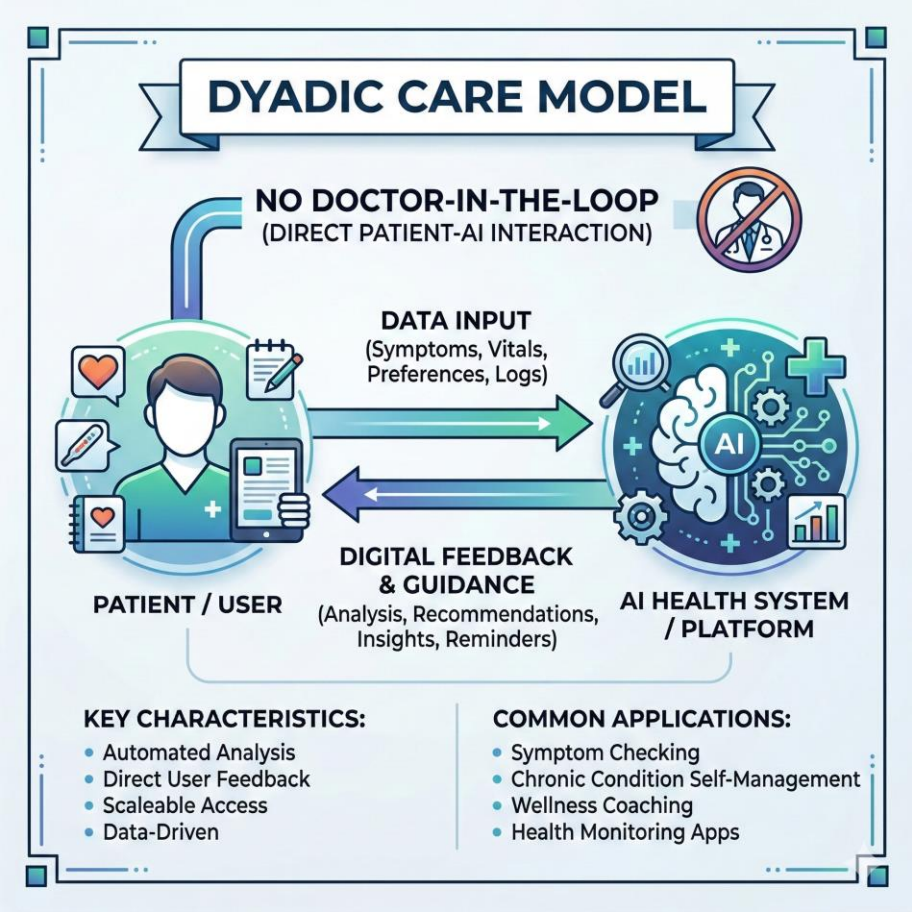


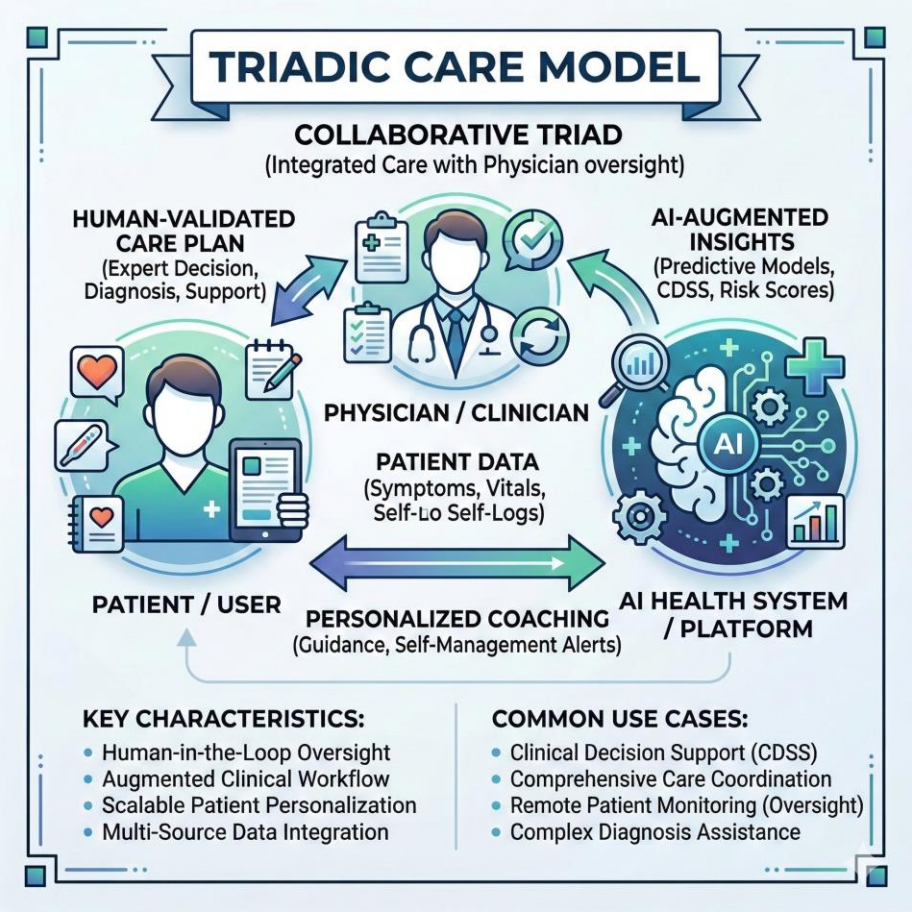
46% of U.S. counties don't have a cardiologist. ARPA-H's new agentic AI program could bring them specialized care

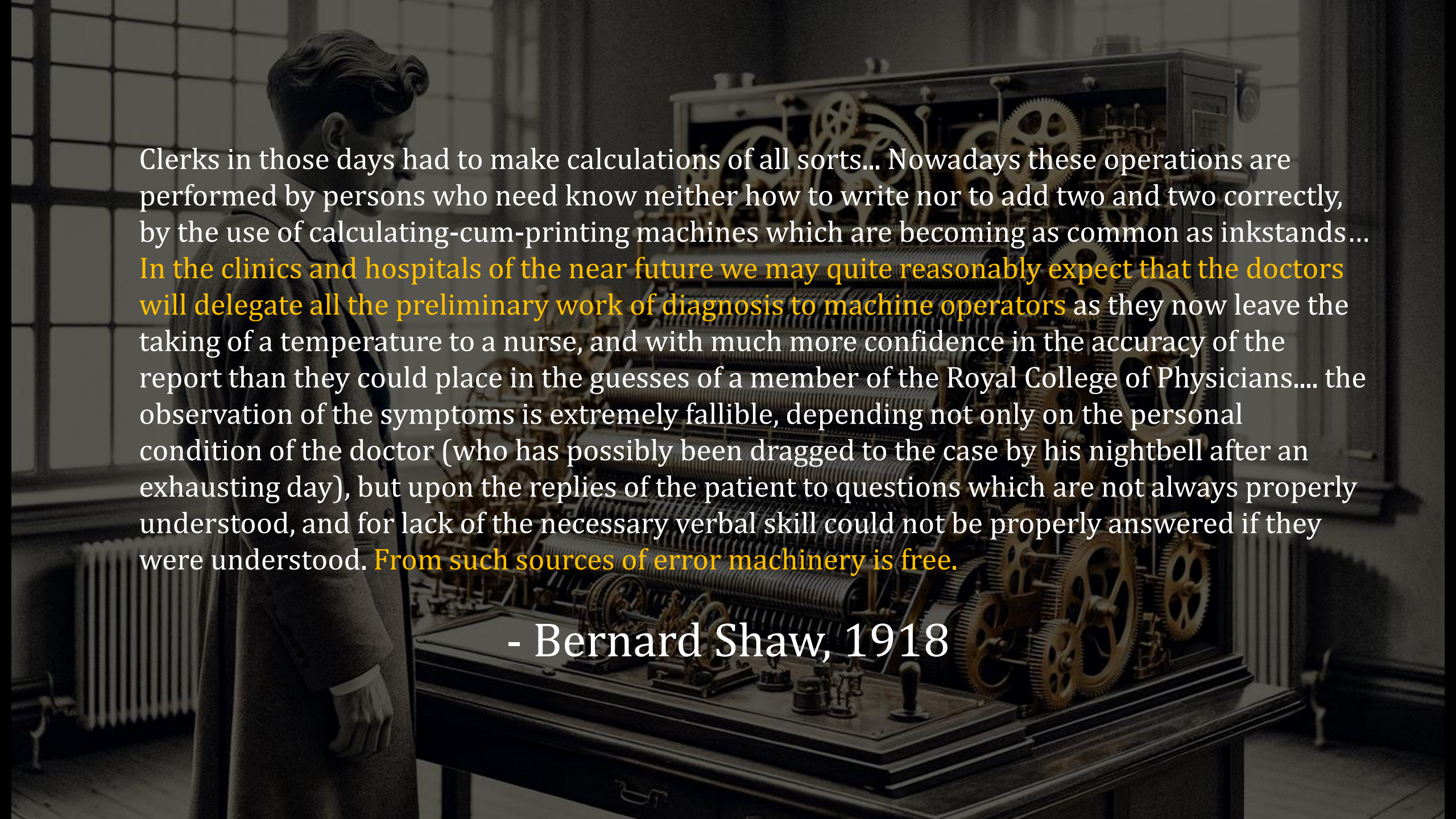
ADVOCATE aims to support every American living with advanced heart disease



