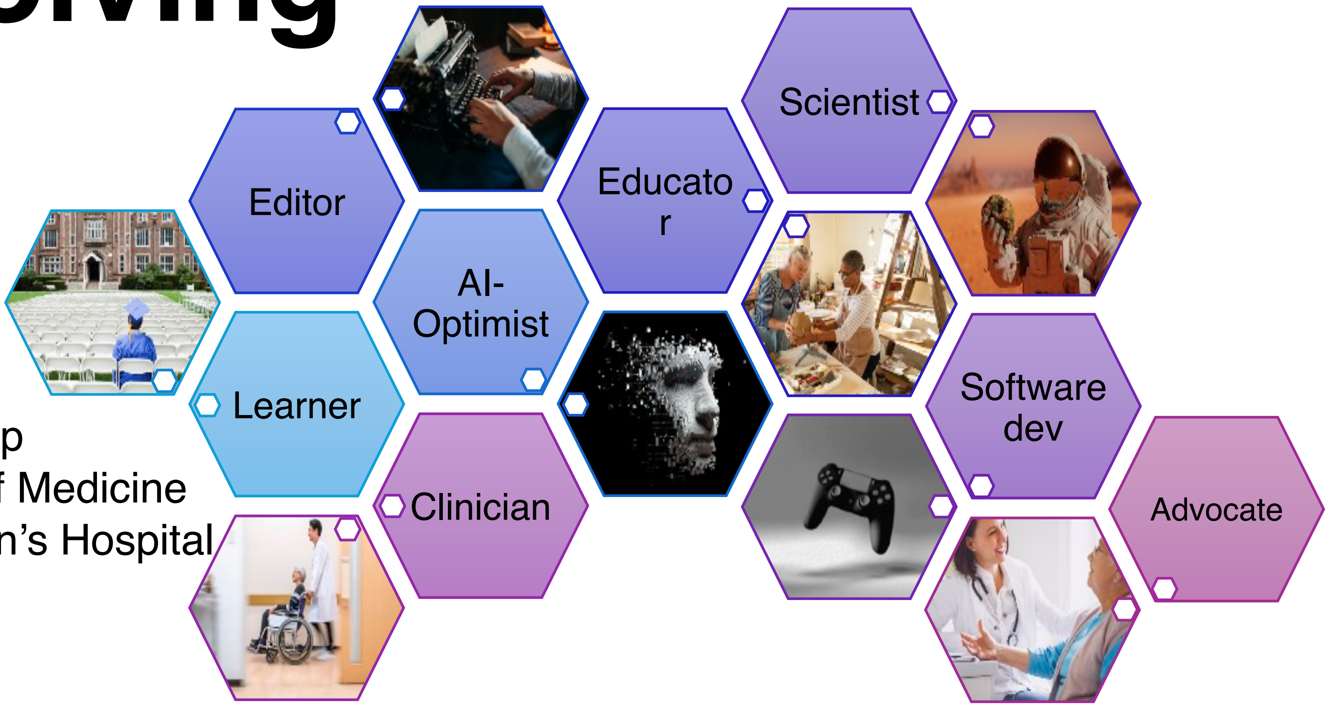
The Natural Language of Clinical Problem-Solving

Raja-Elie E. Abdulnour, M.D.

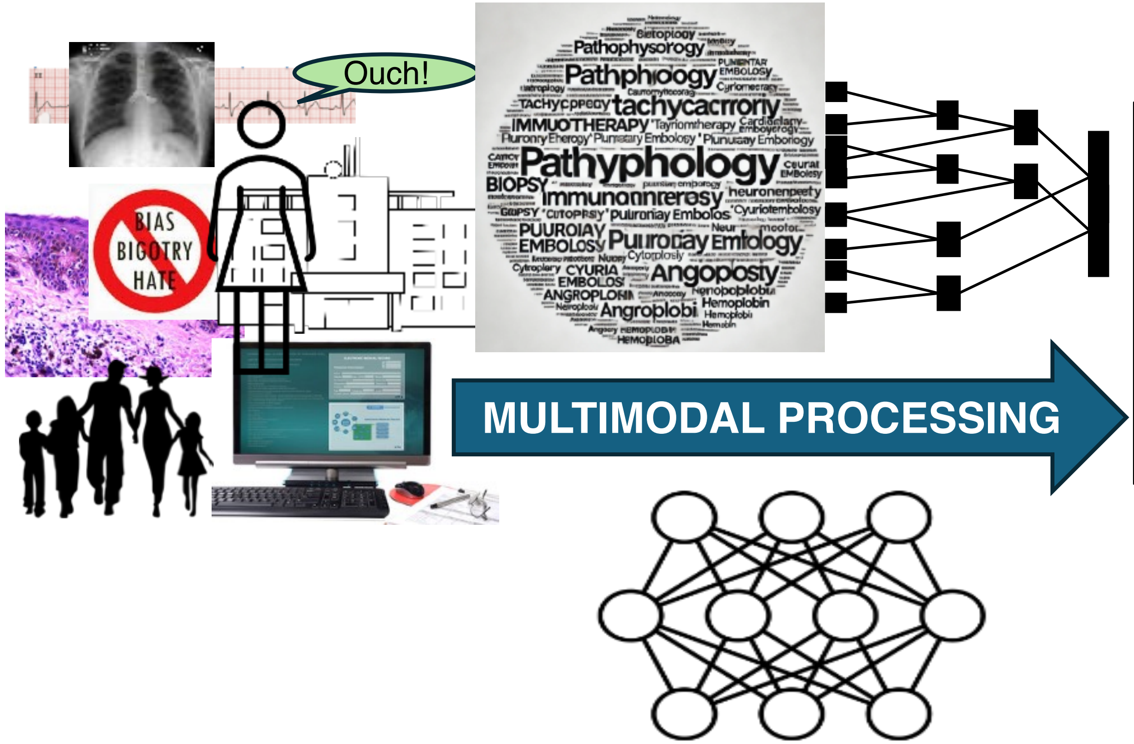
Editor-in-Chief, NEJM Journal Watch
Editor, Clinical Development and AI Innovation, NEJM Group
’23-’24 Diagnostic Excellence Scholar, National Academy of Medicine
Pulmonary and Critical Care Medicine, Brigham and Women’s Hospital
Assistant Professor of Medicine, Harvard Medical School



Disclosures

- MMS / NEJM Group
  - NEJM Healer
  - NEJM Healer was acquired by Lecturio (no financial COI)
- Lumeris, consultant
- I am NOT a computer scientist

Clinical Reasoning: How doctors think



Inductive reasoning

Problem representation

67 year old main with acute dyspnea, chest pain, fever, and hypoxemia. Recent sick contact. Normal chest Xray.

THINK FAST!!!

Deductive reasoning

What is causing problem?

Differential Diagnosis

I have a high suspicion for Covid pneumonia and possible bacterial superinfection

Think sloooow….

TREATMENT

Solving

Corticosteroids
Remdesivir
Levofloxacin

Clinical reasoning. INTELLIGENCE = measure of reasoning

"To Err Is Human"

Cardiac arrest from PE

Inductive reasoning
Problem representation
67 year old main with acute dyspnea, chest pain, fever, and hypoxemia. Recent sick

Deductive reasoning
What is causing problem?
Differential Diagnosis
I have a high suspicion for Covid pneumonia

TREATMENT Solving
Corticosteroids
Remdesivir
Levofloxacin

795,000 / year permanently disabled or die.

3rd cause of death

The Burden

JAMA Internal Medicine | Original Investigation

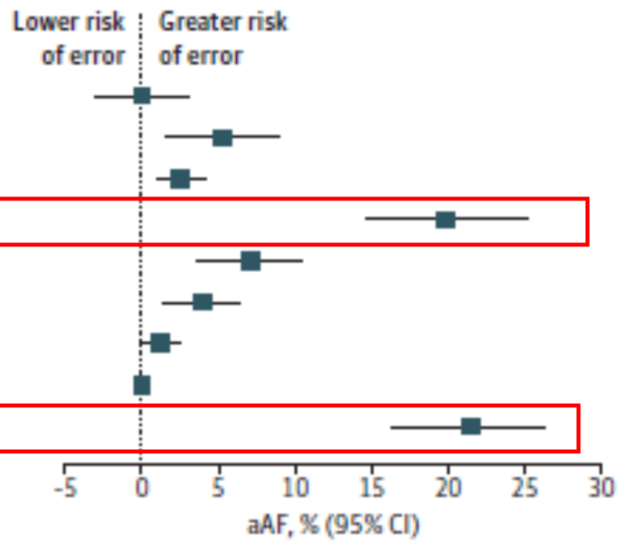
Diagnostic Errors in Hospitalized Adults Who Died or Were Transferred to Intensive Care

Andrew D. Auerbach, MD, MPH; Tiffany M. Lee, BA; Colin C. Hubbard, PhD; Sumant R. Ranji, MD; Katie Raffel, MD; Gilmer Valdes, PhD, DABR; John Boscardin, PhD; Anuj K. Dalal, MD; Alyssa Harris, MPH; Ellen Flynn, RN, MBA, JD; Jeffrey L. Schnipper, MD, MPH; for the UPSIDE Research Group

- 550 (23.0%)out of 2428 experienced diagnostic error
- 22.7% of inpatient patients who died or transferred to the ICU experienced diagnostic error

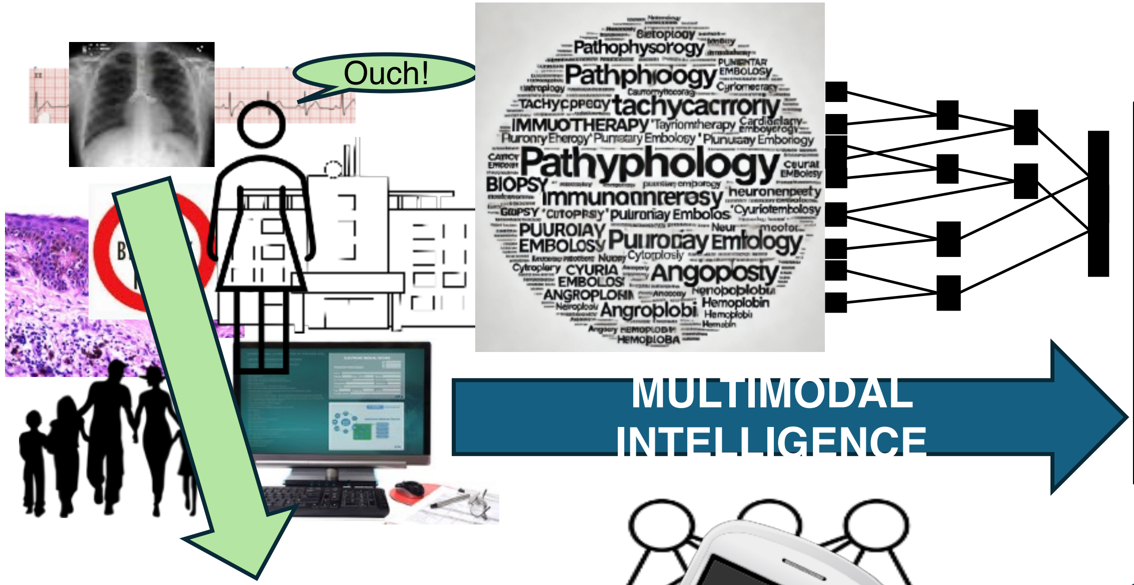
Figure 2. DEER Process Fault Dimensions: Prevalence, Adjusted Associations With Diagnostic Errors, and Adjusted Attributable Fractions (aAFs) (N = 2428)

Diagnostic process dimensions	Patients, No. (%) No error	Error	aRR (95% CI)	aAF, % (95% CI)
Access/presentation	232 (14.1)	36 (8.8)	0.79 (0.60 to 1.05)	0.0 (−3.1 to 3.0)
History taking	80 (5.0)	54 (12.9)	1.71 (1.23 to 2.36)	5.2 (1.6 to 8.9)
Physical examination	19 (0.9)	45 (7.8)	1.86 (1.39 to 2.49)	2.5 (0.9 to 4.1)
Testing	60 (4.4)	158 (34.1)	2.85 (2.16 to 3.76)	19.9 (14.7 to 25.1)
Patient follow-up and monitoring	47 (3.0)	81 (15.4)	1.94 (1.45 to 2.60)	7.0 (3.7 to 10.4)
Obtaining referrals	42 (1.9)	58 (10.0)	1.54 (1.13 to 2.09)	3.8 (1.3 to 6.3)
Teamwork	4 (0.1)	11 (2.3)	2.89 (1.62 to 5.16)	1.2 (−0.1 to 2.5)
Communication with patient/caregiver	5 (0.3)	2 (0.3)	1.47 (0.66 to 3.32)	0.0 (−0.2 to 0.2)
Assessment	87 (3.7)	220 (37.1)	2.89 (2.23 to 3.73)	21.4 (16.4 to 26.4)



Auerbach. JAMA IM. 2024

Artificial Clinical Intelligence



Inductive reasoning
Problem representation
67 year old main with acute dyspnea, chest pain, fever, and hypoxemia. Recent sick contact. Normal chest Xray.

Deductive reasoning
What is causing problem?
Differential Diagnosis
I have a high suspicion for Covid pneumonia and possible bacterial superinfection

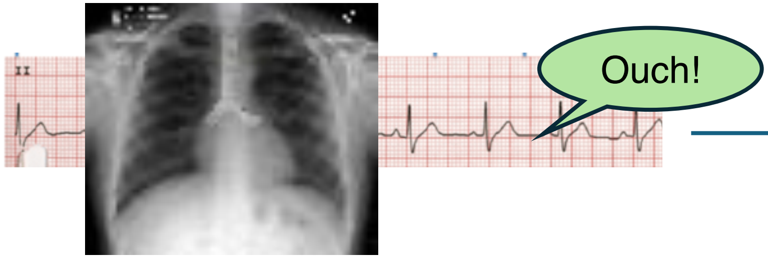
TREATMENT
Solving
Corticosteroids
Remdesivir
Levofloxacin

Multimodal GenAI
Scribe, imaging, text

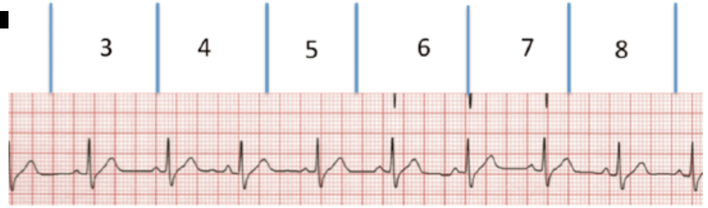
Hey Doc! Patients with Covid19 are at risk of PE. This patient's hypoxemia is out of proportion to their CXR findings. Consider a CTPE

PE diagnosed
IV heparin

Generative AI: A paradigm shift

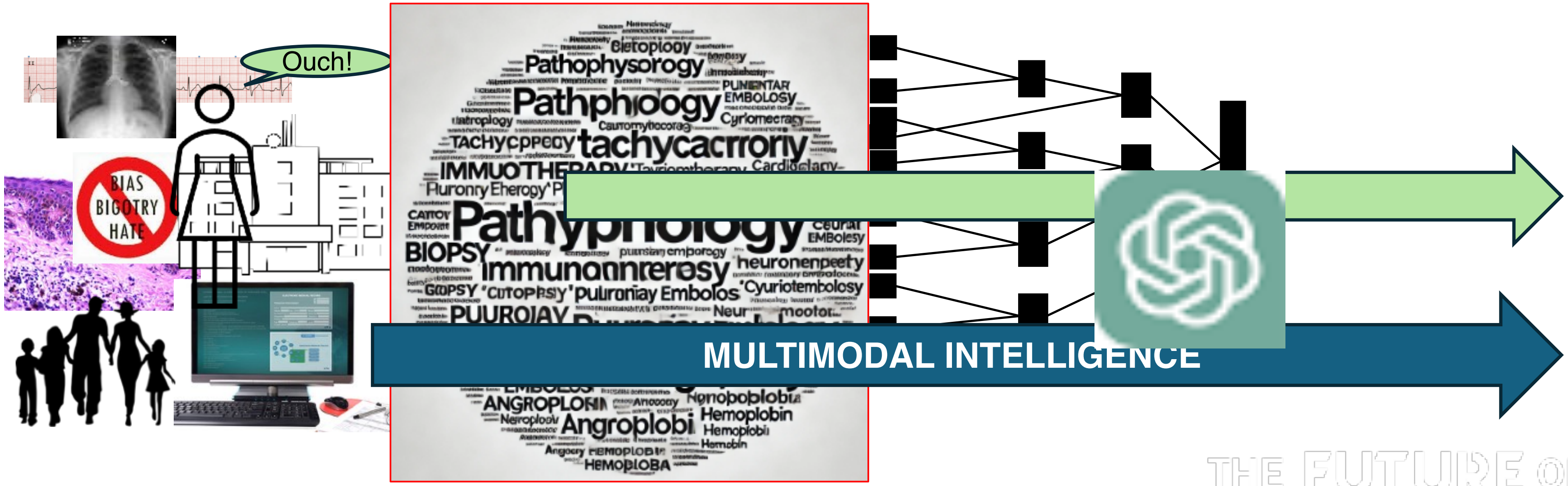


Unimodal Intelligence



14 R waves in 10 secs

14 x 6 = 84 BPM



Clinical assessment and management plan

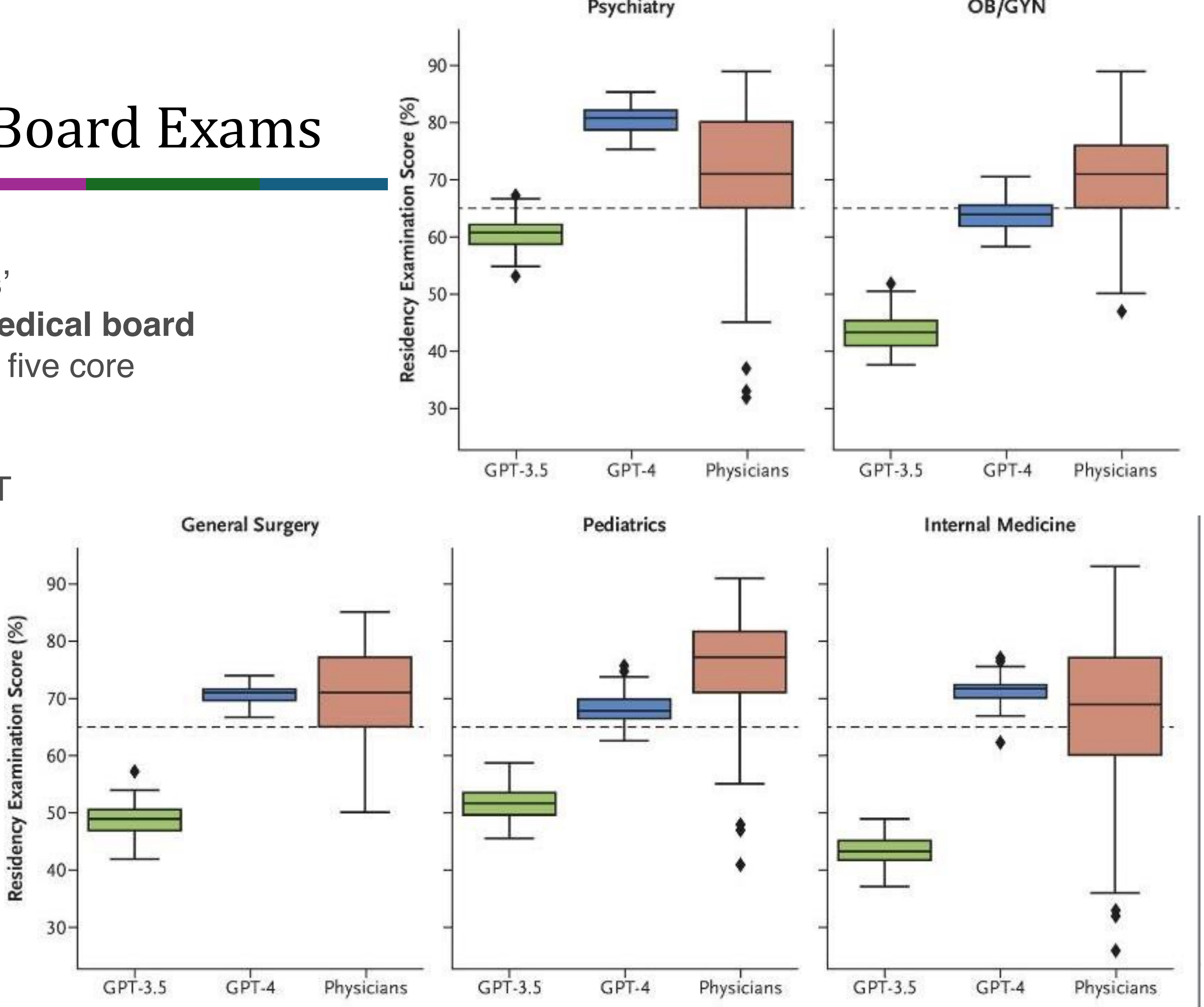
GPT vs Physicians on Board Exams

retrospective analysis of physicians' performance on the **2022 Israeli medical board certification examinations** across five core medical specialties.

Compared 849 physicians with GPT

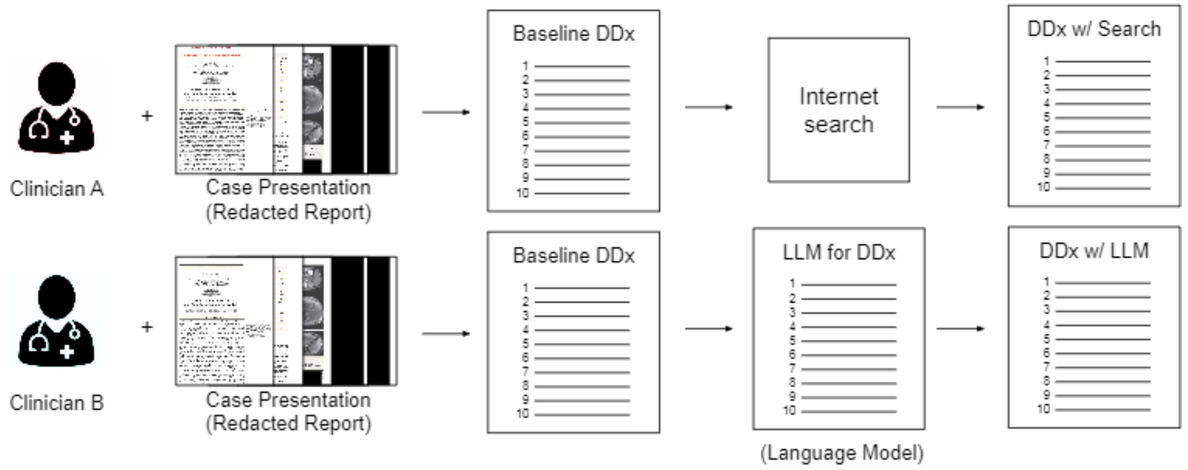
Accounted for model stochasticity
by GPT model on 120 attempts

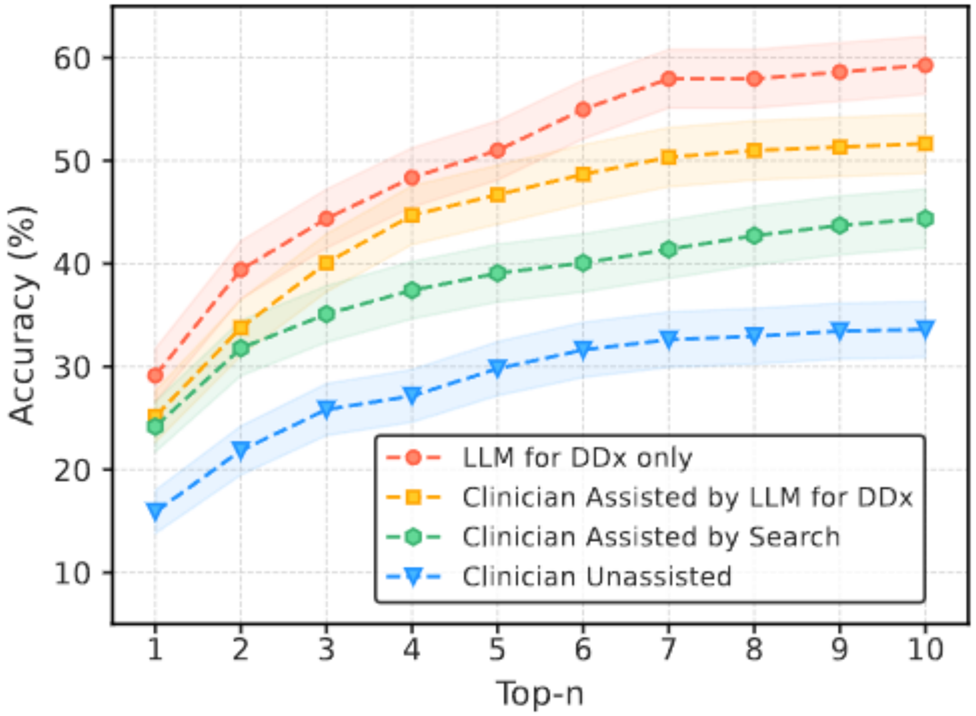
Katz et al. NEJM AI, 2024



LLMs Can Solve Case Reports

- 358 NEJM CPCs, including 56 not included in training data, using a fine-tuned Palm2 compared to human clinicians.