Dyadic versus Triadic Care



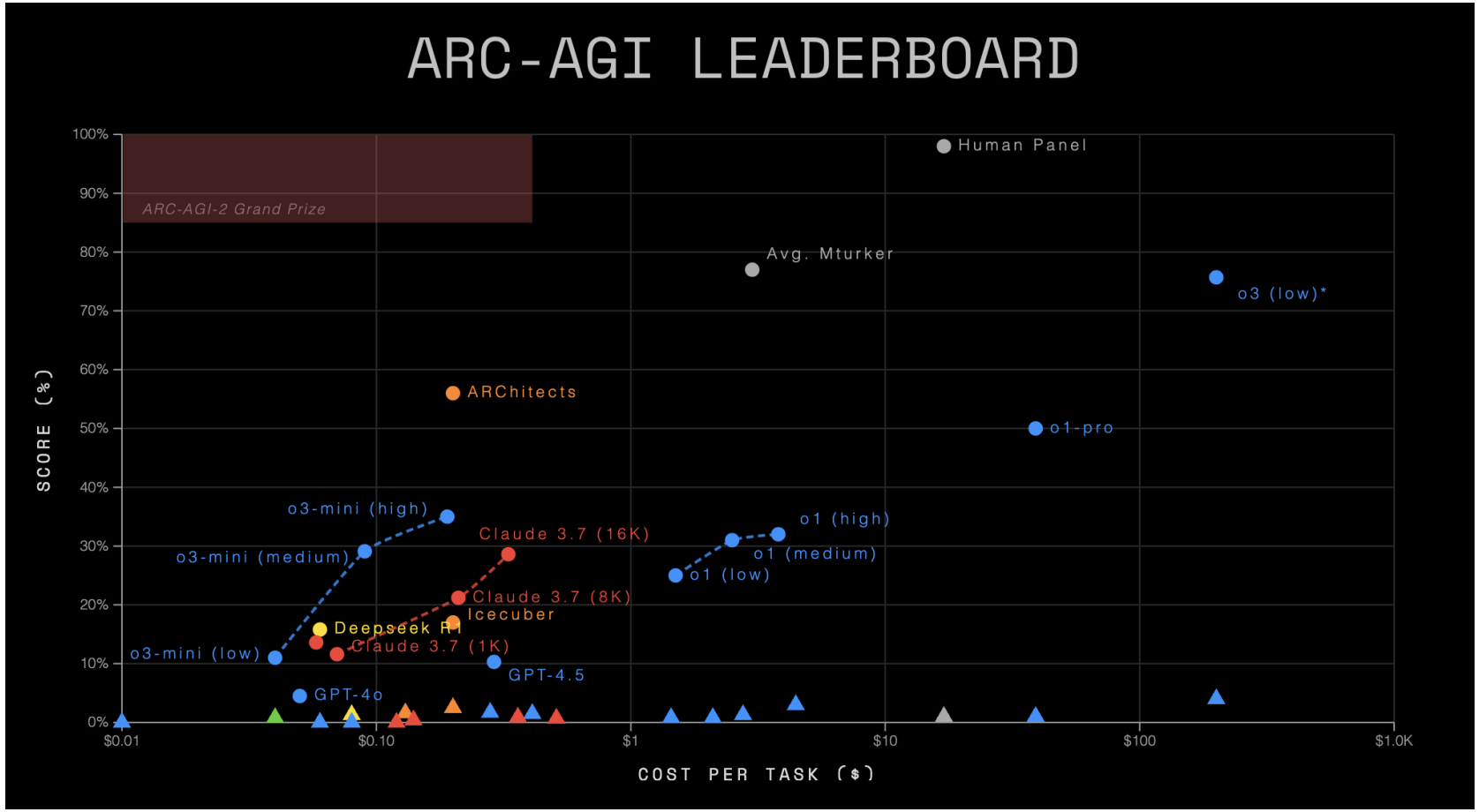




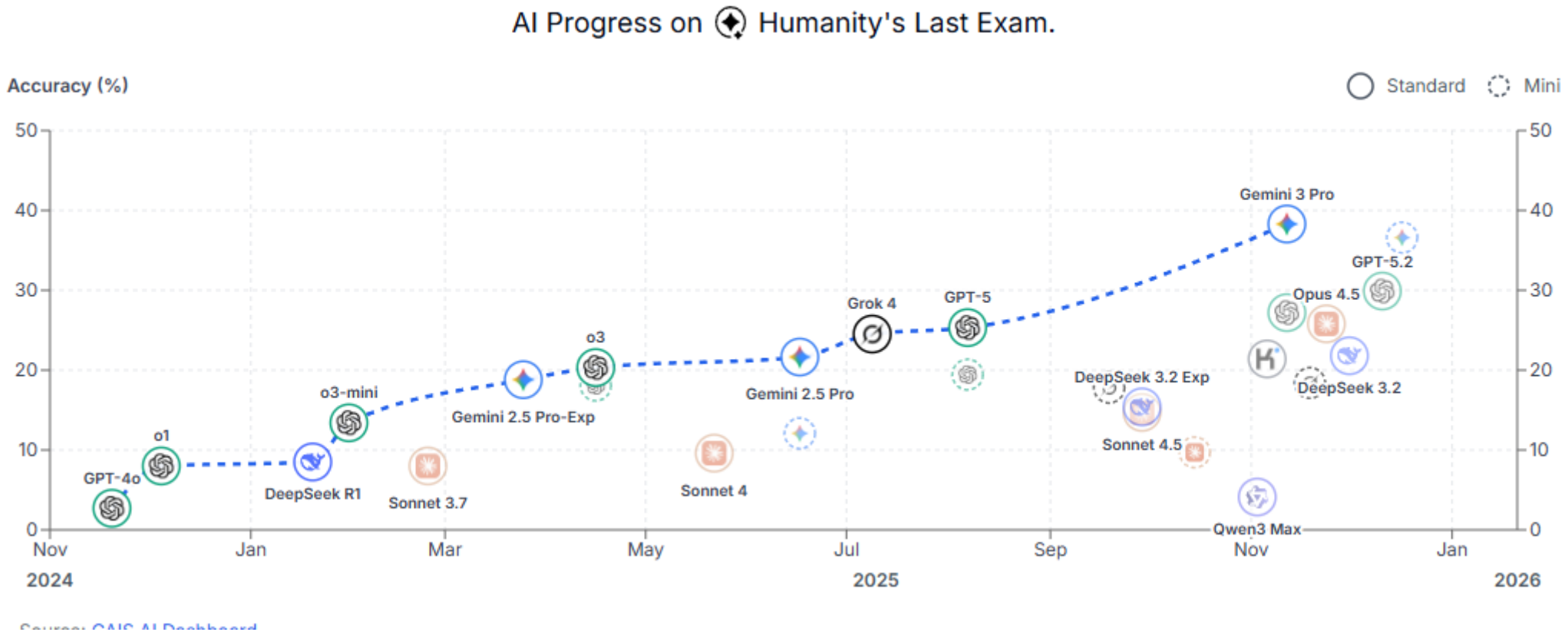
Clerks in those days had to make calculations of all sorts... Nowadays these operations are performed by persons who need know neither how to write nor to add two and two correctly, by the use of calculating-cum-printing machines which are becoming as common as inkstands… In the clinics and hospitals of the near future we may quite reasonably expect that the doctors will delegate all the preliminary work of diagnosis to machine operators as they now leave the taking of a temperature to a nurse, and with much more confidence in the accuracy of the report than they could place in the guesses of a member of the Royal College of Physicians.... the observation of the symptoms is extremely fallible, depending not only on the personal condition of the doctor (who has possibly been dragged to the case by his nightbell after an exhausting day), but upon the replies of the patient to questions which are not always properly understood, and for lack of the necessary verbal skill could not be properly answered if they were understood. From such sources of error machinery is free.

- Bernard Shaw, 1918

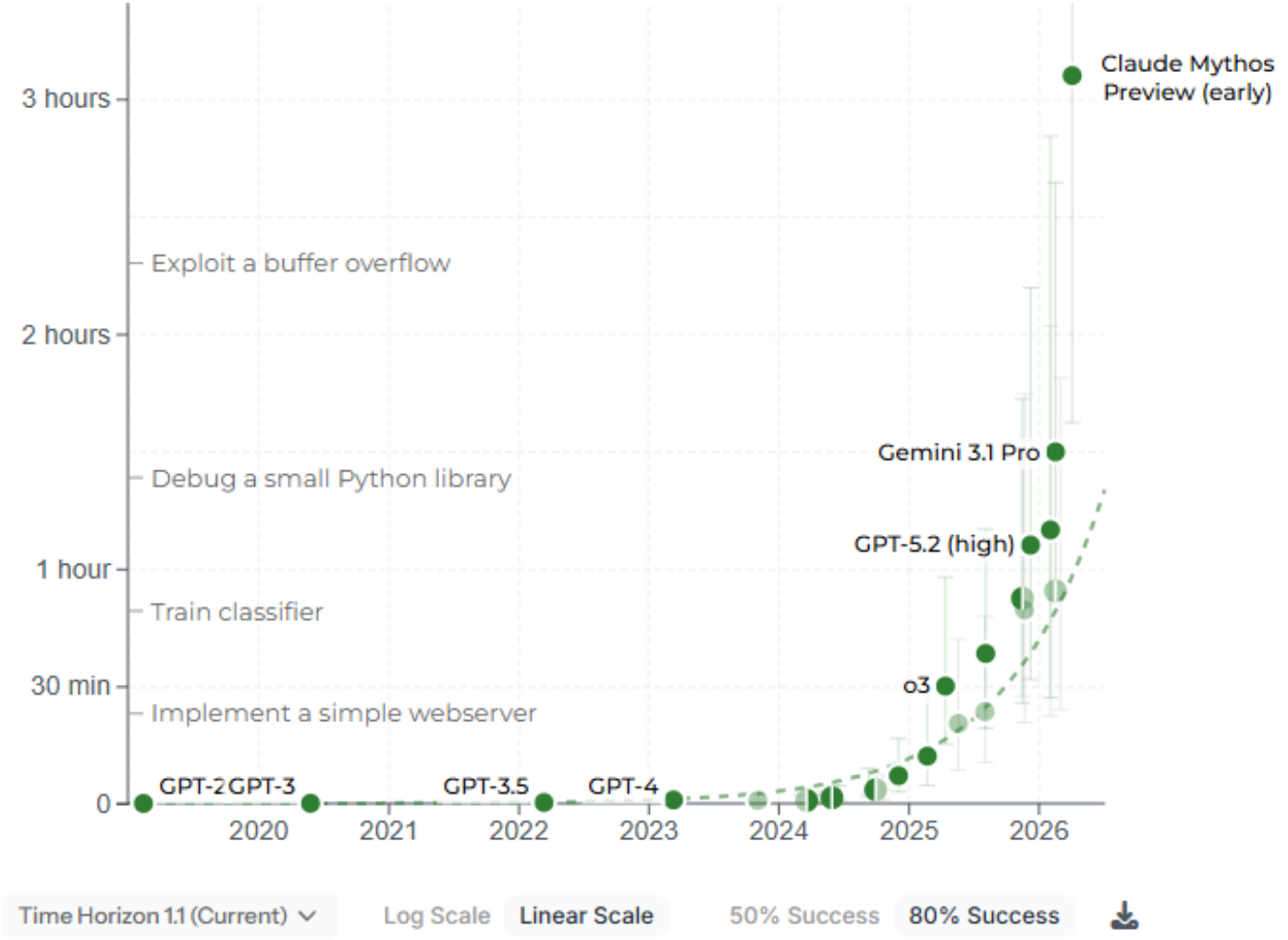
Reasoning models have changed the game …



… real-time inference has not yet hit a wall

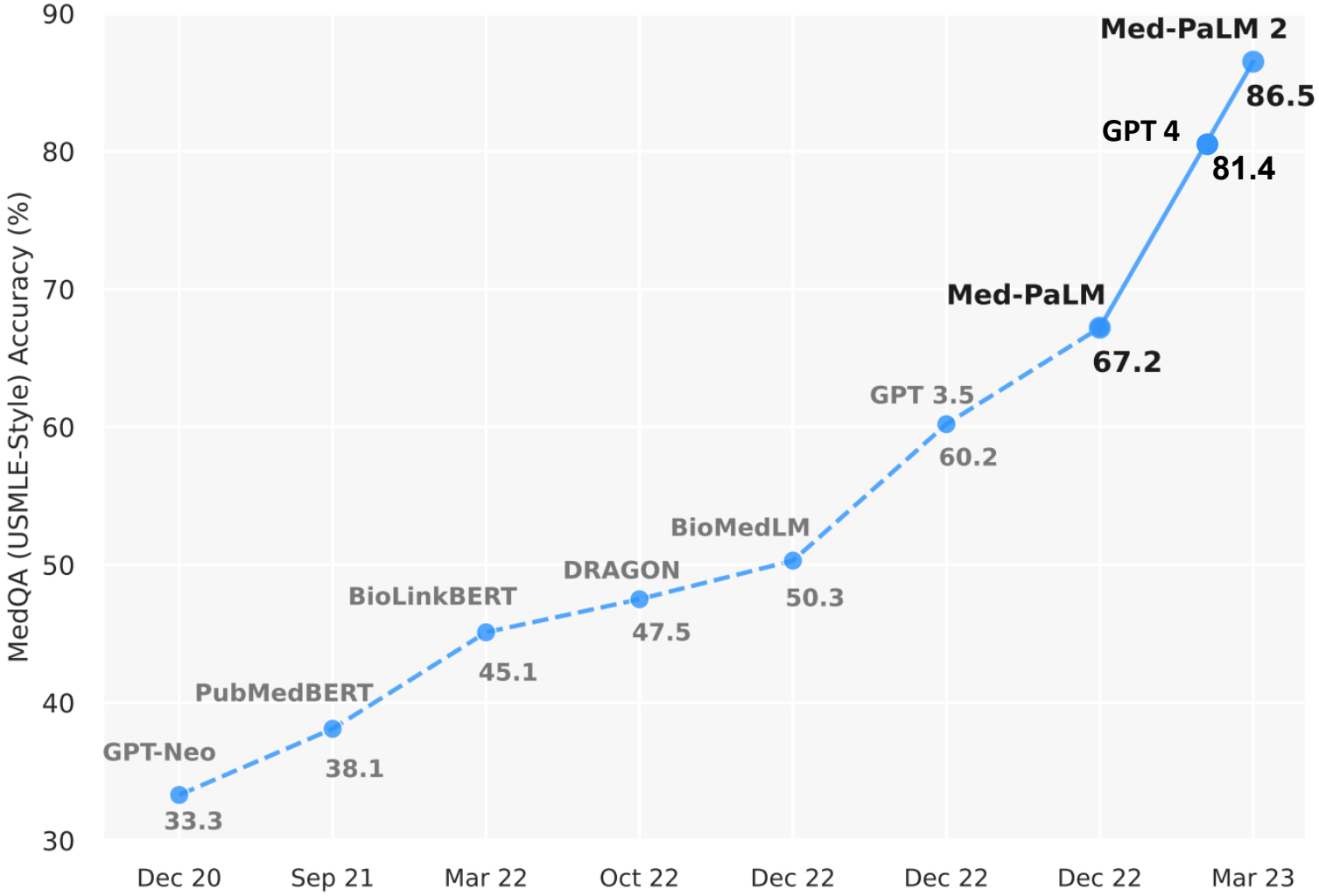


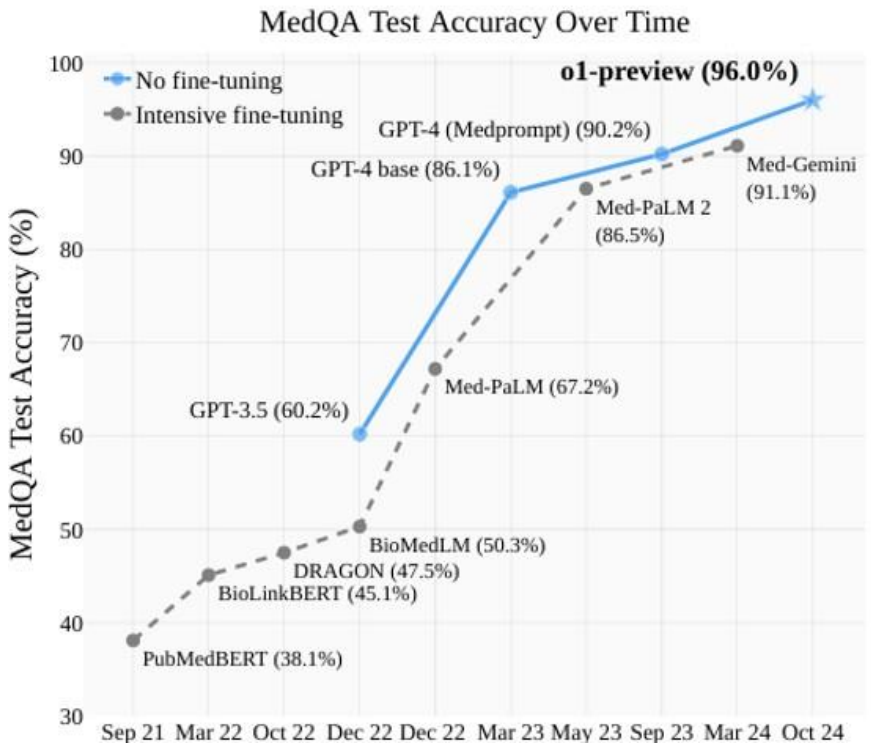
Agents are here and capabilities continue to increase