McDuff D, Schaekermann M et al, Towards Accurate Differential Diagnosis with Large Language Models. Preprint available at https://arxiv.org/pdf/2312.00164.pdf. **Not yet peer reviewed.**

LLMs Express Clinical Reasoning

- Residents, attending, and GPT-4 solving NEJM Healer cases – 236 sections in total
- Assessed expression of reasoning process with r-IDEA

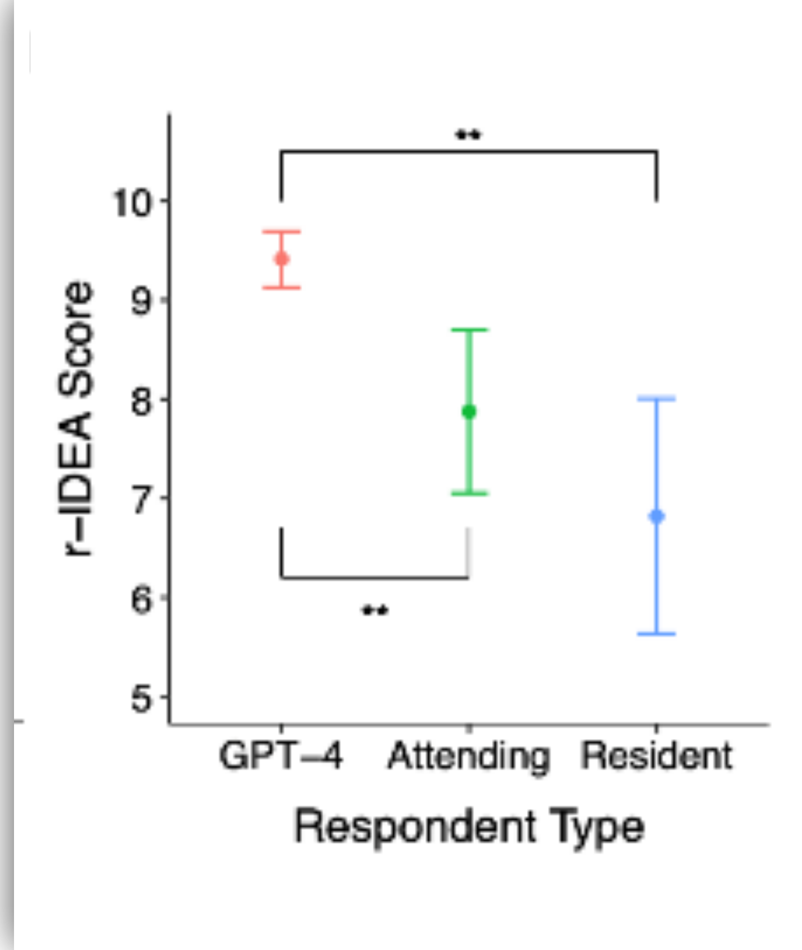
Domain	Description	Assessment	Points
I - Interpretive Summary	Provides a concise summary statement that uses semantic vocabulary to highlight the most important elements from history, exam, and testing and to interpret and represent the patient's main problem(s). The presence or absence of the following features is assessed: a) Key risk factors; b) Chief complaint; c) Illness time course; and d) Use of semantic qualifiers (e.g. monoarticular vs polyarticular) or unified medical concepts (e.g. volume overload, cardiovascular risk factors). NB: Some problems have an implied time course (e.g. syncope, seizure).	No features present	0
		1 feature present	1
		2 features present	2
		3 features present	3
		4 features present	4
D – Differential Diagnosis	Offers more than one relevant diagnostic possibility, committing to what is most likely and considering what is less likely or unlikely yet important to consider for the main chief complaint. If the chief complaint is a diagnosis or syndrome (e.g. acute on chronic systolic heart failure) then differential to rate may be around the differential for that exacerbation (e.g. medication non-compliance vs. arrhythmia).	No differential	0
		Differential is implicitly stated, given as a diagnostic category (e.g. "cardiac"), OR implicitly prioritized	1
		Differential is explicitly stated AND explicitly prioritized	2
E – Explanation of Lead Diagnosis	Explains the reasoning behind the lead diagnosis, including the epidemiology and key features and how these compare with the patient's presentation. If objective data points are not clearly linked to the lead diagnosis or alternative diagnosis, then only designate points to lead OR alternative diagnosis and NOT both.	No explanation	0
		1 objective data point in explanation of lead diagnosis	1
		≥2 objective data points in explanation of lead diagnosis	2
A – Alternative Diagnosis Explained	Explains the reasoning behind alternative diagnoses, including the epidemiology and key features and how these compare with the patient's presentation and alternative diagnosis. If objective data points are not clearly linked to the lead diagnosis or alternative diagnosis, then only designate points to lead OR alternative diagnosis and NOT both.	No explanation for any alternative diagnosis	0
		1 objective data point in explanation of at least one alternative diagnosis	1
		≥2 objective data points in explanation of at least one alternative diagnosis	2
Revised-IDEA Score	Overall evaluation of demonstration of clinical reasoning in the assessment section of admission notes.	Sum of I + D + E + A points (score ≥6 indicates high-quality clinical reasoning documentation)	0-10

Figure 1 The Revised-IDEA assessment tool for clinical reasoning documentation.

Cabral S, […], Abdulnour RE, Rodman A. JAMA IM, April 1, 2024

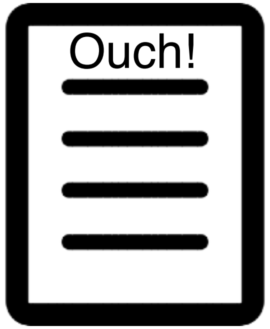
LLMs Express Clinical Reasoning

- Residents, attending, and GPT-4 solving NEJM Healer cases – 236 sections in total
- Assessed expression of reasoning process with r-IDEA
- 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



Cabral S, […], Abdulnour RE, Rodman A. JAMA IM, April 1, 2024

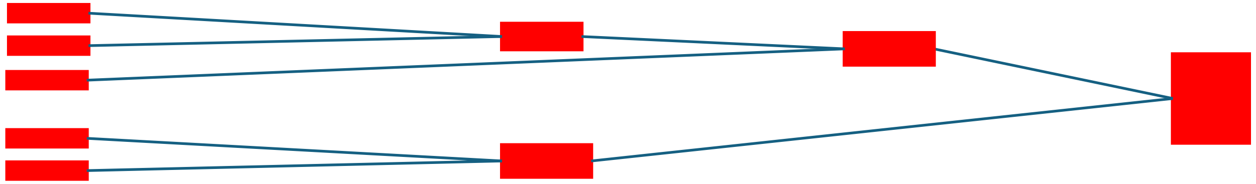
Patient Triage In The ED



1’000 patients

Clinical history and physical exam only

“You are an Emergency Department physician. Below are the symptoms of a patient presenting to the Emergency Department.



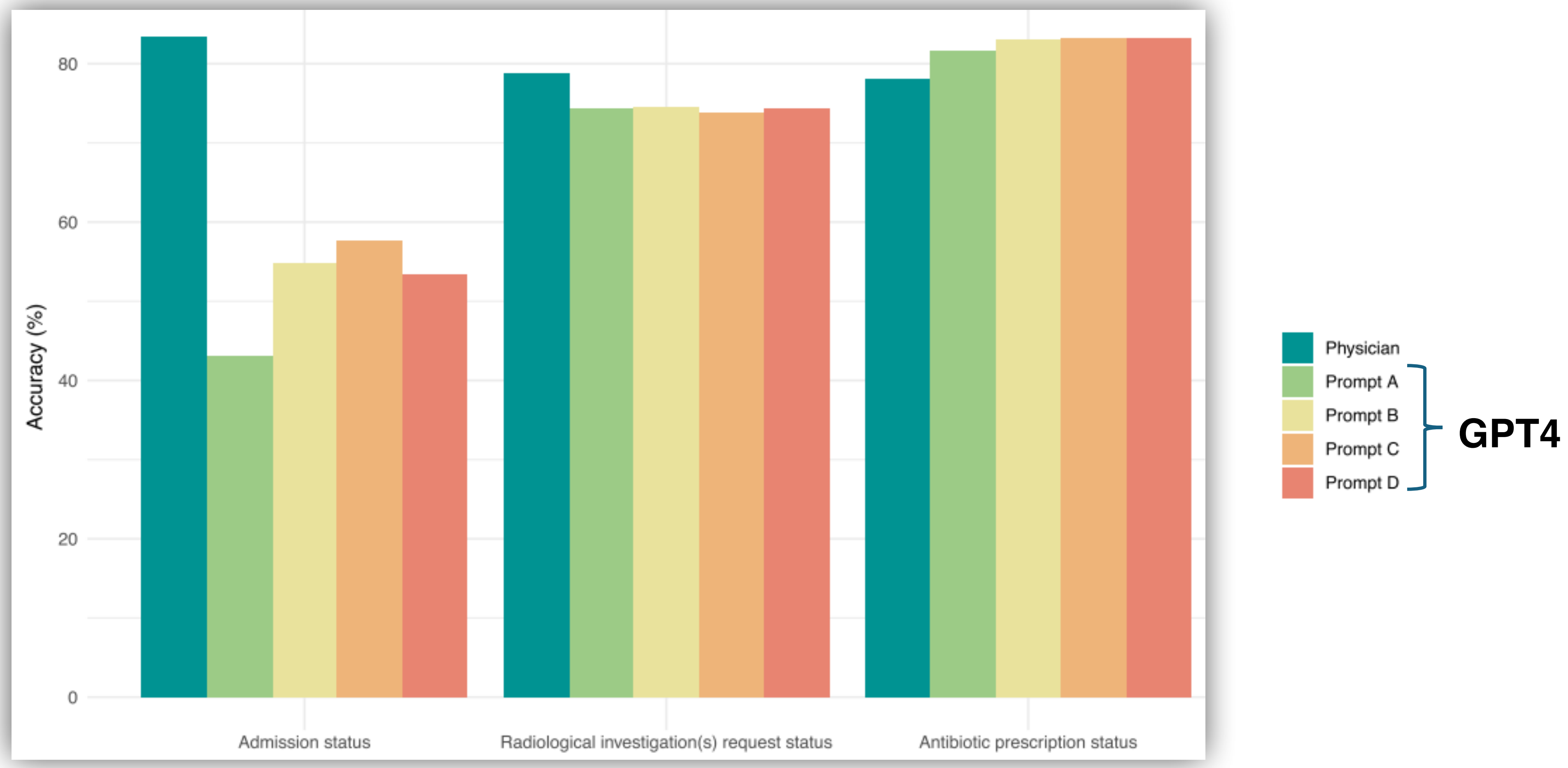
Please return whether the patient should be admitted to hospital.

Please return whether the patient requires radiological investigation (e.g X-ray, ultrasound scan, CT scan or MRI scan)

Please return whether the patient requires antibiotics

Williams, C. Y. K., Miao, B. Y., Kornblith, A. E., & Butte, A. J. (2024). *Nature Communications 2024 15:1, 15(1)*

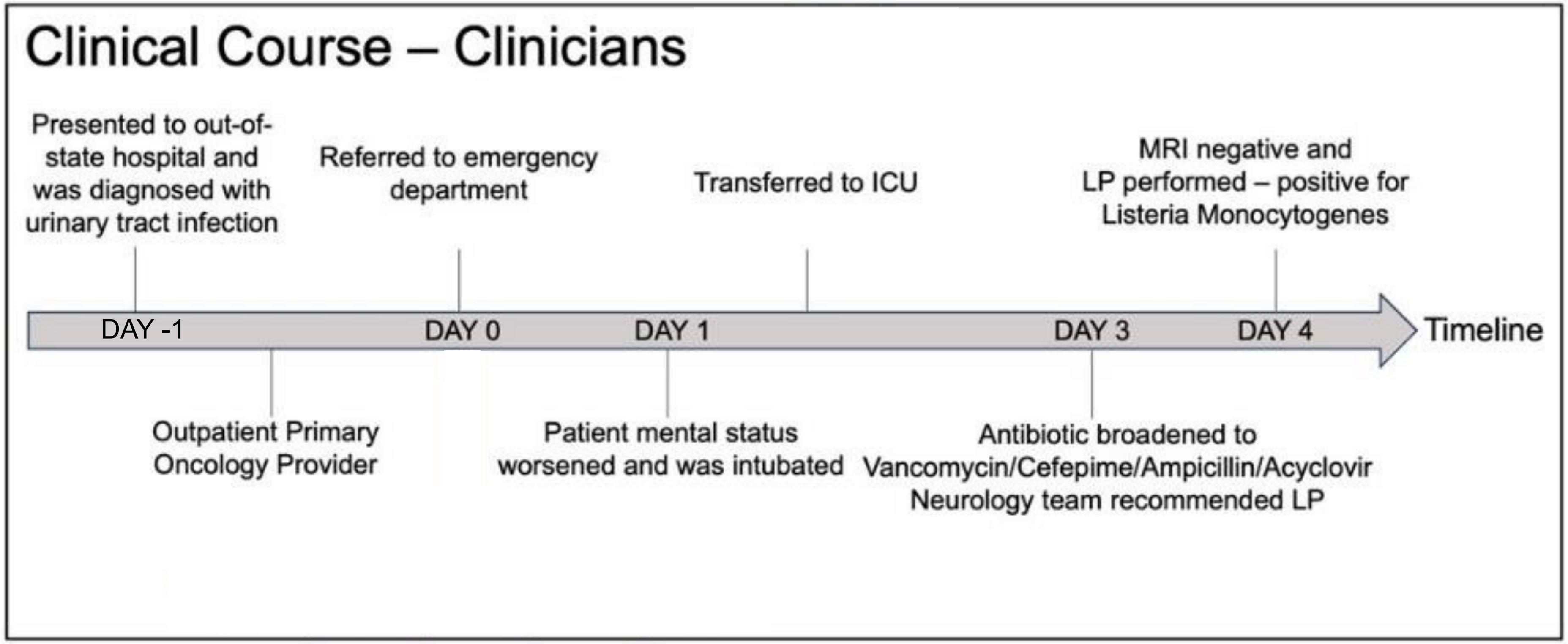
Patient Triage In The ED



Williams, C. Y. K., Miao, B. Y., Kornblith, A. E., & Butte, A. J. (2024). *Nature Communications 2024 15:1, 15(1)*

A Day In The ICU

61-year-old female with a history of hypertension, breast cancer, and recurrent ovarian cancer, presenting with a recent history of nausea, poor oral intake, and altered mental status following a ferry ride



"I want you to be the consultant on the team. I will provide you the history of present illness with the vitals, physical examination findings, labs, and imaging data. Following that, I want you to provide me with 1) a brief summary of the patient, 2) 10 Differential diagnosis (ranked by percentage likelihood), 3) further diagnostic tests you would obtain and 4) an initial management plan."

nature medicine

Article

https://doi.org/10.1038/s41591-024-03097-1

Evaluation and mitigation of the limitations of large language models in clinical decision-making

Received: 26 January 2024

Accepted: 29 May 2024

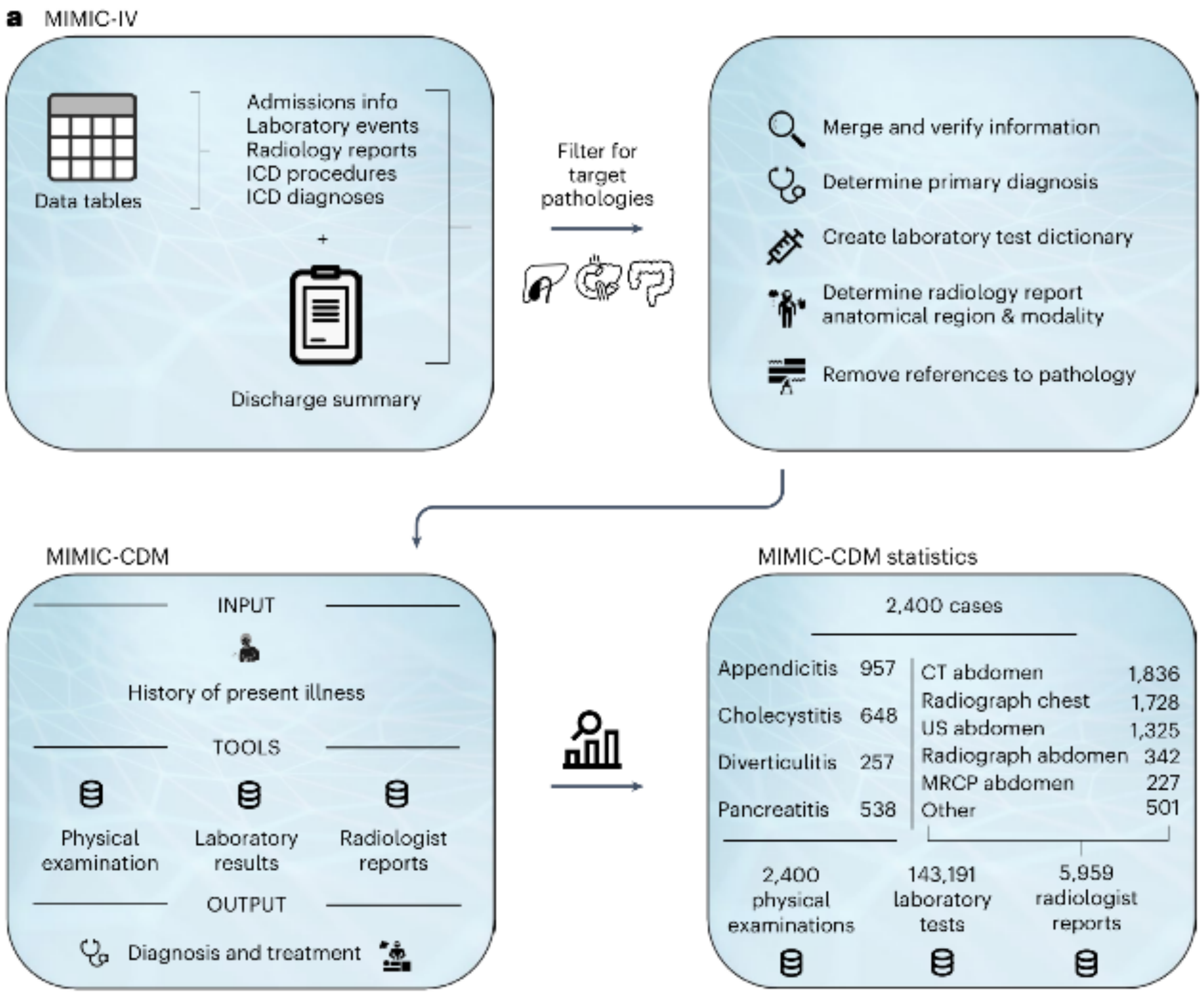
Published online: 04 July 2024

Paul Hager[1,2,8] ✉, Friederike Jungmann[1,2,8], Robbie Holland[3], Kunal Bhagat[4], Inga Hubrecht[5], Manuel Knauer[5], Jakob Vielhauer[6], Marcus Makowski[2], Rickmer Braren[2,9], Georgios Kaissis[1,2,3,7,9] & Daniel Rueckert[1,3,9]

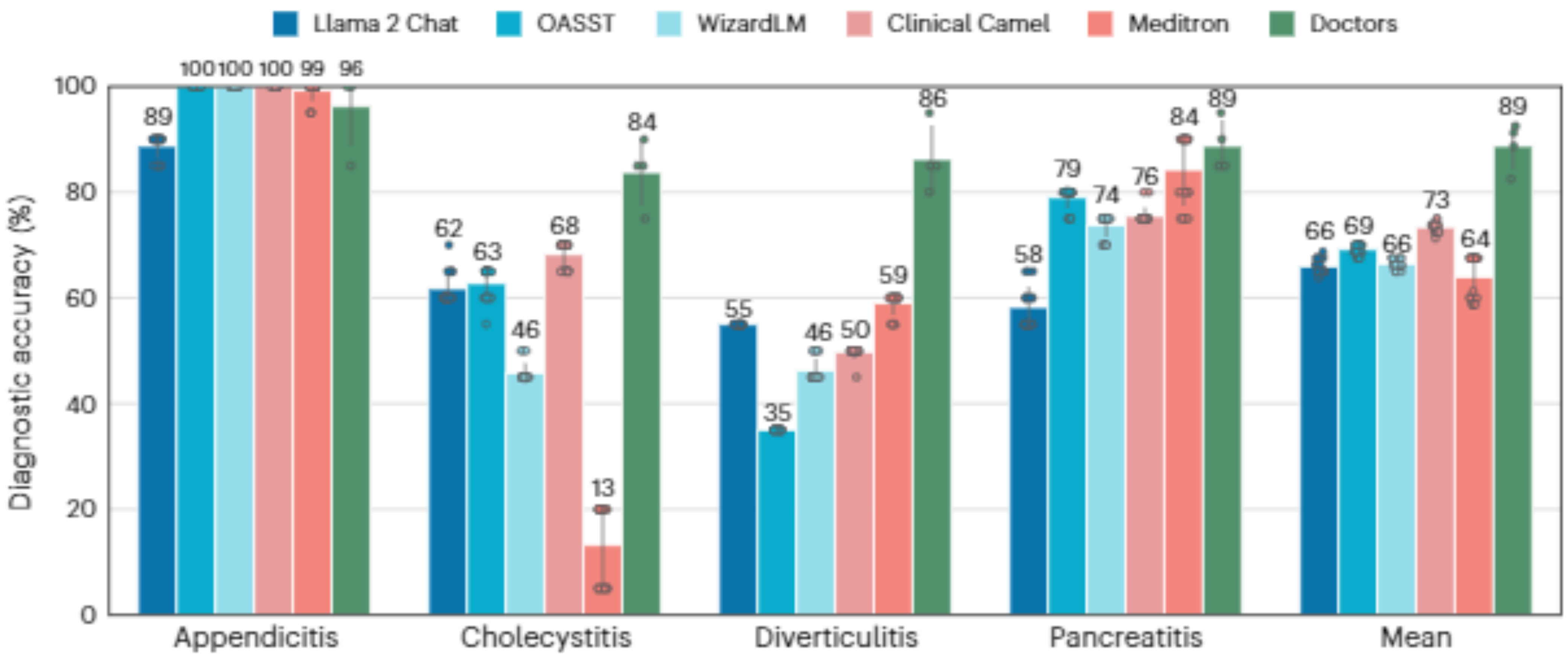
Table 1 | An overview of the considered LLMs and their properties

Model	Base	Parameters	Training dataset	
Llama 2 Chat[32]	Llama 2 (ref. 32)	70B	Public data[a]	
OASST[33]	Llama 2 (ref. 32)	70B	Public data[a], https://huggingface.co/O 70b-oasst-sft-v10/, open-source data	
WizardLM[34]	Llama 2 (ref. 32)	70B	Public data[a], Evol-Instruct generated[34]	
Clinical Camel[19]	Llama 2 (ref. 32)	70B	Public data[a], https://sharegpt.com/; Sh 2021)[19], MedQA[13]	
Meditron[35]	Llama 2 (ref. 32)	70B	Public data[a], https://huggingface.co/da guidelines, public PubMed abstracts[35],	
Chat-GPT[59]	GPT3.5 (ref. 60)	???	User conversations[b], Common Crawl[61], Books2 (ref. 63), Wikipedia	
GPT-4 (ref. 64)	???	???	???	
Med-PaLM[9]	Flan-PaLM[65]	540B	Webpages[b], Wikipedia[b], social media[b], 473 instruction fine-tuning datasets[65], HealthSearchQA[9], MedicationQA[66], LiveQA[67]	
Med-PaLM 2 (ref. 8)	PaLM 2 (ref. 68)	340B	Web Documents[b], books[b], code[b], mathematics[b], conversational data[b], MedQA[13], HealthSearchQA[9], MedicationQA[66], LiveQA[67]	✗

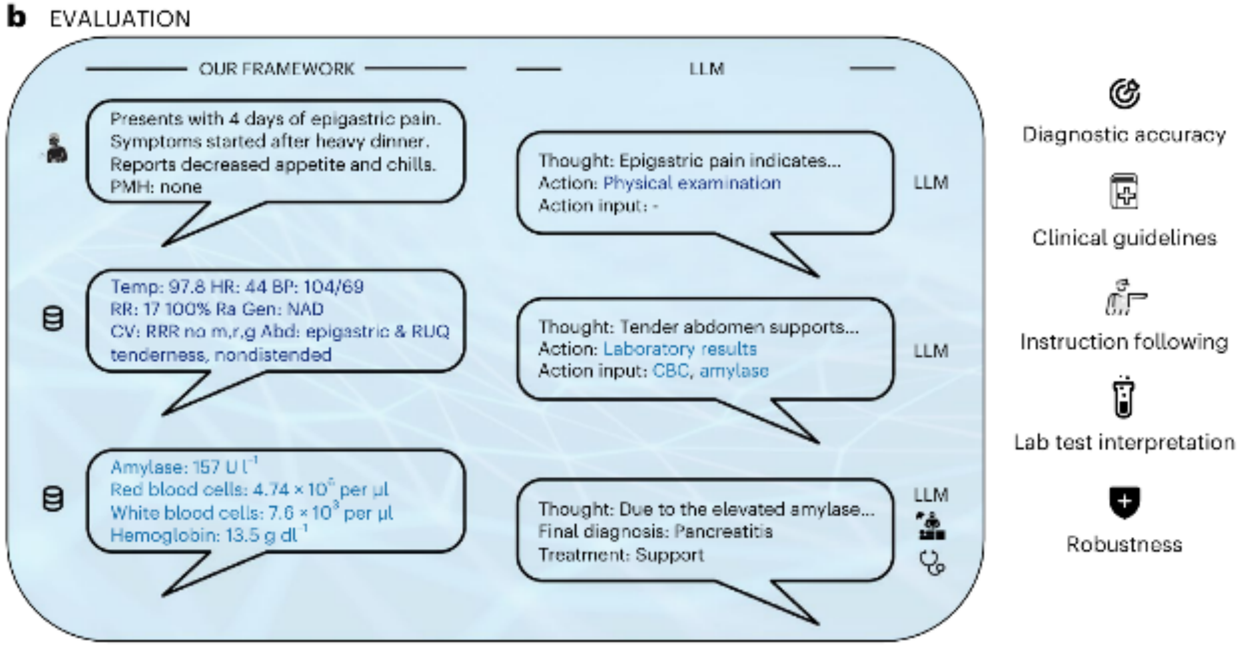
65K ICU patients → 2400 ICU patients with 4 common Dx



Evaluation of LLMs



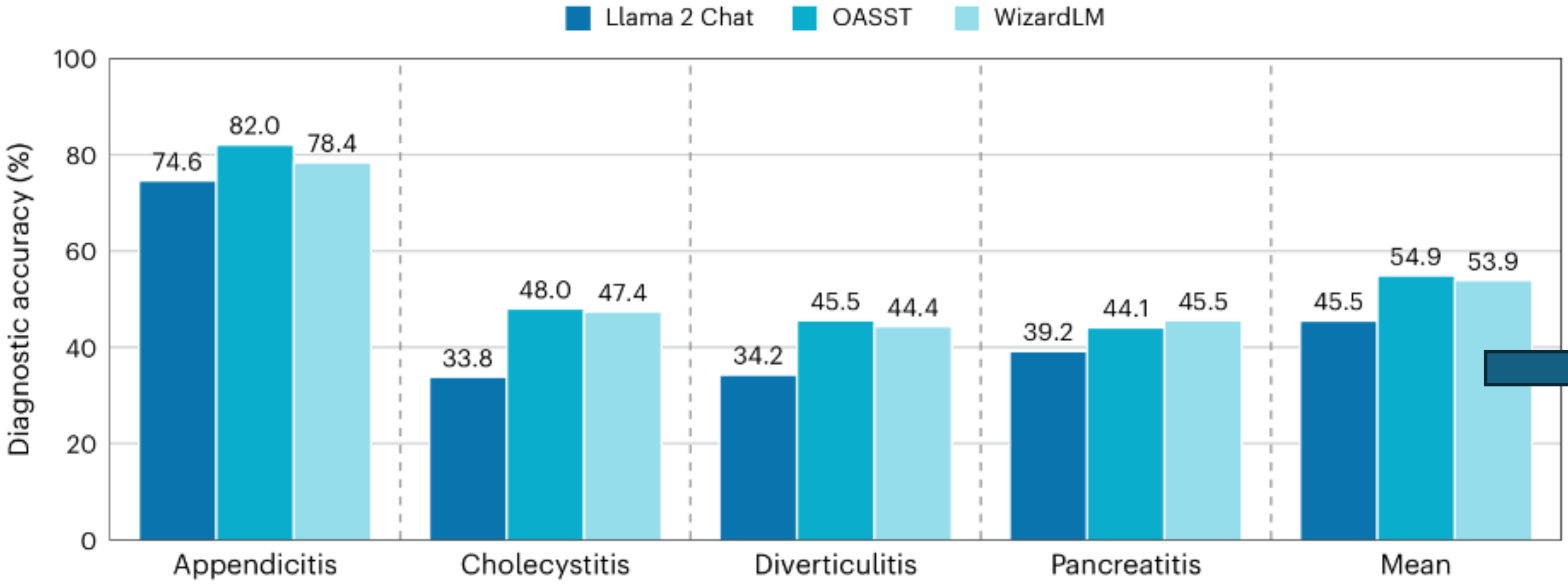
When provided with all information on a subset (n = 80), LLMs diagnose significantly worse than doctors. Mean diagnostic accuracy of LLMs over multiple seeds (n = 20) compared to clinicians (n = 4)