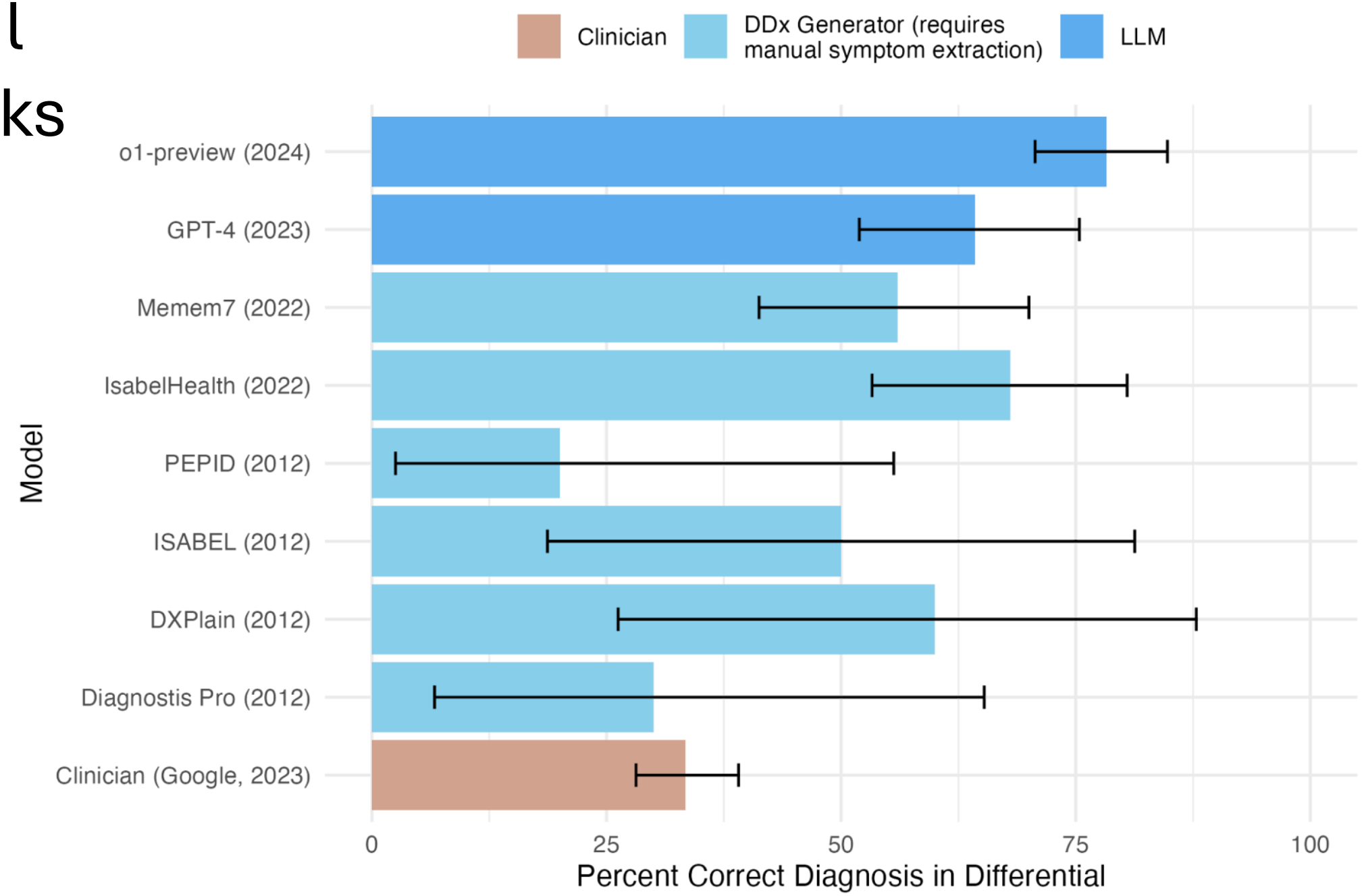
What do we know about LLM performance in medicine?

Board exams are solved



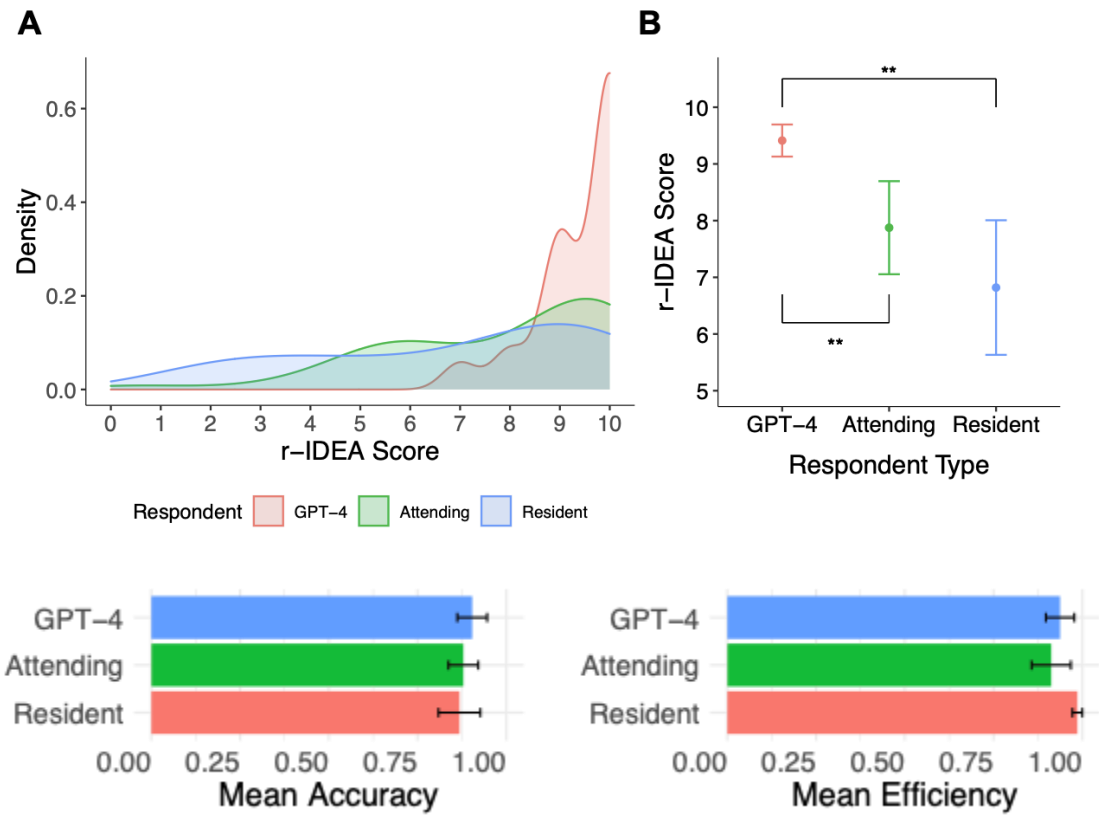


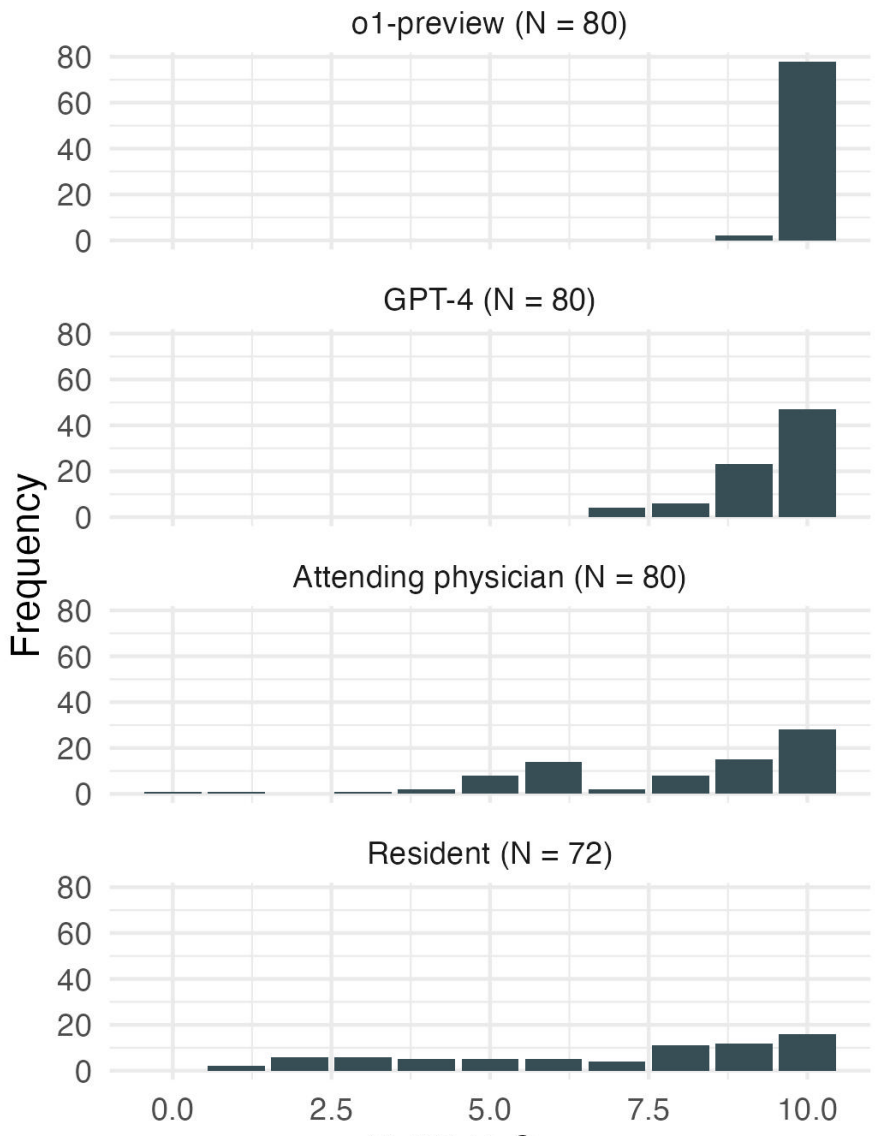
Differential benchmarks are solved



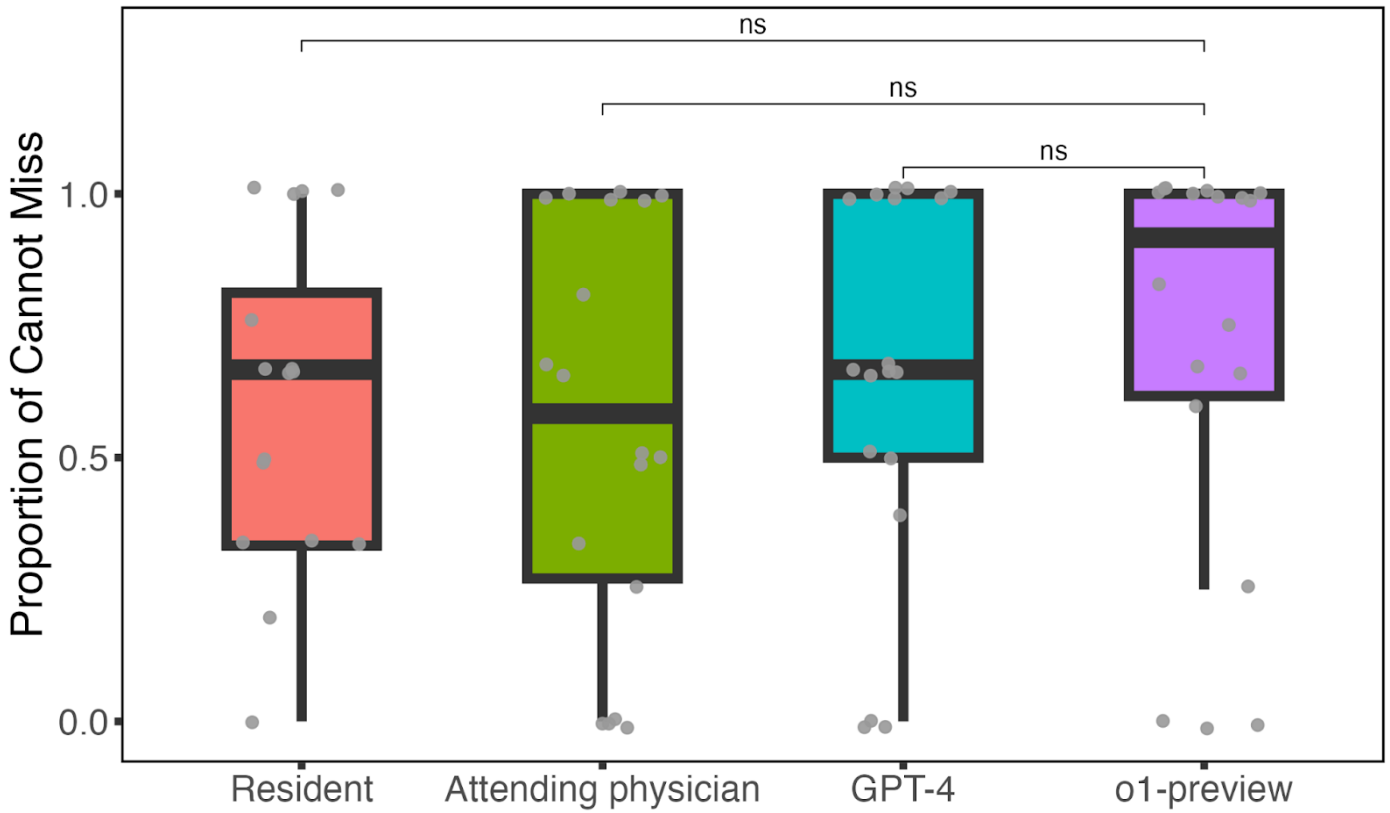
LLMs demonstrate superior reasoning to humans – and are equivalent in process*

- Prospective study of residents, attending, and GPT-4 solving NEJM Healer cases – 236 sections in total
- GPT-4 had significantly higher r-IDEA scores (9.41 vs 7.83 for attendings and 6.82 for residents)
- No difference in efficiency, accuracy, quality, cannot miss
- *Increase of incorrect reasoning (12% vs 3%), though all minor examples





… and these are saturated too.