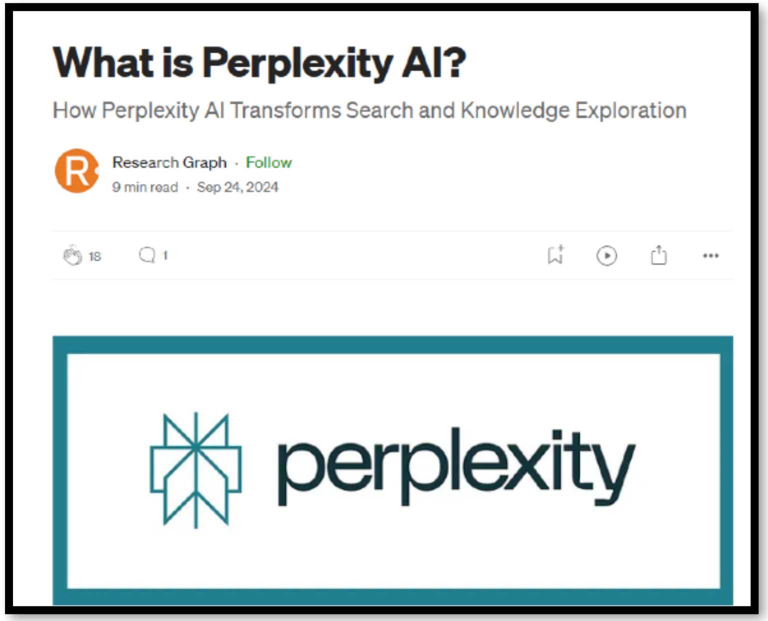
Worse performance if autonomous *(20% loss)*

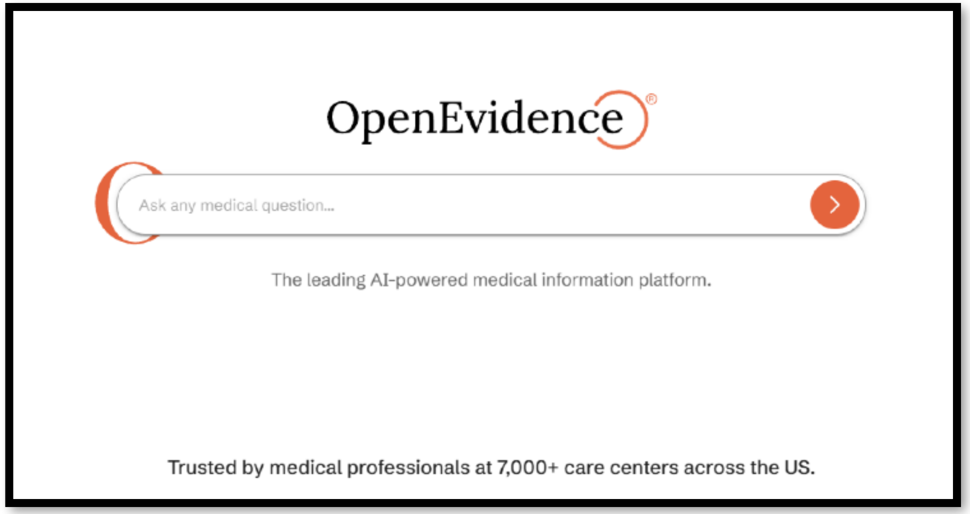
- Inability to consistently follow clinical guidelines.
- Struggles with interpreting laboratory results.
- Sensitivity to the order and amount of information provided.

minor changes in instructions can greatly change diagnostic accuracy such as asking for the 'main diagnosis' or 'primary diagnosis' instead of 'final diagnosis



Mean with all data at once






NEJM AI

NEJM AI 2024; 1 (4)
DOI: 10.1056/AIcs2300191

CASE STUDY

GPT-4 in a Cancer Center — Institute-Wide Deployment Challenges and Lessons Learned

Renato Umeton, Ph.D.,[1] Anne Kwok, B.A.,[1] Rahul Maurya, B.S.,[1] Domenic Leco, Esq., M.B.A.,[1] Naomi Lenane, B.A.,[1] Jennifer Willcox, Esq.,[1] Gregory A. Abel, M.D., M.P.H.,[1] Mary Tolikas, Ph.D., M.B.A.,[1] and Jason M. Johnson, Ph.D.,[1] for the Dana-Farber Generative AI Governance Committee*

Received: October 14, 2023; Revised: January 1, 2024; Accepted: January 22, 2024; Published: March 28, 2024

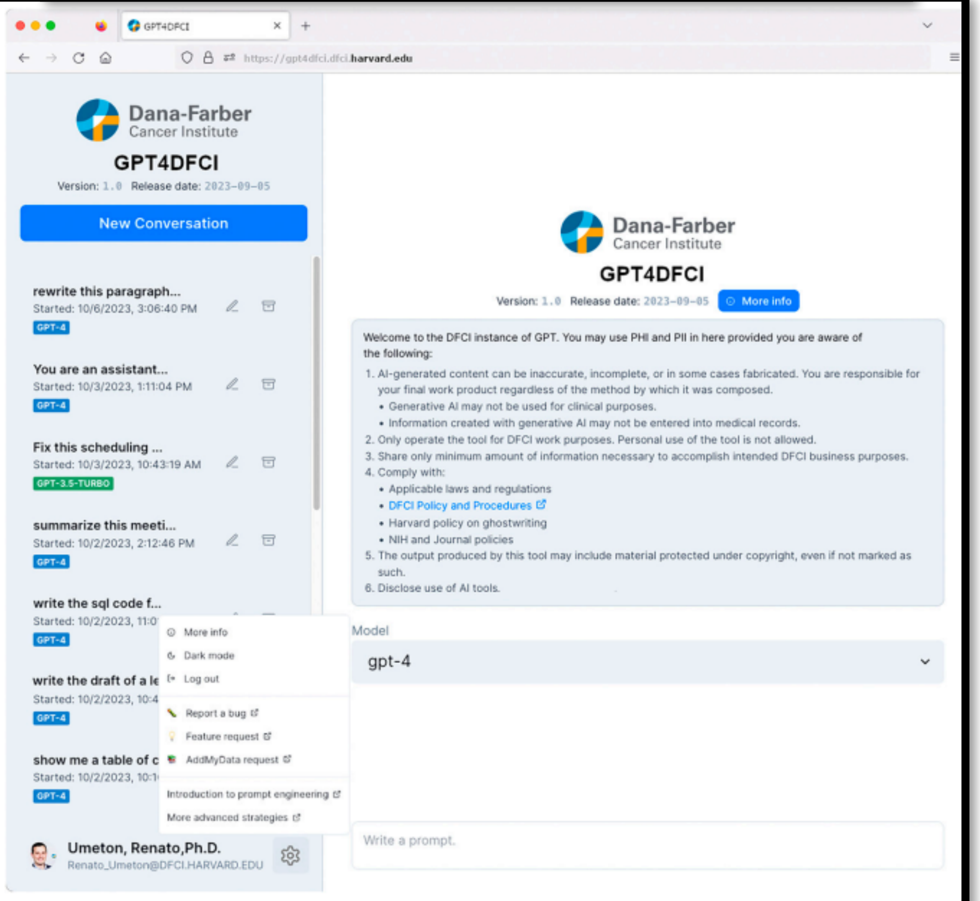
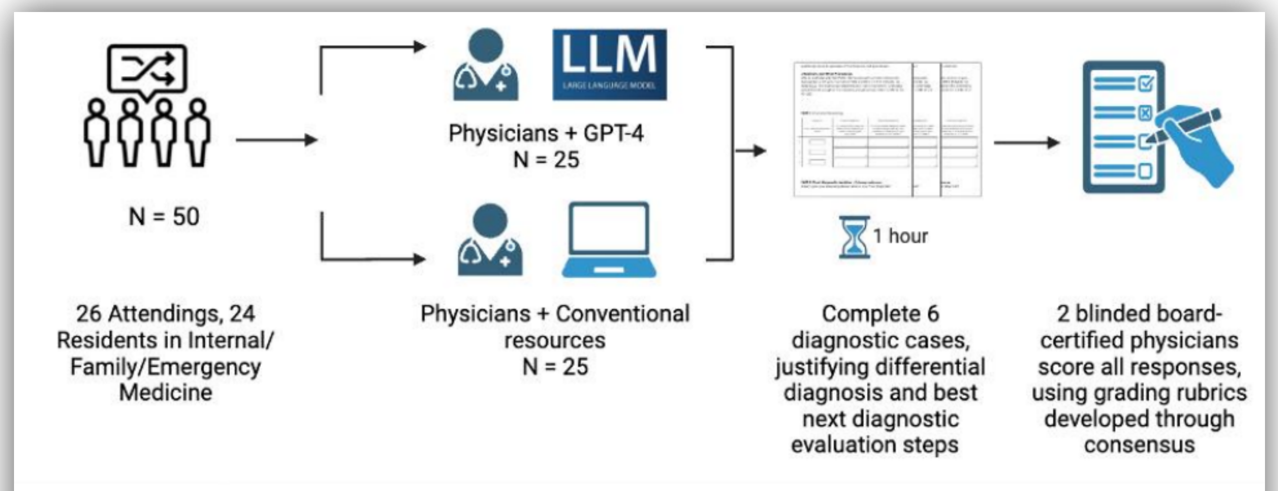
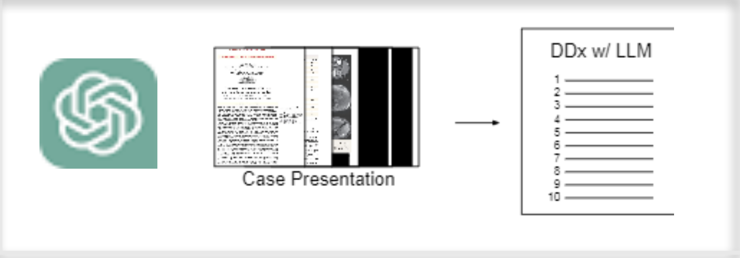


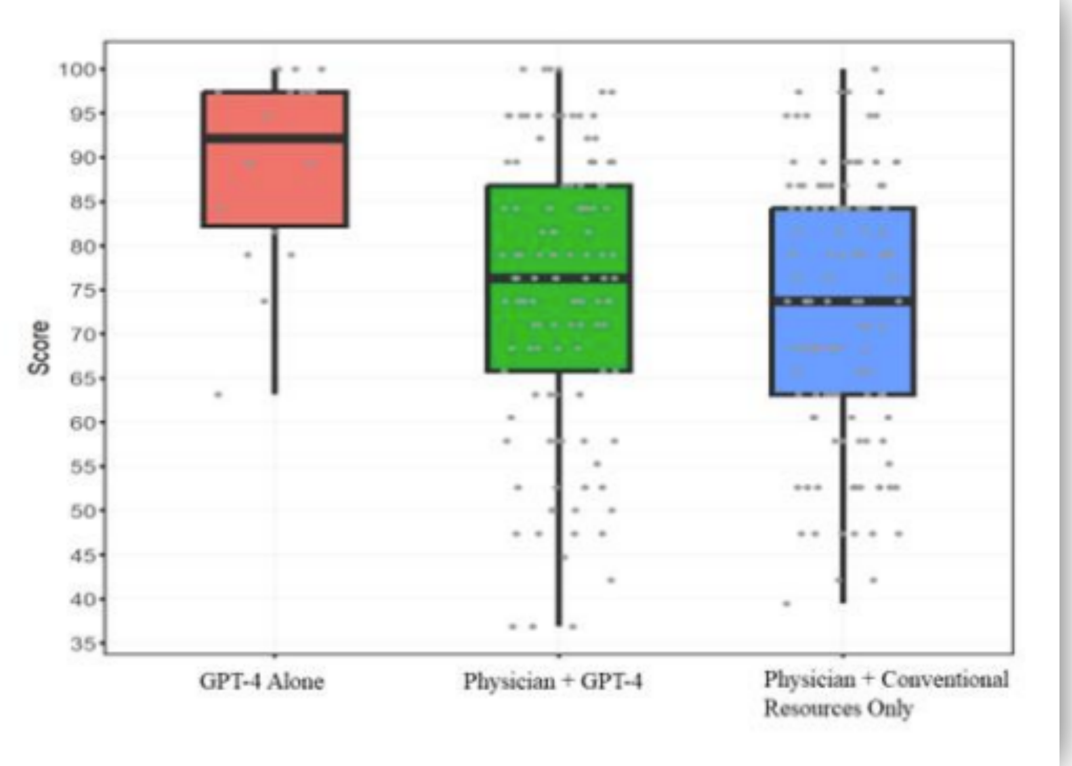
Figure 1. User Interface of GPT4DFCI v1.0 as Released on September 5, 2023.
Previous conversations appear on the left, and a summary of the terms of use appears on the right, along with model selection. Also shown is a mechanism for users to report bugs, request features, consult the user guide on prompt engineering and "More advanced strategies," and initiate the process for the creation of a retrieval-augmented generation project. GPT4DFCI denotes Generative Pretrained Transformer models for Dana-Farber Cancer Institute.

LLM Influence on Diagnostic Reasoning



50 Physicians randomized to complete a diagnosis quiz with GPT-4 vs. conventional resources. Participants were asked to offer differential diagnosis with supporting statements of findings in favor or against each differential, and to propose best next diagnosis evaluation steps.

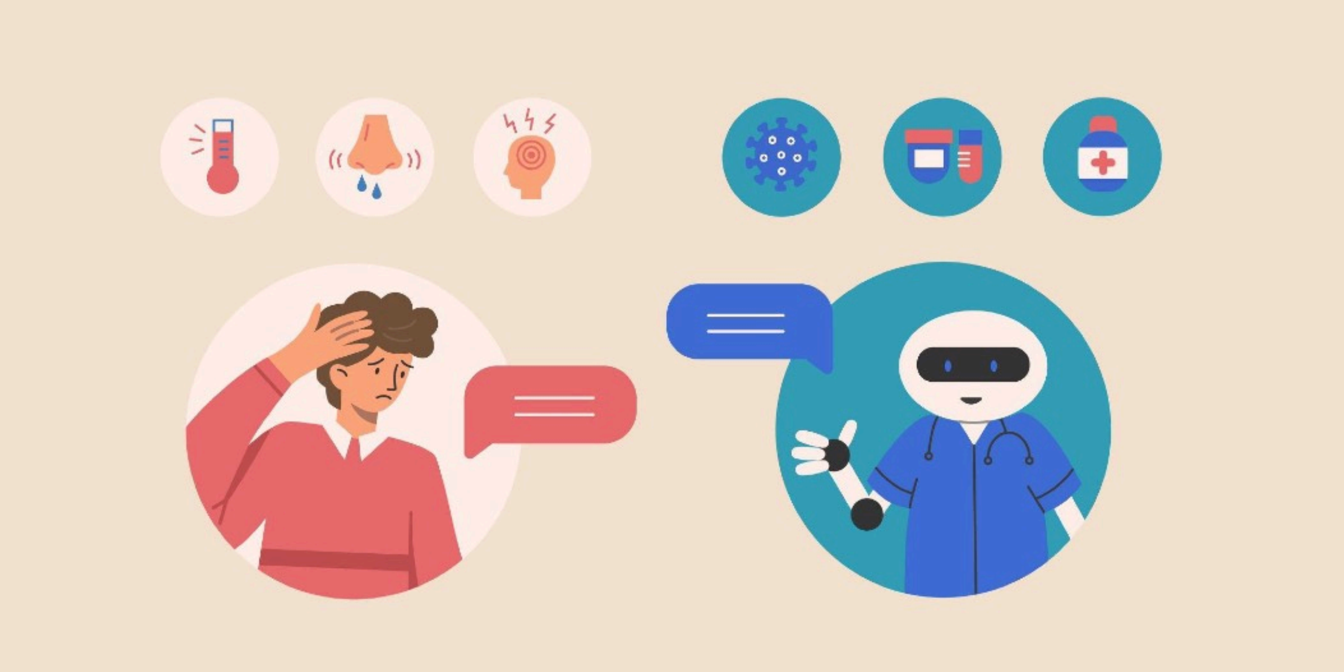


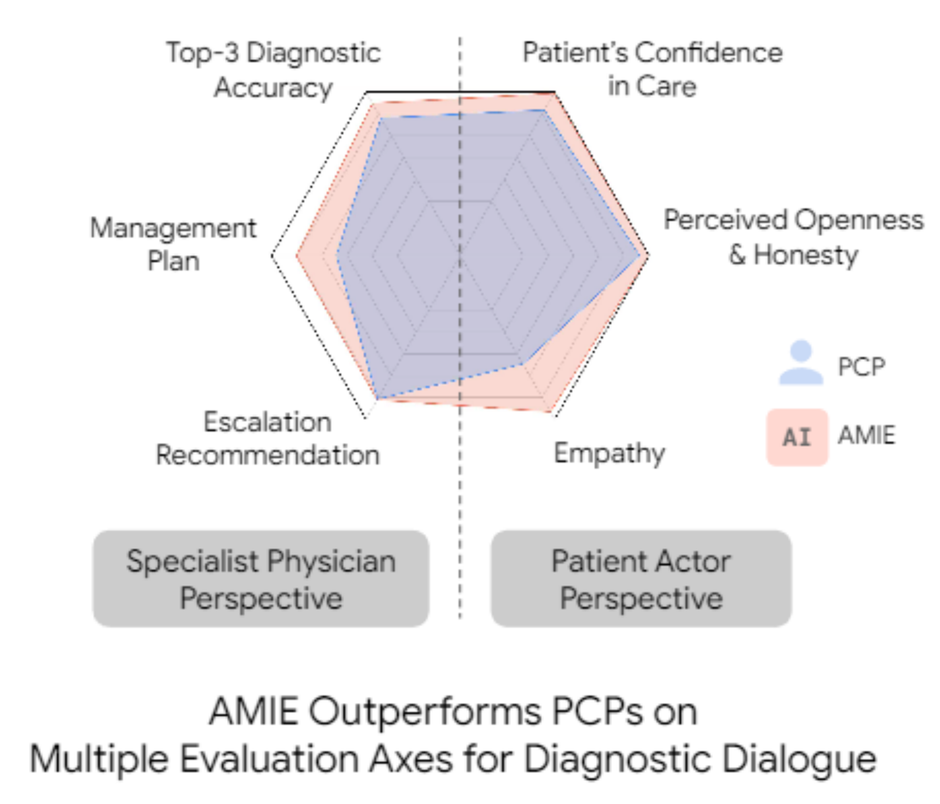


Distribution of Diagnostic Performance Scores

Goh, E., et al. (2024). "Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial." JAMA Netw Open **7(10): e2440969.**

Conversational Diagnostic AI

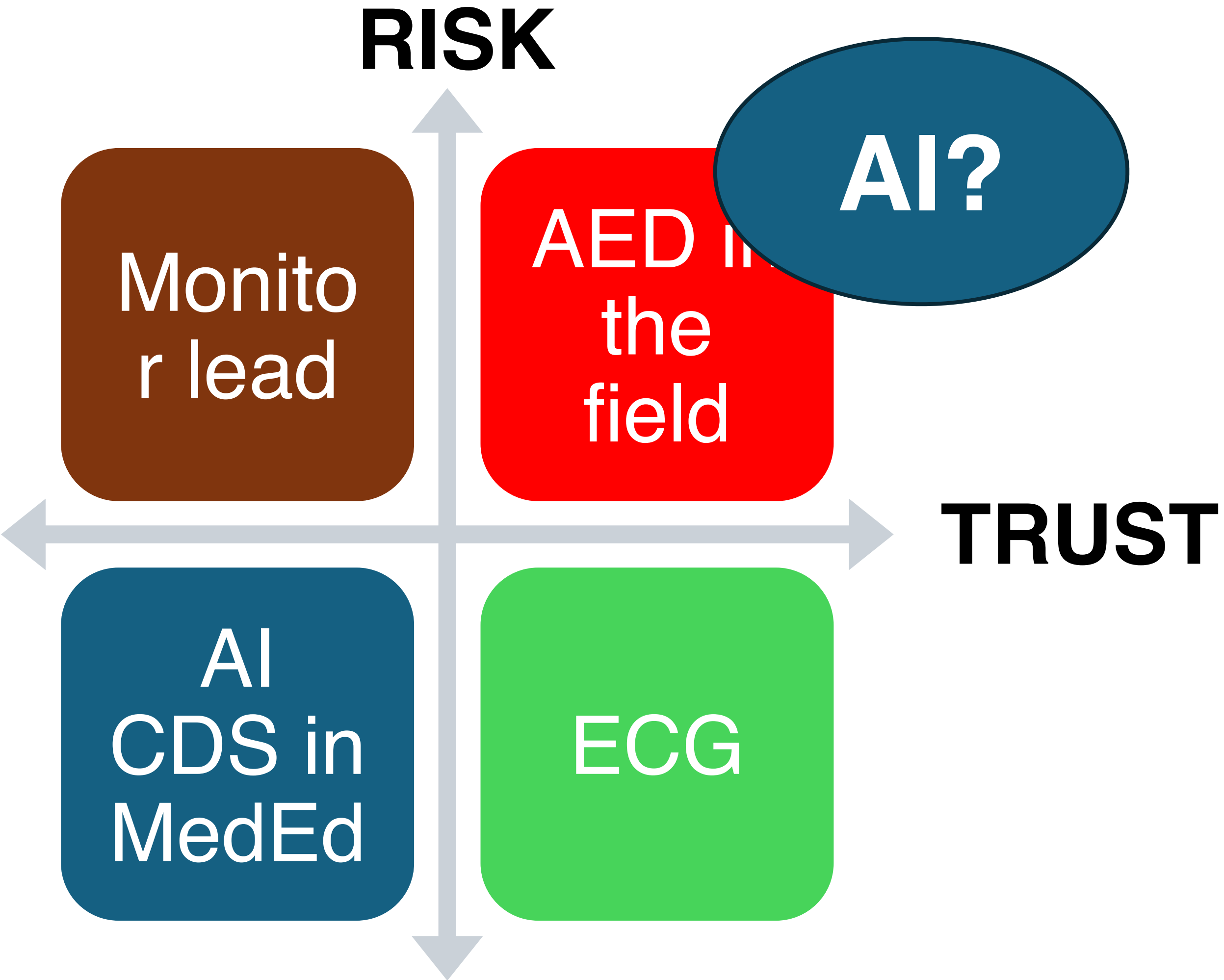




Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, *et al.* Towards Conversational Diagnostic AI. *ArXiv* 2024.

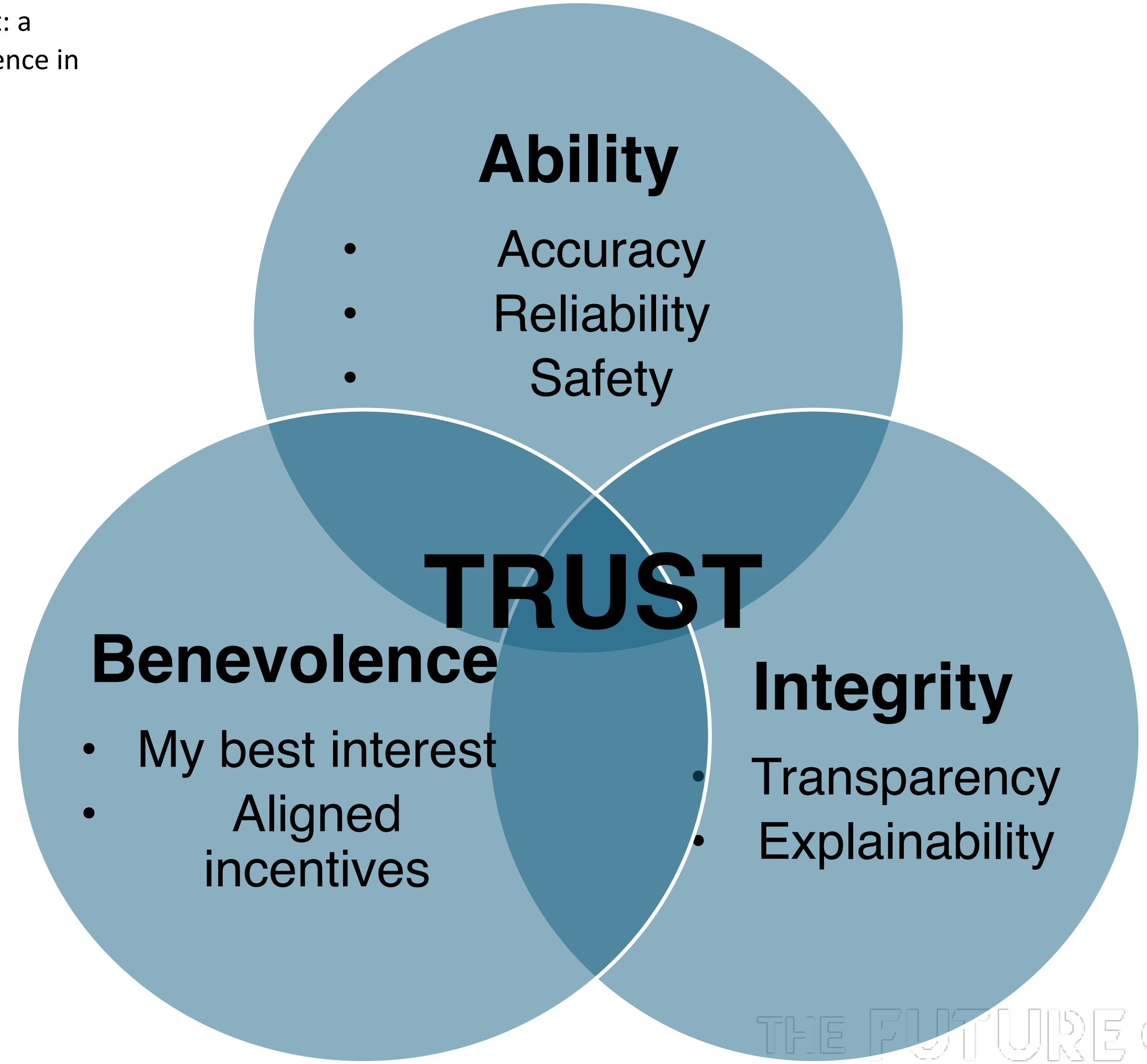
Entrustment In Patient Care

Is this patient having ventricular tachycardia?