LLMs can outperform the best doctors on displays of reasoning

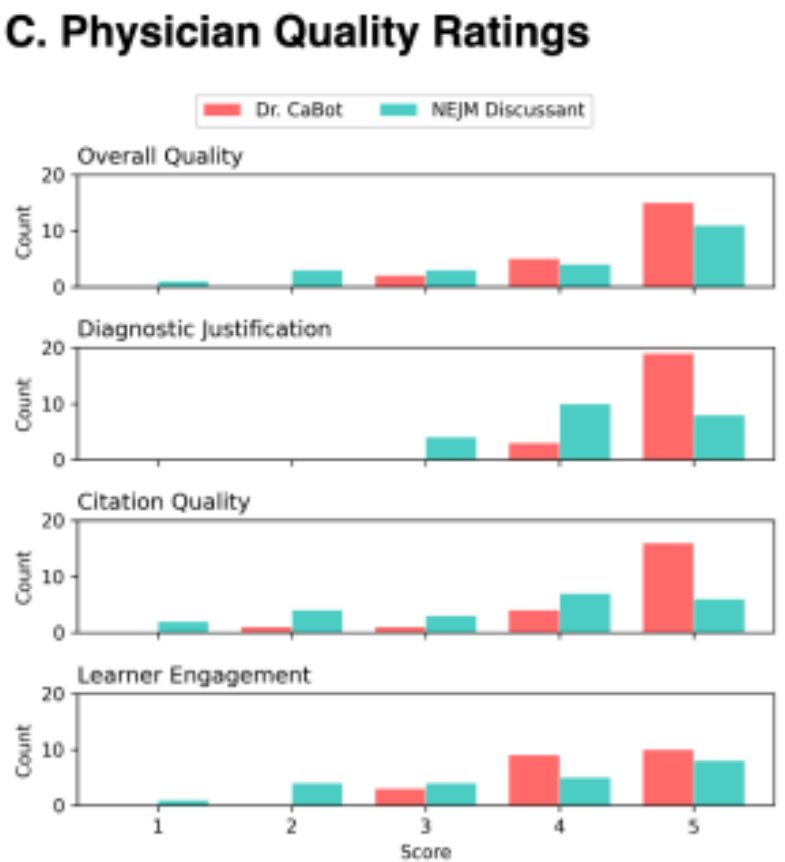
BRAVE NEW WORLD DEPT.

IF A.I. CAN DIAGNOSE PATIENTS, WHAT ARE DOCTORS FOR?

Large language models are transforming medicine—but the technology comes with side effects.

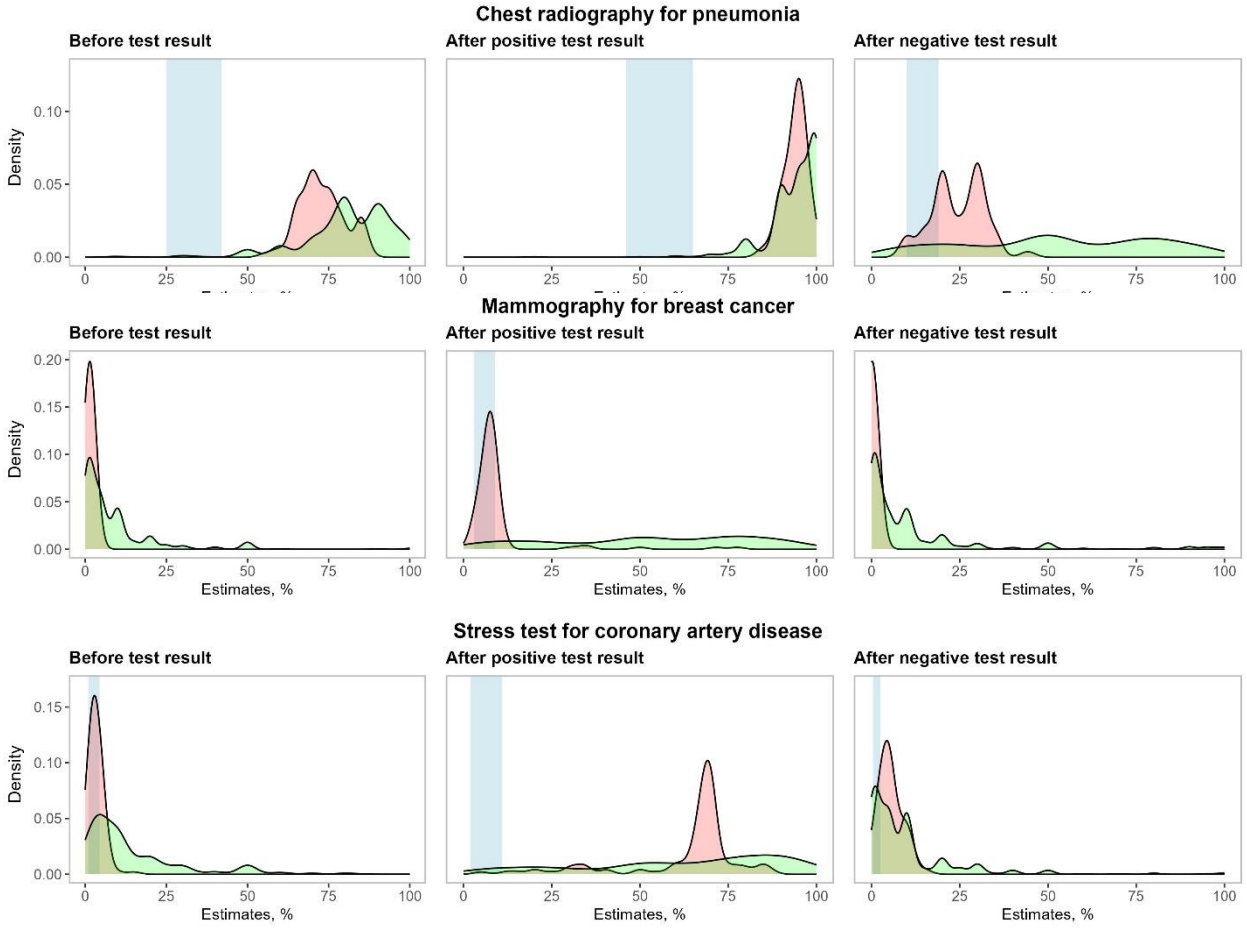
By Dhruv Khullar

September 22, 2025

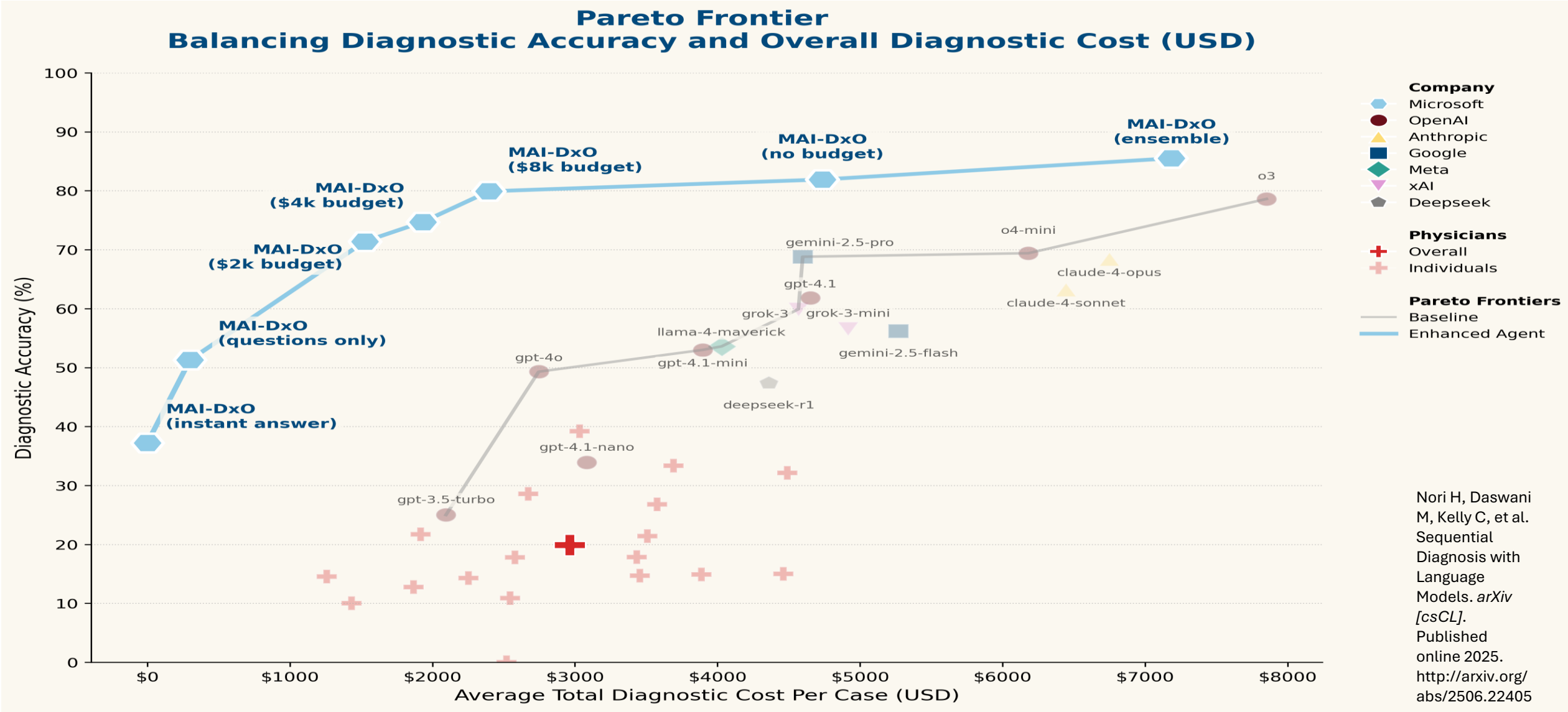


LLMs have emergent probabilistic reasoning…

- Error distributions of LLM (GPT-4) and 553 clinicians
- GPT-4 with much less MAE in all cases of pre-test probability and post-test after a negative; equivalent after positive

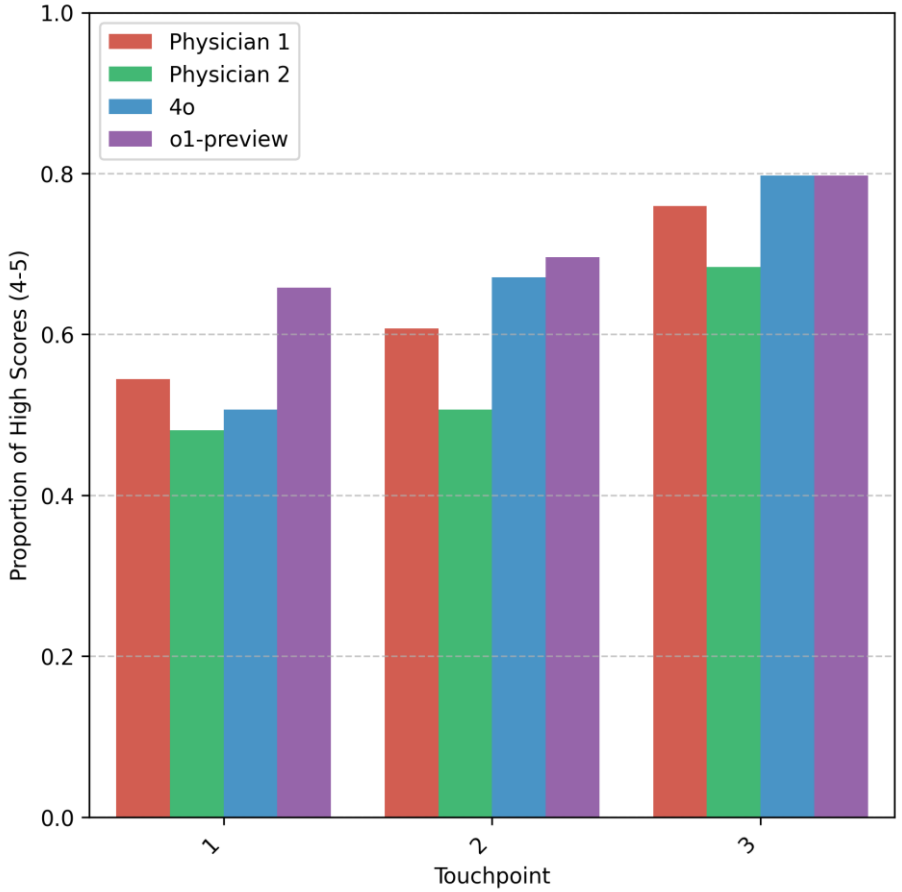


… and can choose the “next best test.”



Nori H, Daswani M, Kelly C, et al. Sequential Diagnosis with Language Models. *arXiv [csCL]*. Published online 2025. http://arxiv.org/abs/2506.22405

They perform well on real patients



- "All comers" to emergency room over a two week period; at three pre-defined touchpoints both experienced hospitalists **and** AI models gave "second opinions"
- Reasoning models outperformed at lower information density and were equivalent at higher information density.