Entrustment In Patient Care

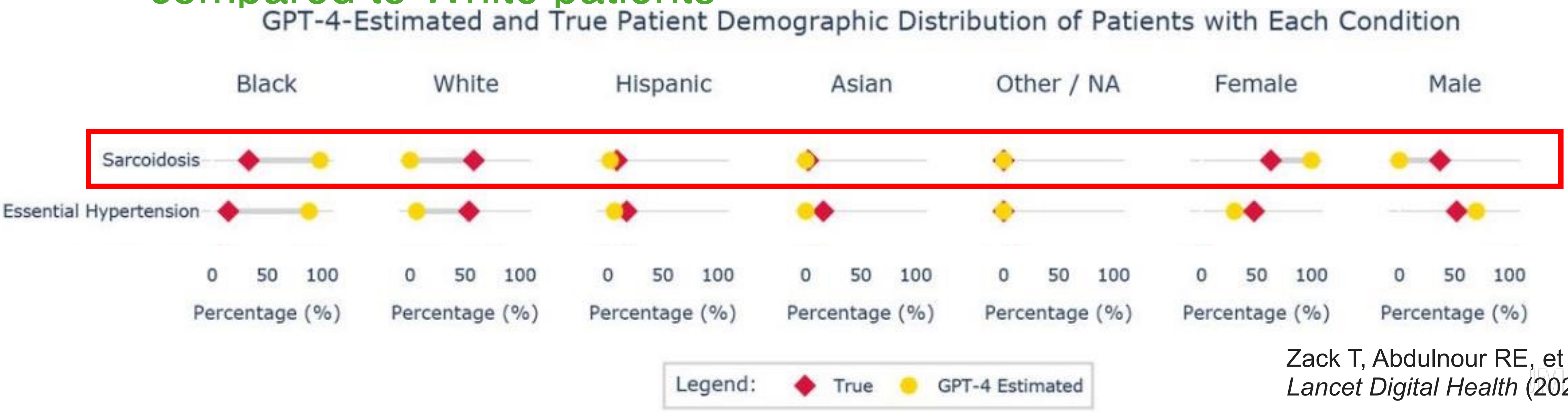
Brian C. Gin, M., PhD, et al. (In press). "Entrustment: a framework to safeguard the use of artificial intelligence in health professions education." Acad Med.



Is AI safe?

LLMs contain the bias of their training

- Asked GPT-4 to create clinical vignettes
  - Over-represented demographic stereotypes of diseases
- Asked GPT-4 to give management plans for cases while substituting gender and race/ethnicity
  - Less likely to recommend advanced imaging for Black patients compared to White patients

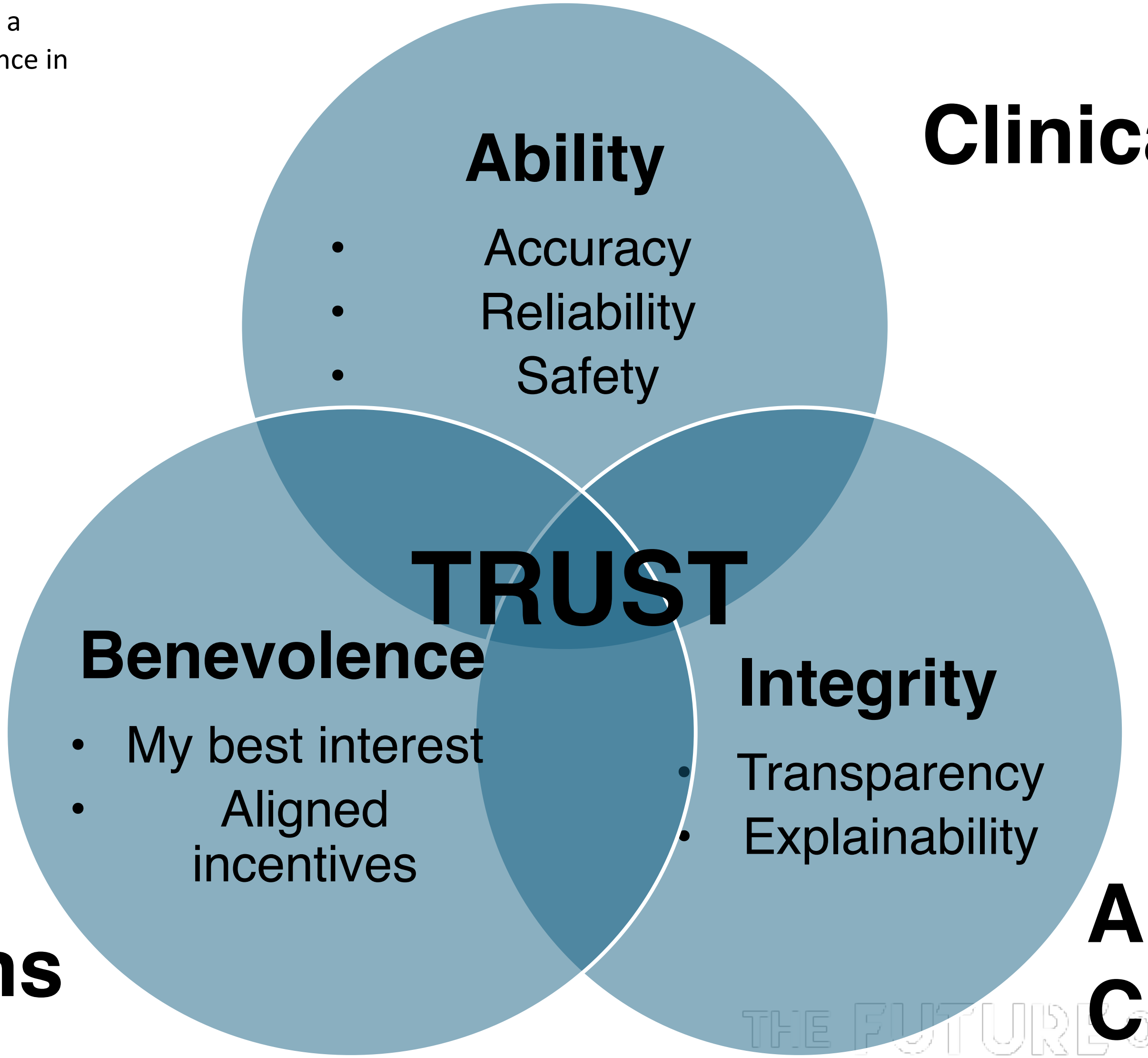


Zack T, Abdulnour RE, et al. *Lancet Digital Health* (2023).

Entrustment In Patient Care

Brian C. Gin, M., PhD, et al. (In press). "Entrustment: a framework to safeguard the use of artificial intelligence in health professions education." Acad Med.

Clinical trials



Regulation by third-party (e.g., FDA) and patients/clinicians

AI report cards
Chain-of-Thought

Do you want more?

- **Option1**: Thank you!!!

- **Option2**: Next slide.

Clinical Trials

- Summarize existing evaluations of LLMs in health care
- A systematic search of PubMed and Web of Science was performed for studies published between January 1, 2022, and February 19, 2024
- 519 studies reviewed, published between January 1, 2022, and February 19, 2024

Health care tasks	Accuracy	Comprehensiveness	Factuality	Robustness	Fairness, bias, and toxicity evaluation	Deployment metrics	Calibration and uncertainty
Enhancing medical knowledge	222	91	44	33	16	10	3
Making diagnoses	100	38	11	11	14	4	0
Educating patients	88	68	32	22	18	3	2
Making treatment recommendations	47	22	9	8	3	1	0
Communicating with patients	35	29	8	15	22	1	0
Care coordination and planning	36	24	4	5	7	1	0
Triaging patients	24	7	5	2	8	8	0
Carrying out a literature review	18	7	3	2	2	2	0
Synthesizing data for research	16	7	2	3	2	2	0
Generating medical reports	8	8	2	0	3	0	0
Conducting medical research	8	7	3	3	3	0	0
Providing asynchronous care	8	5	3	3	1	1	0
Managing clinical knowledge	5	5	1	1	0	0	0
Clinical note-taking	6	2	1	1	0	0	1
Generating clinical referrals	3	0	0	0	0	0	0
Enhancing surgical operations	3	3	1	1	0	0	0
Biomedical data mining	2	0	0	0	0	0	0
Generating billing codes	1	0	0	0	0	0	0
Writing prescriptions	1	0	0	0	0	0	0
NLP and NLU tasks							
Question answering	398	194	71	61	54	14	5
Text classification	29	10	6	5	10	2	0
Information extraction	29	12	8	5	4	6	0
Summarization	29	21	7	3	8	0	1
Conversational dialogue	6	6	1	1	5	1	0
Translation	5	1	2	2	1	2	0

Dimension of evaluation

Bedi, S., et al. (2024). "Testing and Evaluation of Health Care Applications of Large Language Models." JAMA.

Chain-of-Thought prompting

What is the cause of this patient's respiratory failure?

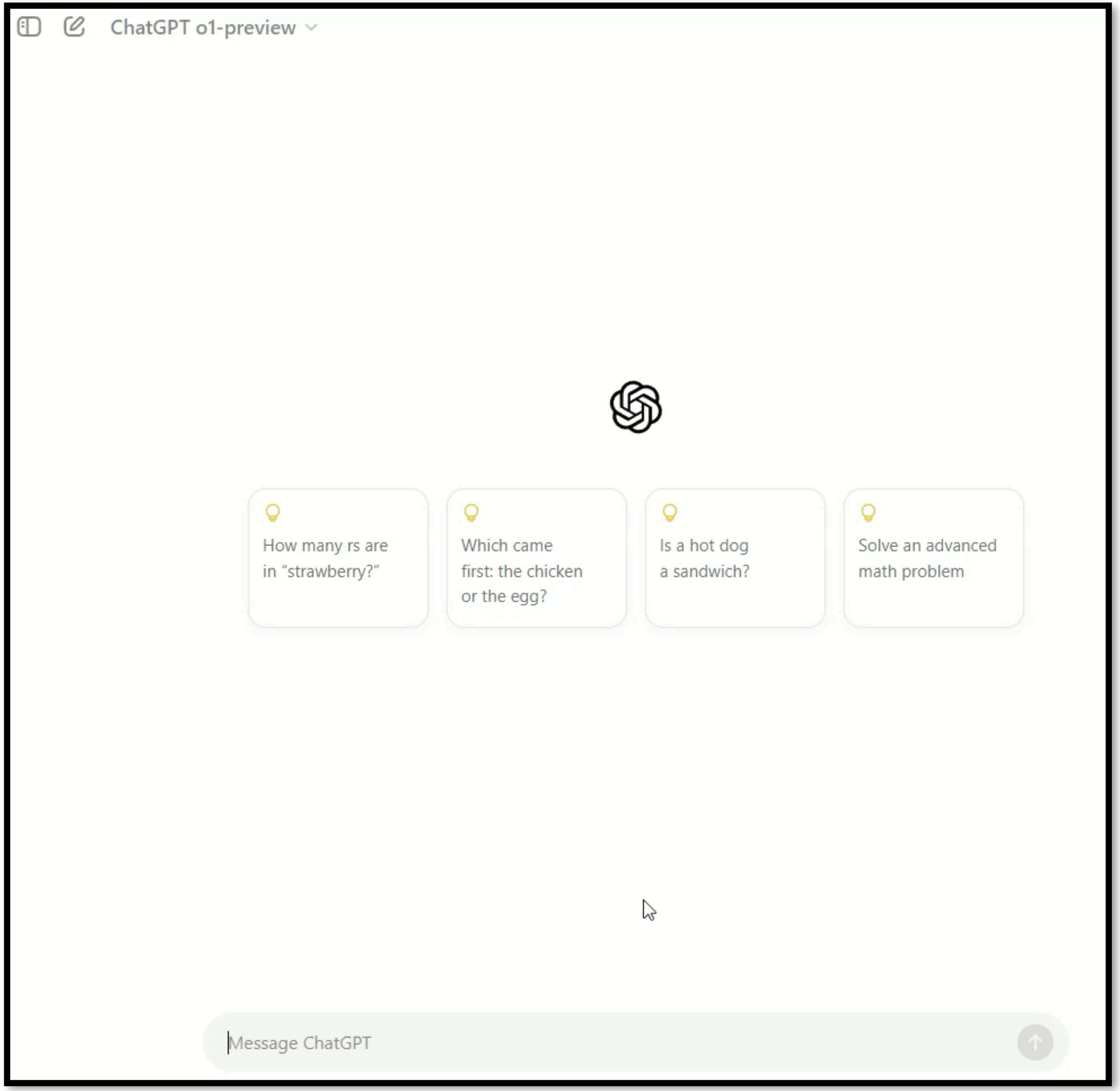
This patient has ARDS

Explain your reasoning

Because their symptoms started three months ago and they like Candy

Chain-of-Thought prompting

What is the cause of this patient's respiratory failure? **Explain your reasoning.**