And they have real clinical impacts

- Randomized QI study in at Penda Health in Nairobi with 40,000 patients using GPT-4o
- Significant error decreases – 16% fewer diagnostic errors and 13% fewer treatment errors

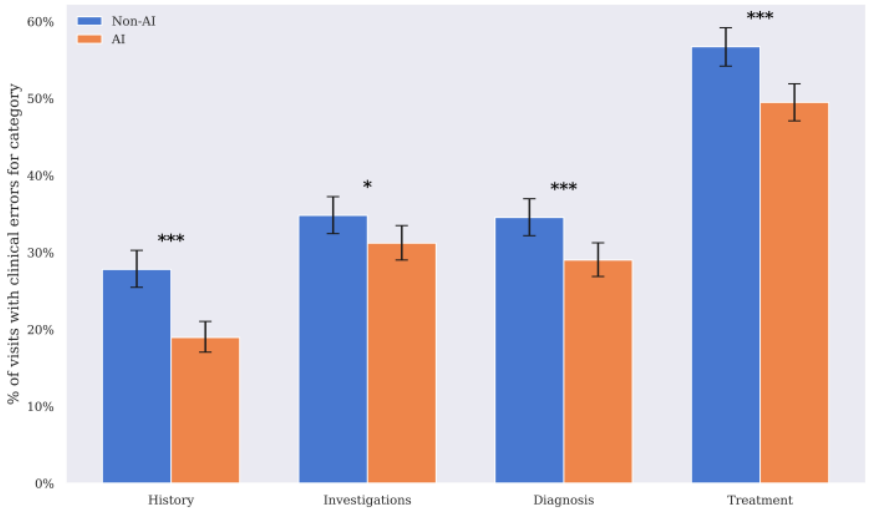


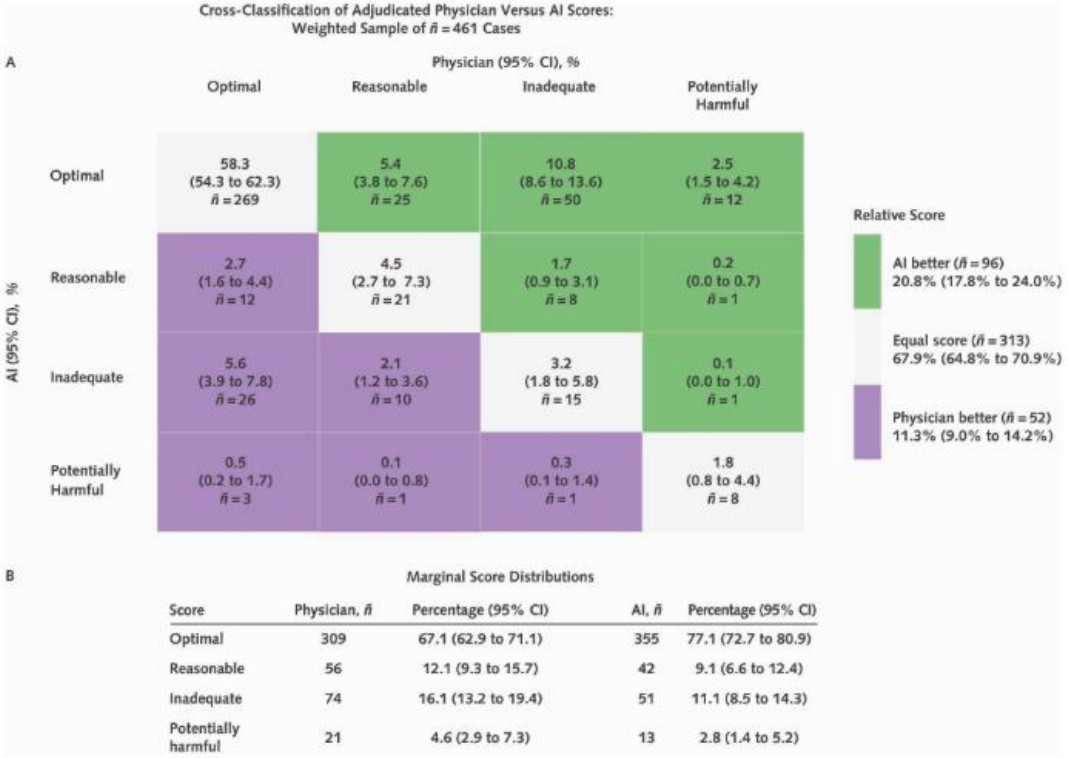
Figure 4: Clinical error rates for history-taking, investigations, diagnosis, and treatment, comparing the AI group to the non-AI group. Error bars show 95% Wilson confidence intervals. * indicates $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Main period, all visits	Induction period	Main period, only visits with reds
History	31.8% (21.9%-40.5%)	16.7% (4.5%-27.3%)	30.8% (12.8%-45.1%)
Investigations	10.3% (1.0%-18.8%)	13.8% (2.7%-23.6%)	17.9% (-72.6%-60.9%)
Diagnosis	16.0% (6.9%-24.2%)	6.4% (-5.6%-17.1%)	31.5% (14.0%-45.5%)
Treatment	12.7% (6.8%-18.3%)	4.3% (-3.0%-11.1%)	18.0% (9.4%-25.9%)

Table 4: Relative risk reduction in clinical errors across each category for all visits during the main study period, all visits during the induction period, and only visits with reds for the relevant category during the main study period.

They are seeing patients **right now** in the real world

- Cedars-Sinai deployment of K-Health platform (telehealth urgent care)
- High concordance between physician and AI (56.8%); AI management decisions more frequently rated as “optimal” (77% versus 67%)



Cross-Classification of Adjudicated Physician Versus AI Scores: Weighted Sample of $\tilde{n} = 461$ Cases

A

AI (95% CI), % \ Physician (95% CI), %	Optimal	Reasonable	Inadequate	Potentially Harmful
Optimal	58.3 (54.3 to 62.3) $\tilde{n} = 269$	5.4 (3.8 to 7.6) $\tilde{n} = 25$	10.8 (8.6 to 13.6) $\tilde{n} = 50$	2.5 (1.5 to 4.2) $\tilde{n} = 12$
Reasonable	2.7 (1.6 to 4.4) $\tilde{n} = 12$	4.5 (2.7 to 7.3) $\tilde{n} = 21$	1.7 (0.9 to 3.1) $\tilde{n} = 8$	0.2 (0.0 to 0.7) $\tilde{n} = 1$
Inadequate	5.6 (3.9 to 7.8) $\tilde{n} = 26$	2.1 (1.2 to 3.6) $\tilde{n} = 10$	3.2 (1.8 to 5.8) $\tilde{n} = 15$	0.1 (0.0 to 1.0) $\tilde{n} = 1$
Potentially Harmful	0.5 (0.2 to 1.7) $\tilde{n} = 3$	0.1 (0.0 to 0.8) $\tilde{n} = 1$	0.3 (0.1 to 1.4) $\tilde{n} = 1$	1.8 (0.8 to 4.4) $\tilde{n} = 8$

Relative Score

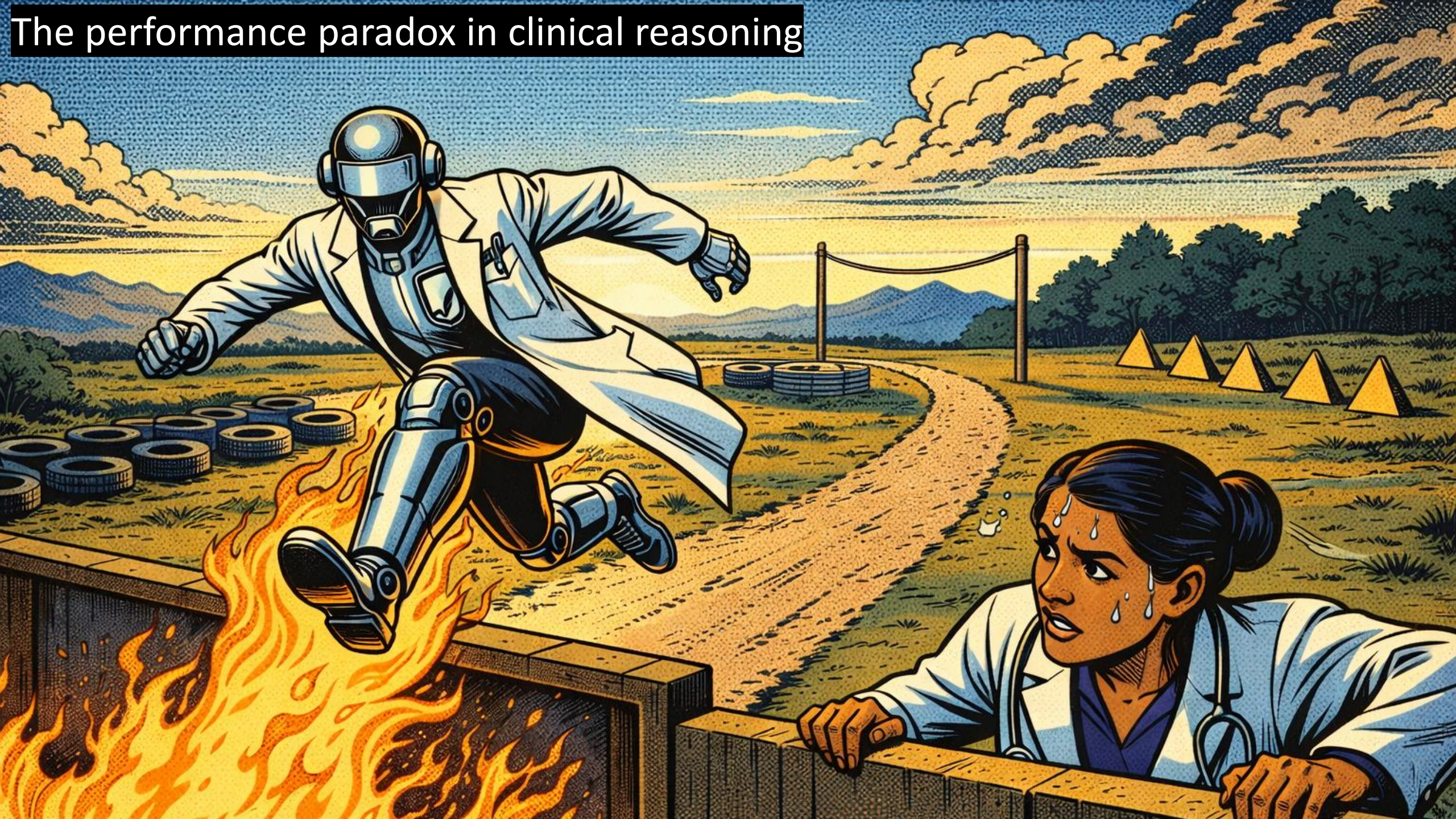
- AI better ($\tilde{n} = 96$) 20.8% (17.8% to 24.0%)
- Equal score ($\tilde{n} = 313$) 67.9% (64.8% to 70.9%)
- Physician better ($\tilde{n} = 52$) 11.3% (9.0% to 14.2%)

B

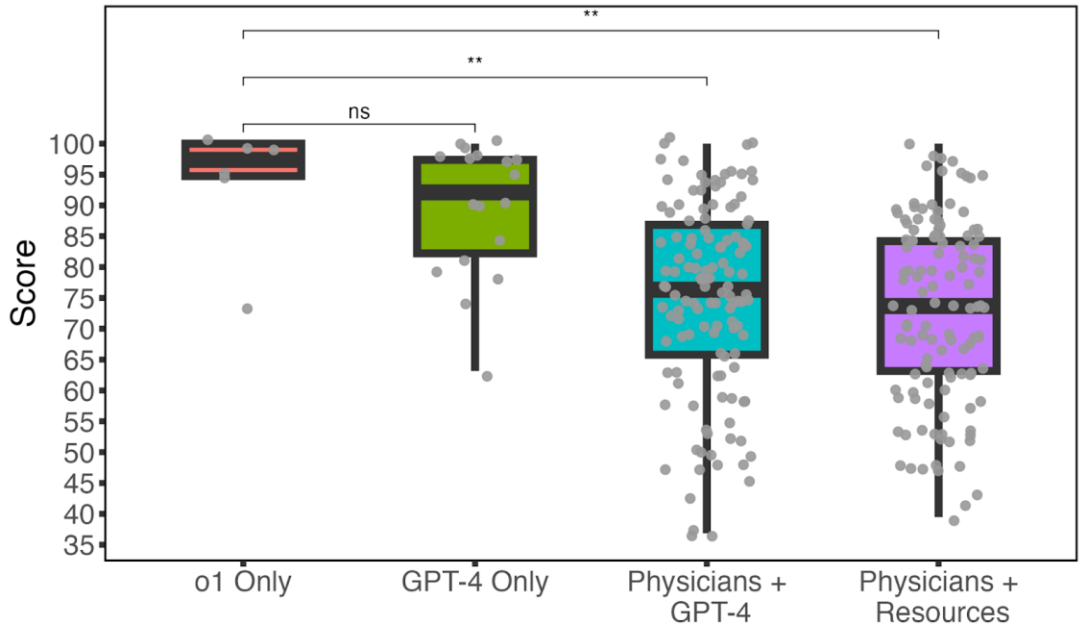
Marginal Score Distributions

Score	Physician, $\tilde{n}$	Percentage (95% CI)	AI, $\tilde{n}$	Percentage (95% CI)
Optimal	309	67.1 (62.9 to 71.1)	355	77.1 (72.7 to 80.9)
Reasonable	56	12.1 (9.3 to 15.7)	42	9.1 (6.6 to 12.4)
Inadequate	74	16.1 (13.2 to 19.4)	51	11.1 (8.5 to 14.3)
Potentially harmful	21	4.6 (2.9 to 7.3)	13	2.8 (1.4 to 5.2)

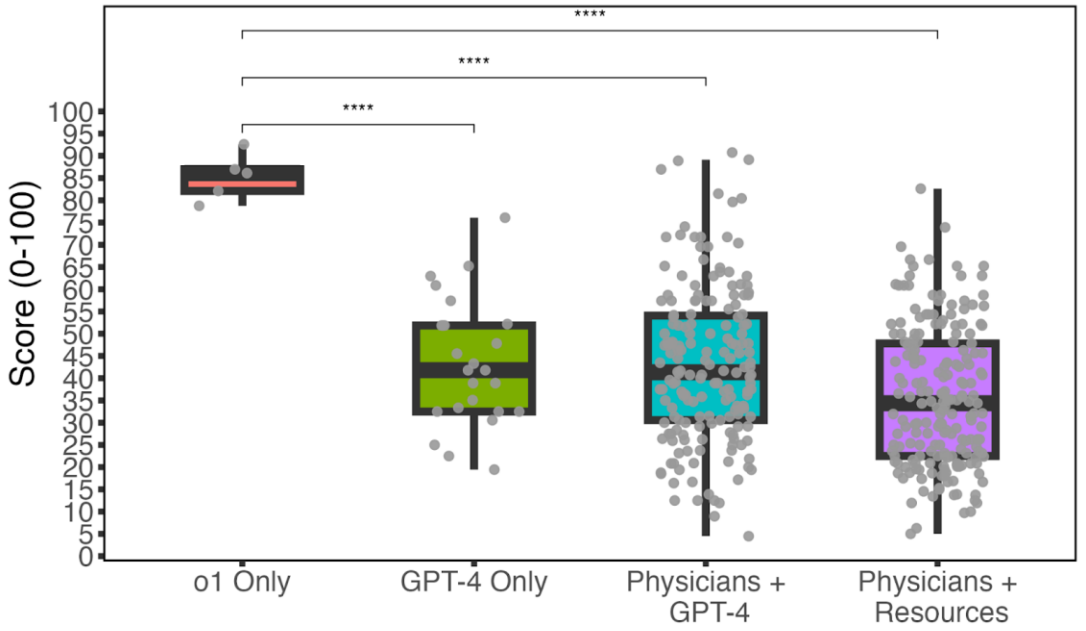
The performance paradox in clinical reasoning



Language models might **already** be better than us – and certainly will soon…

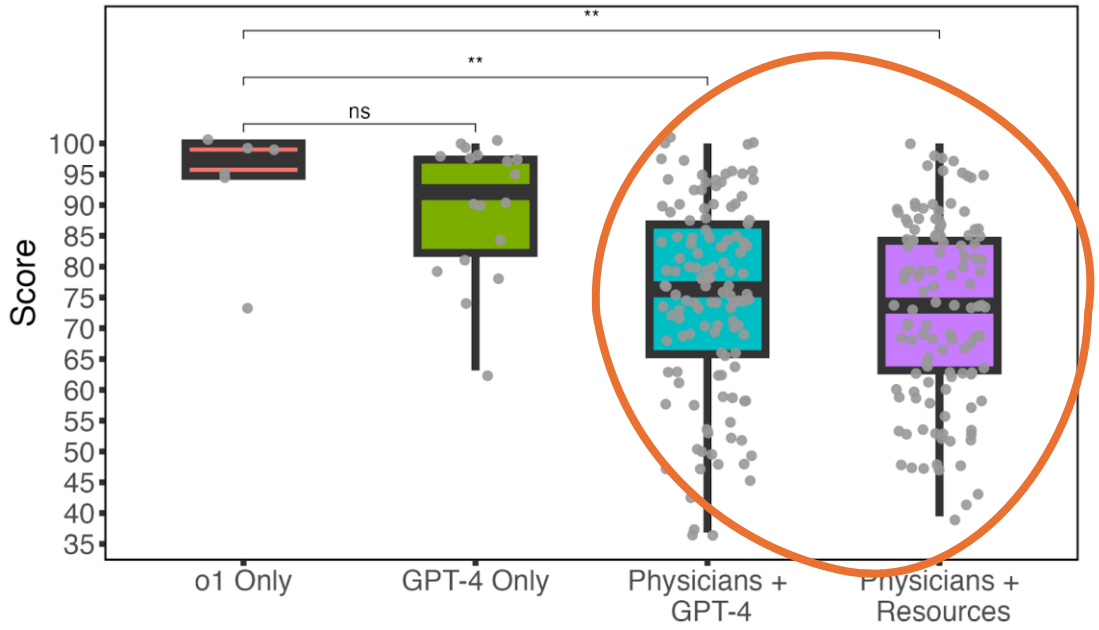


A. o1 Diagnostic Reasoning Scores Compared to GPT-4 and Physicians

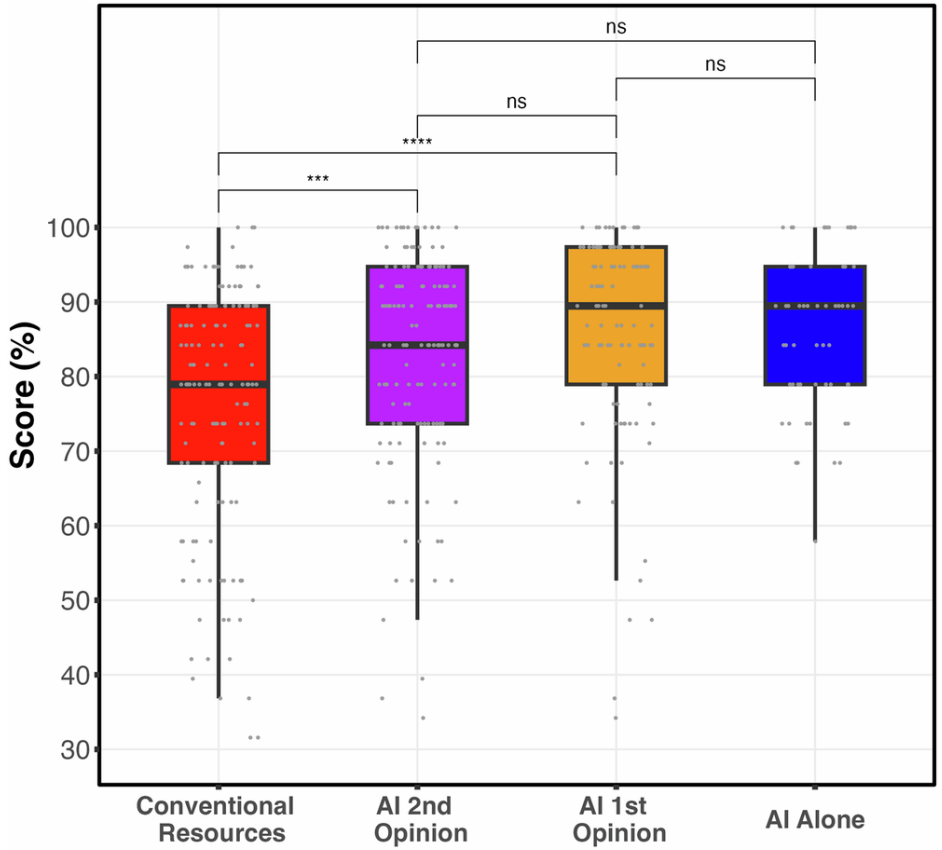


B. o1 Management Reasoning Scores Compared to GPT-4 and Physicians

…but they don’t always make **us** better.



A. o1 Diagnostic Reasoning Scores Compared to GPT-4 and Physicians



LLMs can collect histories

- Double-blind trial using standardized patients of AMIE (Articulate Medical Intelligence Explorer)
- Using standardized rubrics (PACES), performed better than humans in 28 of 32 axes, which significantly improved diagnostic accuracy

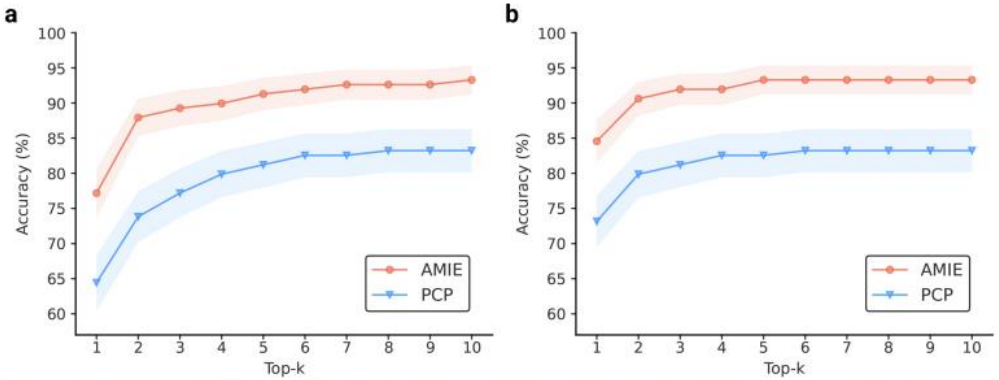


Figure 3 | Specialist-rated top-k diagnostic accuracy. AMIE and PCPs top-k DDx accuracy are compared across 149 scenarios with respect to the ground truth diagnosis (**a**) and all diagnoses in the accepted differential (**b**). Bootstrapping (n=10,000) confirms all top-k differences between AMIE and PCP DDx accuracy are significant with $p < 0.05$ after FDR correction.

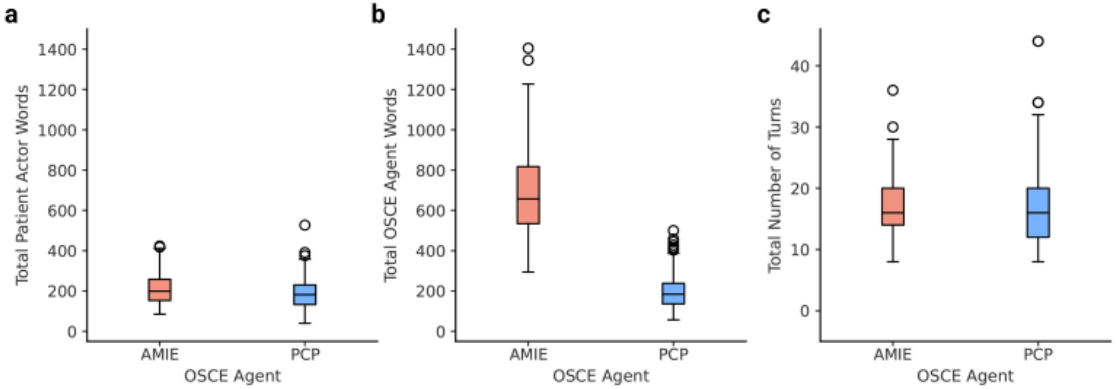
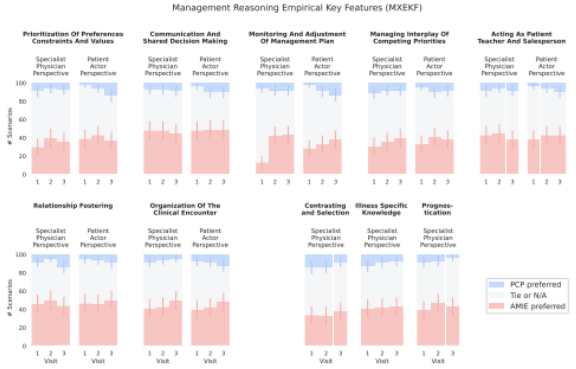
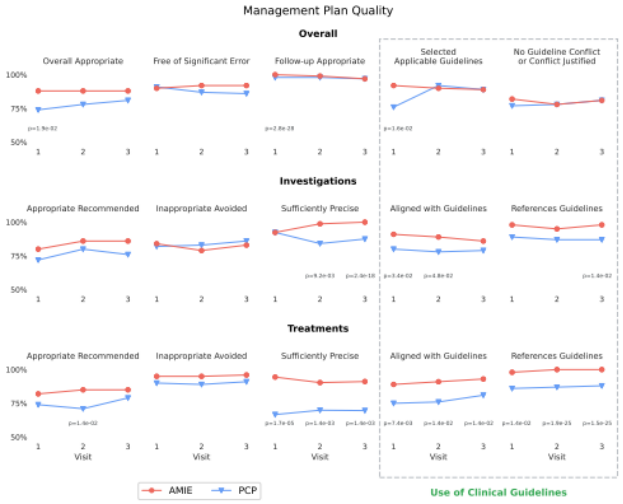


Figure A.11 | Distribution of words and turns in OSCE consultations. (**a**) Total patient actor words elicited by AMIE vs. PCPs. (**b**) Total words sent to patient actor from AMIE vs. PCPs. (**c**) Total number of turns in AMIE vs. PCP consultations.

… and make management decisions.

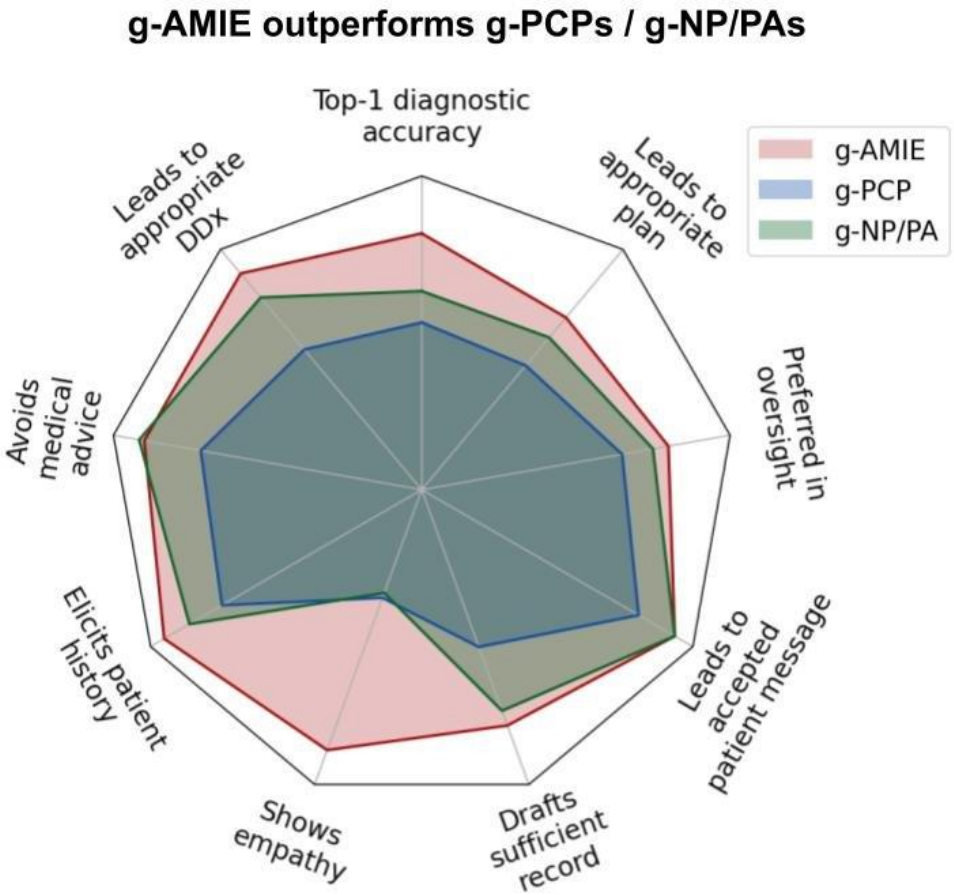
- Double blind follow up study of multi-agent AMIE trained with a “reasoning” Mx Agent that can read clinical guidelines.
- Noninferior or superior to PCPs in every patient actor scenario, on every measure.



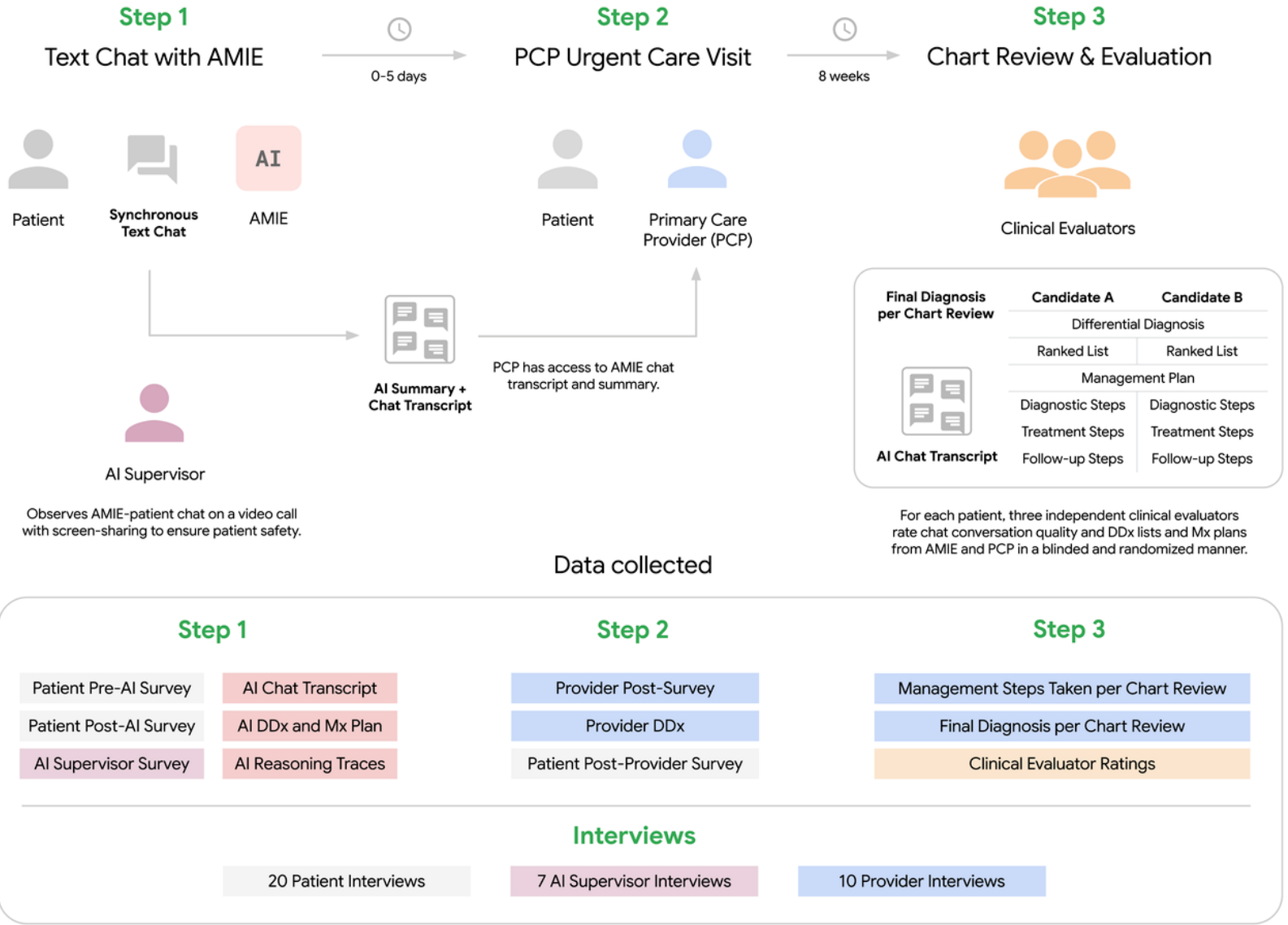


LLMs can participate in (human-like) supervisory systems…

- Double-blind study of AMIE, primary care residents, and NP/Pas performing intake histories and then handing off to experienced PCPs
- AMIE drastically outperforms residents, and generally.

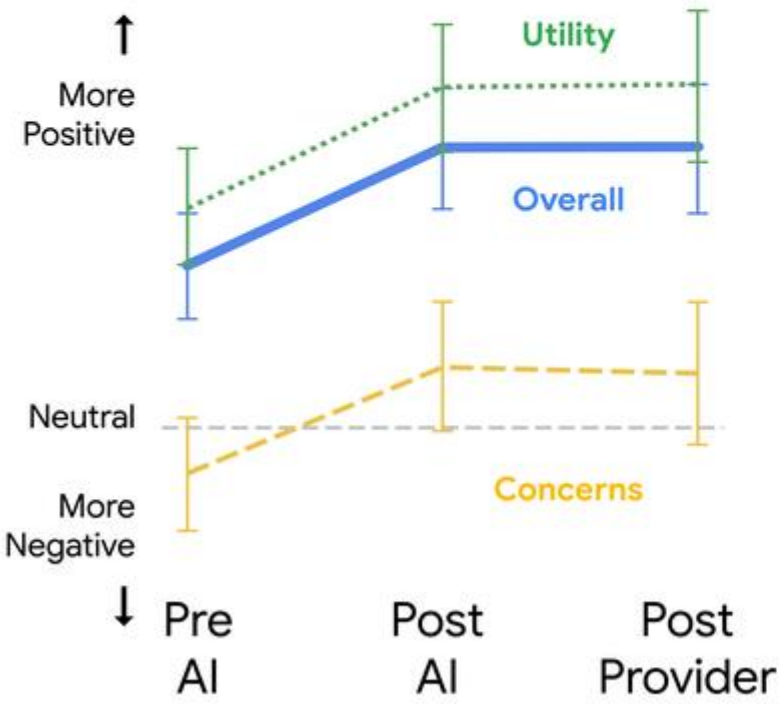


… and they **work** in the real world with real patients.

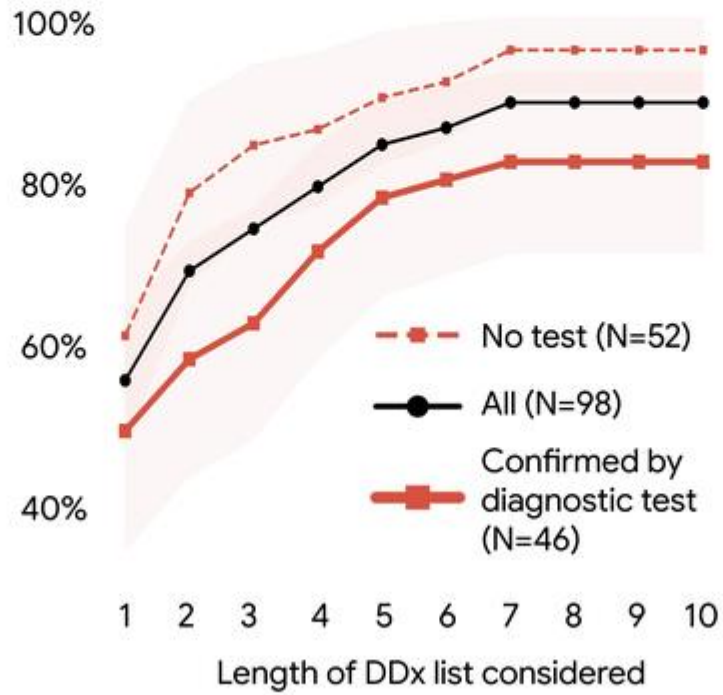


Zero safety stops required
across all patient-AMIE interactions

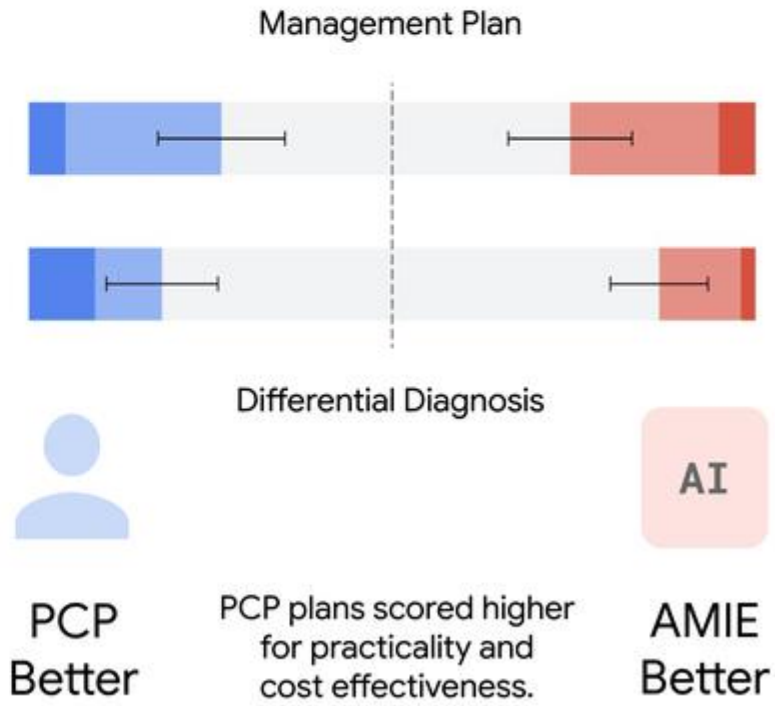
Patients' **trust** in AI increased after AMIE



AI DDx **accuracy** high w.r.t. objective reference standard

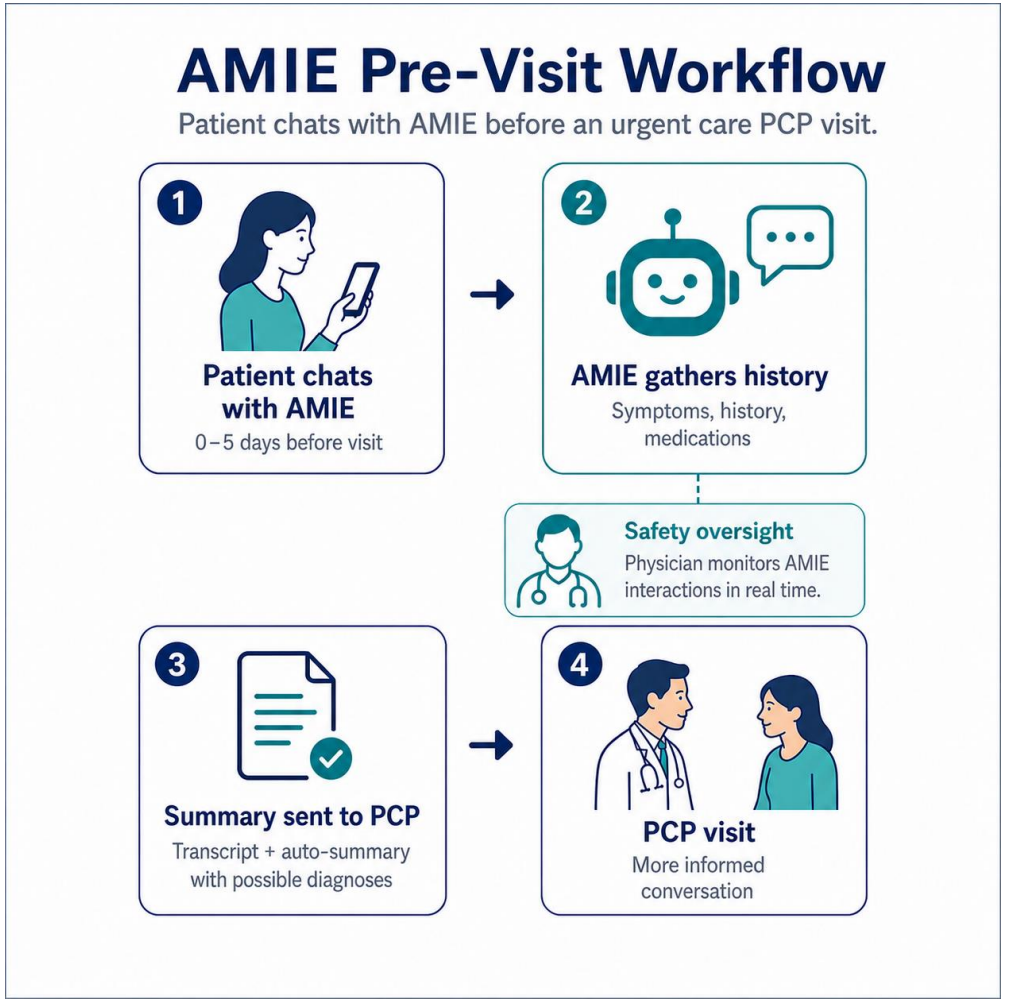


AMIE on par with PCPs for overall **Mx plan and DDx quality**



Discussion & Interpretation

We can study **human-computer interaction** in real-life clinical workflows with meaningful outcomes.



Discussion & Interpretation

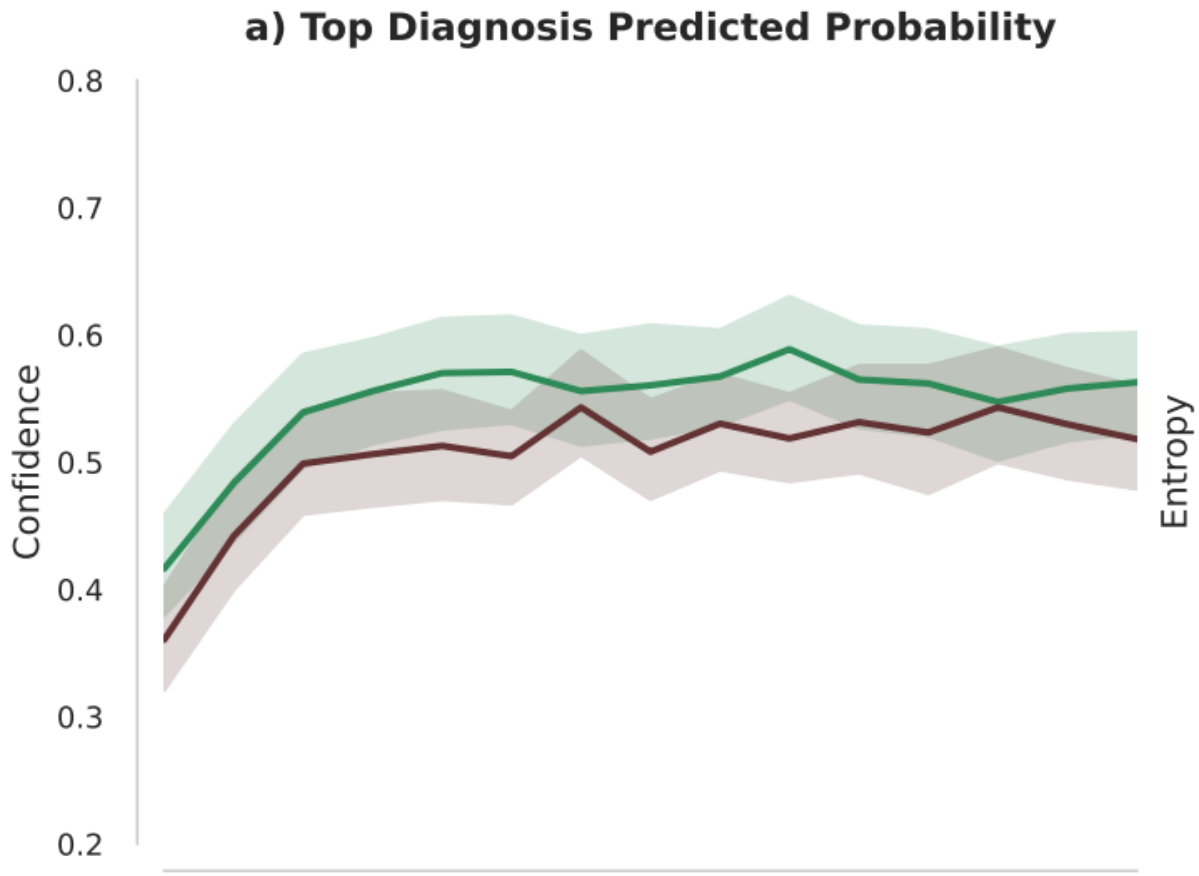
Providing safety infrastructure is the **minimum** we need to do for clinical implementations – but different safety infrastructures come with very different trade offs.

Close collaboration between the **IRB**, **IT**, **quality and safety**, and investigators allows for nimble and flexible AI studies, even in experimental workflows.

Discussion & Interpretation

LLMs can **meaningfully** collect data from patients and make diagnoses.

They also display some of the same challenges as human clinicians – which may be a topic for HCI research.



Discussion & Interpretation

Management reasoning is **measurable**, even if those measures are imperfect.

Constructs adapted from the psychological literature can be adapted to tools that impact reasoning.

Discussion & Interpretation

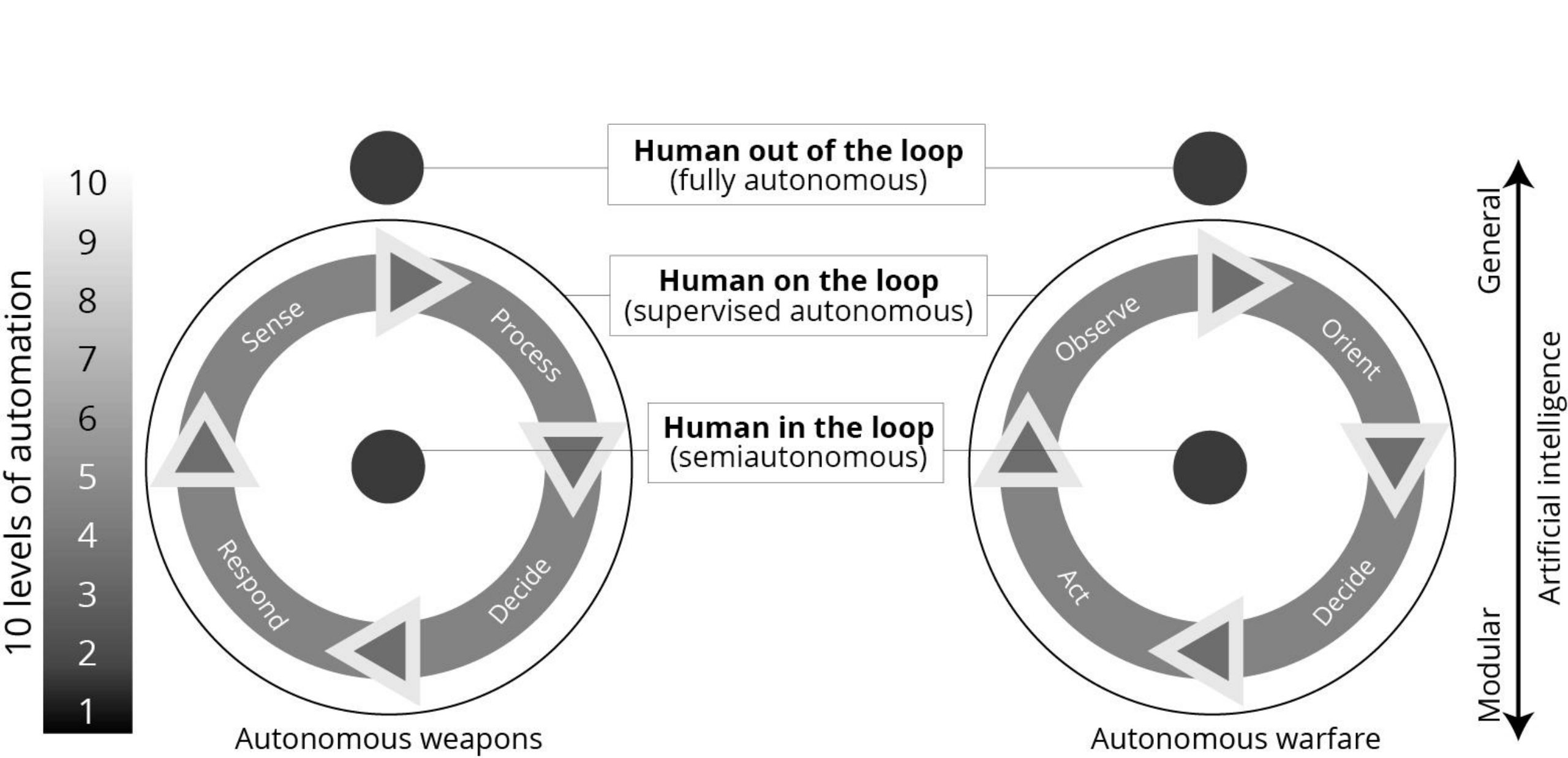
It is fully possible to **match** the degree of clinical claims with meaningful evidence.

AI co-clinician

Case 1: Abdominal pain

This video is for research purposes only and does not involve real patients. It is being shared to demonstrate the capabilities and limitations of the technology today. Our initial research collaborations do not involve the depicted capabilities, which are not intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease, or to provide medical advice.


FINISH
Can we solve the performance paradox by movin
beyond “human in the loop”?


10 levels of automation
10
9
8
7
6
5
4
3
2
1
Human out of the loop
(fully autonomous)
Human on the loop
(supervised autonomous)
Human in the loop
(semiautonomous)
Sense
Process
Respond
Decide
Observe
Orient
Act
Decide
Autonomous weapons
Autonomous warfare
General
Artificial intelligence
Modular

Moving from HITL to HOTL

- Current paradigm in **human-in-the-loop** for medicine… but this is probably not sufficient
- 1) Humans as bottleneck – doesn’t scale
- 2) Automation bias
- 3) Cognitive deskilling can degrade HITL systems

Cognitive deskilling is **already** here in medicine

- Retrospective observational study of ACCEPT RCT
- Over three-month period, AI-assisted colonoscopy tools were randomly assigned to endoscopists
- Adenoma detection rate dropped from 28.4% to 22.4%

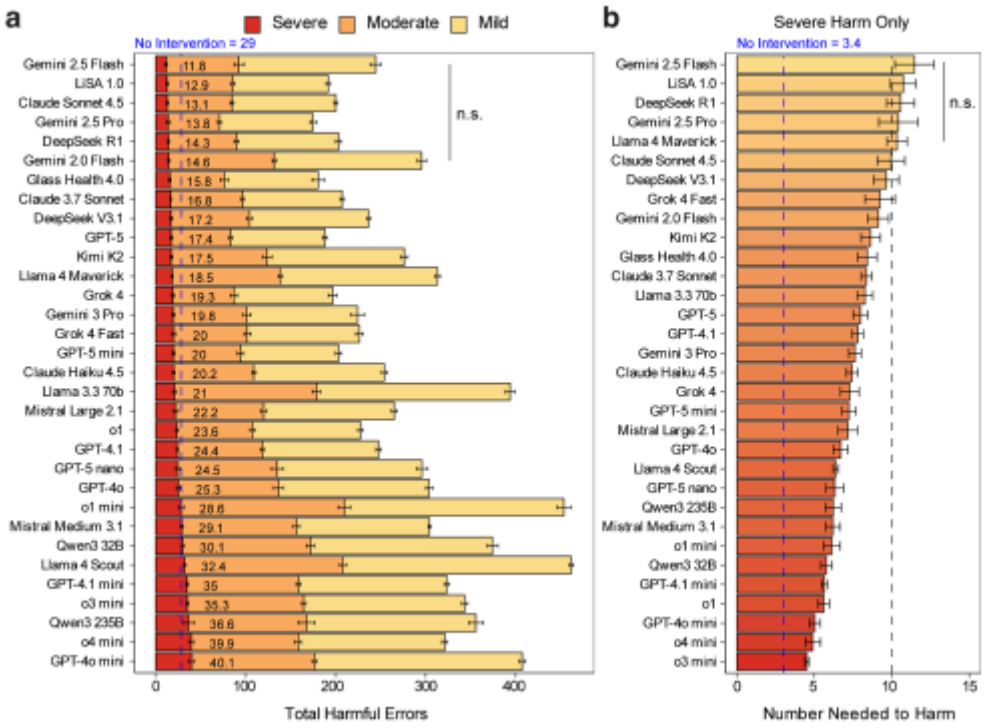
Budzyń K, Romańczyk M, Kitala D, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol Hepatol*. Published online August 12, 2025. doi:10.1016/S2468-1253(25)00133-5

Major challenges for 2026 and beyond

- Significant computational work for **long context** for health records, compounded with low overall data quality.
- Diagnosis is (theoretically) solved – but **management reasoning** remains a major challenge.

Lower performance on **management reasoning –** but still on average better than humans

- New preprint from my lab, looking at **primary care e-consults** (eg, very complex cases).
- Major findings – RAG leads to safer decisions; top performing models make **errors of omission** while lower performers make both.
- **Human decisions** are in the middle of the pack



Wu D, Haredasht FN, Maharaj SK, et al. First, do NOHARM: towards clinically safe large language models. arXiv [csCY]. Published online 2025. http://arxiv.org/abs/2512.01241

Major challenges for 2026 and beyond

- Significant computational work for **long context** for health records, compounded with low overall data quality.
- Diagnosis is (theoretically) solved – but **management reasoning** remains a major challenge.
- There are no regulatory frameworks for HITL->HOTL transitions
- Technology behind scalable oversight systems remains unproven.
- Will there be trials showing patient outcomes?

Triadic care is a **research agenda**, not a product

- Waymo model – important to start in low-risk situations with “safety drivers”
- **Real world evaluation** is fundamentally necessary – needs to be tuning between physician, patient, and AI system
- Risks are **false positives** (alert fatigue), **false negatives** (loss of trust)


The End … (of this presentation) …
… to be continued …

MOC REFLECTIVE STATEMENT (BRIEF TAKE HOME NOTES FOR REFERENCE)

Many of our assumptions about human-computer interaction do not hold in experimental studies, largely because of the psychological characteristics of human clinicians. While this has largely not mattered in previous generations of artificial intelligence, the power of modern transformer-based algorithms and their rapid proliferation into healthcare delivery means there is a pressing need for better understanding human-computer interaction in healthcare. Future care models will likely reflect different interaction systems, with important implications for both patients and clinicians.

REFERENCES

1. Brodeur PG, Buckley TA, Kanjee Z, et al. Performance of a large language model on the reasoning tasks of a physician. *Science.* 2026. doi:10.1126/science.adz4433 Science
2. Brodeur PG, et al. A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic. *arXiv.* Preprint posted March 9, 2026. arXiv:2603.08448 arXiv
3. Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic artificial intelligence. *Nature.* 2025. doi:10.1038/s41586-025-08866-7 Nature
4. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open.* 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969 Science
5. Korom R, Kiptinness S, Adan N, et al. AI-based clinical decision support for primary care: a real-world study. *arXiv.* Preprint posted July 22, 2025. arXiv:2507.16947 arXiv arXiv