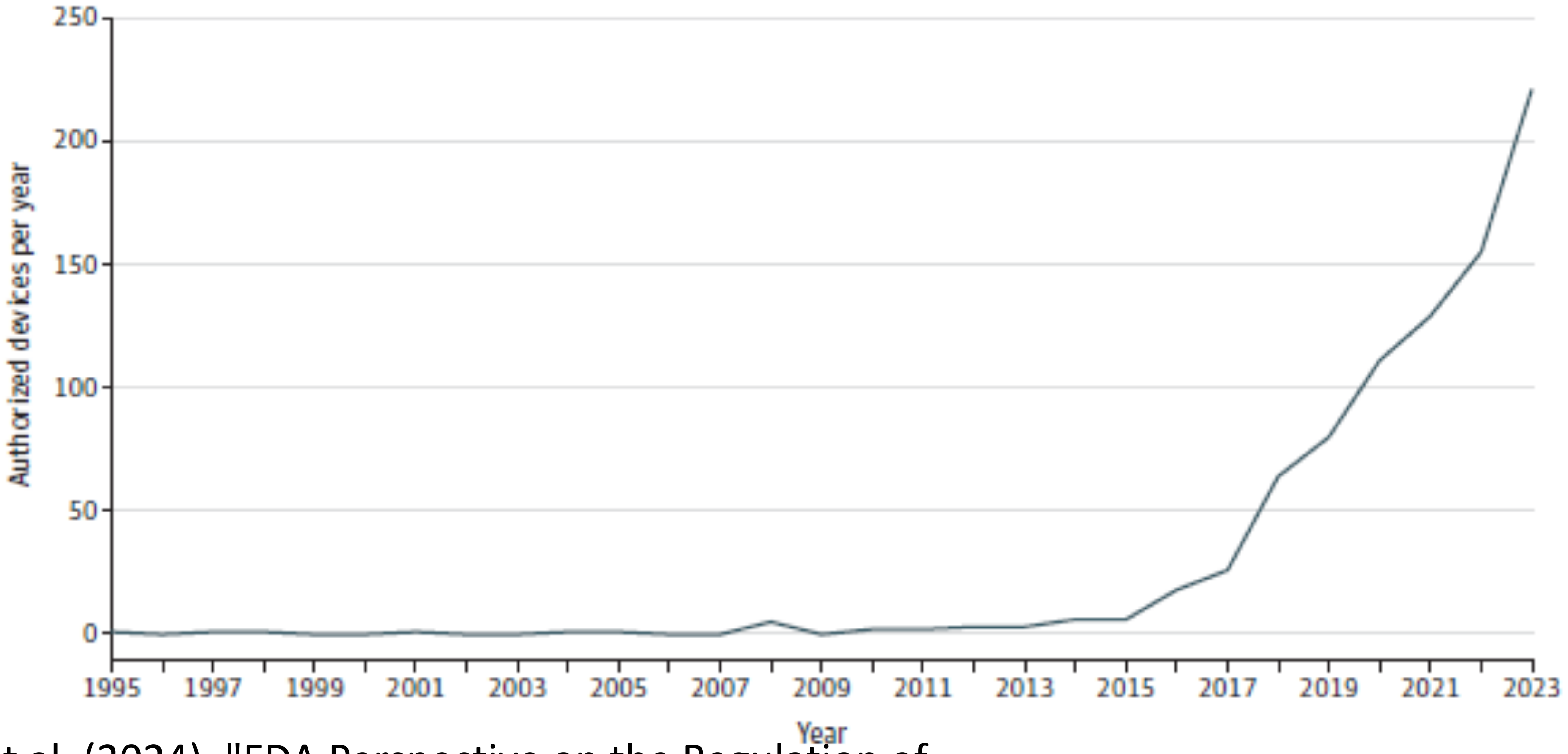

ChatGPT o1-preview
How many rs are in "strawberry?"
Which came first: the chicken or the egg?
Is a hot dog a sandwich?
Solve an advanced math problem
Message ChatGPT

Do you want more?

- **Option1**: Thank you!!!
- **Option2**: Next slide.

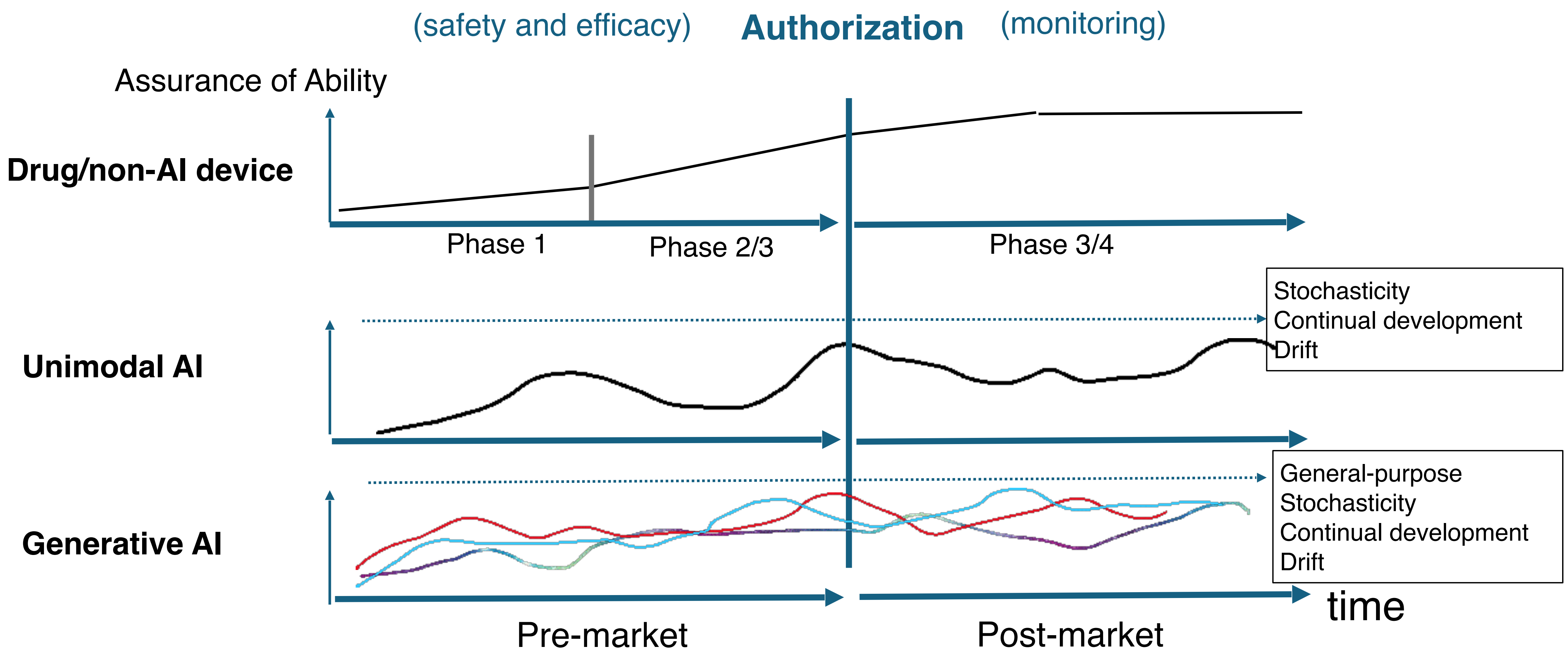
Artificial Intelligence–Enabled Medical Devices

Figure 1. Artificial Intelligence–Enabled Medical Devices Authorized for Marketing by the US Food and Drug Administration, by Year

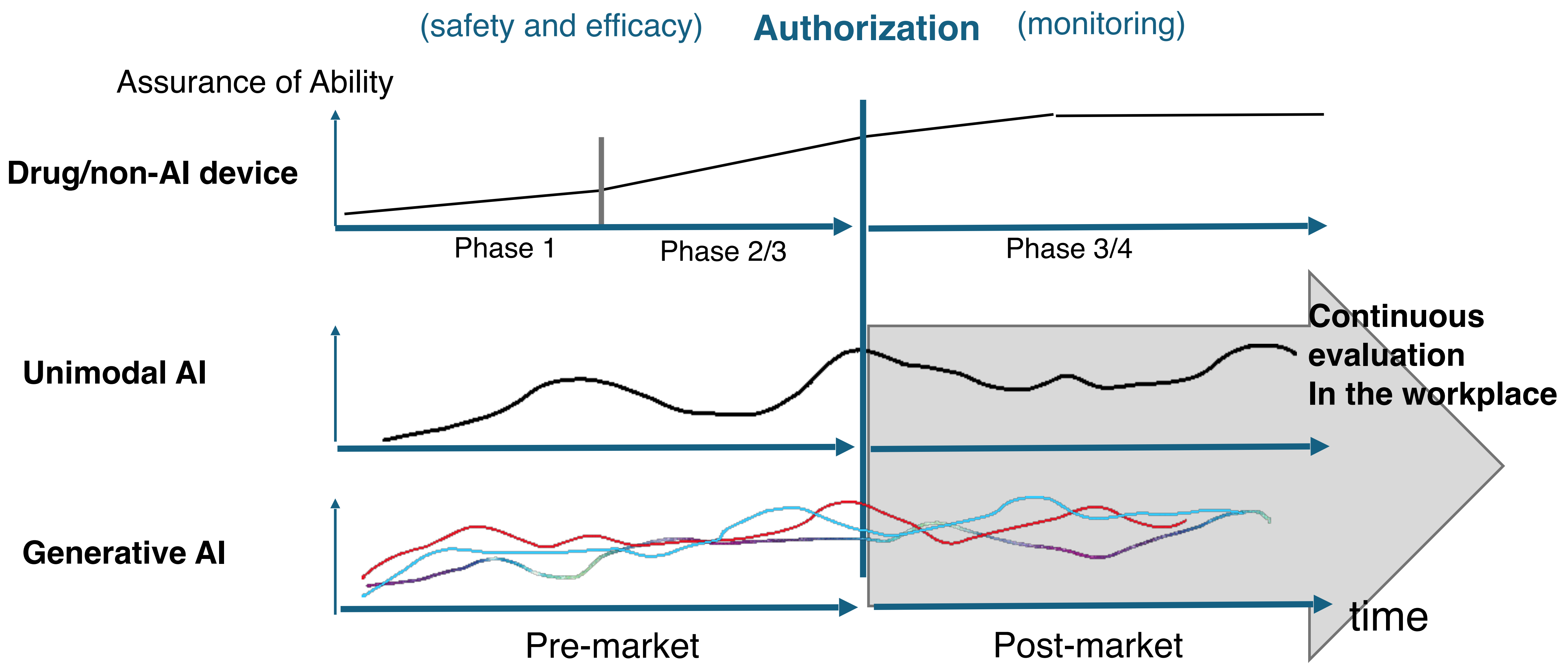


Warraich, H. J., et al. (2024). "FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine." JAMA

Life Cycle Regulation



Life Cycle Regulation



Conclusions

“Continuous complementary efforts to better understand how AI performs in the settings in which it is deployed. This will entail a comprehensive approach reaching far beyond the FDA, spanning the consumer and health care ecosystems to keep pace with accelerating technical progress.”

“Strong oversight by the FDA and other agencies aims to protect the long-term success of regulated products by maintaining a high grade of public trust in the regulated space.”

“Regulated industries, academia, and the FDA will need to develop and optimize the tools needed to assess the ongoing safety and effectiveness of AI in health care and biomedicine. The FDA will continue to play a central role with a focus on health outcomes, but all involved sectors will need to attend to AI with the care and rigor this potentially transformative technology merits.”

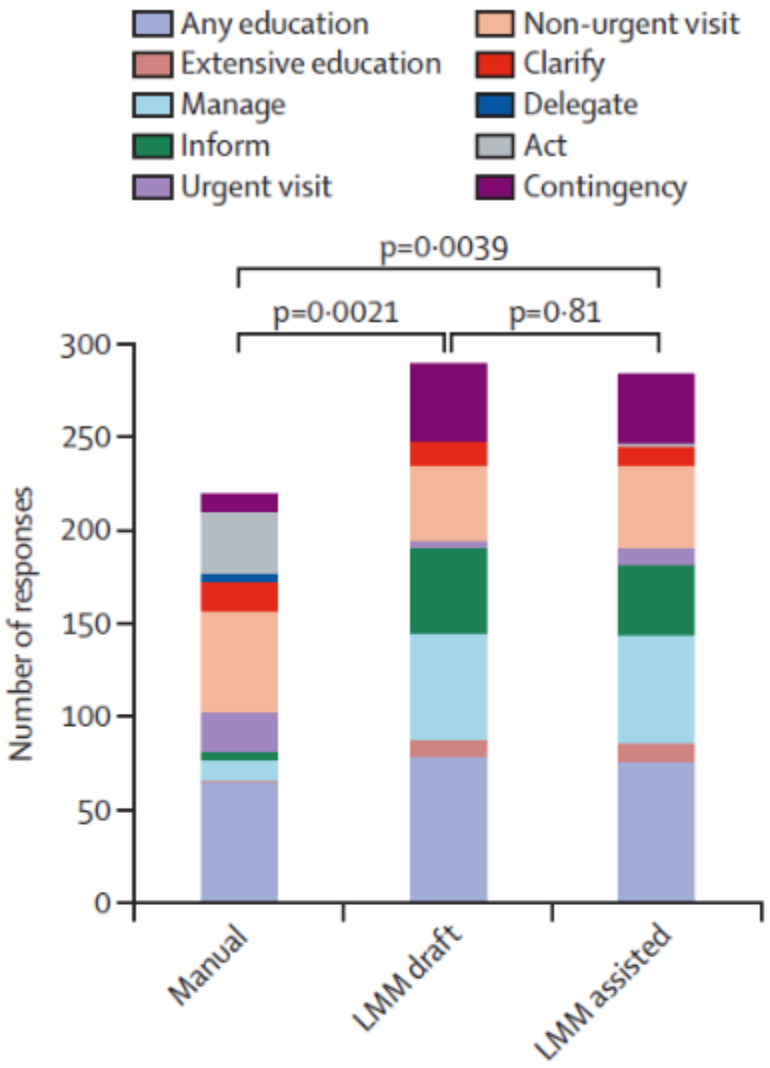
Warraich, H. J., et al. (2024). "FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine." JAMA

Do you want more?

- **Option1**: Thank you!!!
- **Option2**: Next slide.

Impact of using LLM to reply to patient messages

- The mean manual response (34 words) was **shorter** than the LLM draft (169 words) and LLM-assisted responses (160 words; p<0·0001).
- The assessing physicians felt that the **LLM drafts posed a risk of severe harm in 7·1%** of survey responses and **death in one (0·6%)** survey response.
- Most harmful responses were due to *incorrectly determining or conveying the acuity of the scenario and recommended action.*
- The assessing physicians reported that the **LLM draft improved subjective efficiency in 76·9% of cases.**



The effect of using a large language model to respond to patient messages. Chen, Shan et al. The Lancet Digital Health, 2024

NEJM Catalyst | Innovations in Care Delivery

COMMENTARY

Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation

Aaron A. Tierney, PhD, Gregg Gayre, MD, Brian Hoberman, MD, MBA, Britt Mattern, MBA, Manuel Ballesca, MD, Patricia Kipnis, PhD, Vincent Liu, MD, MS, Kristine Lee, MD

Vol. 5 No. 3 | March 2024

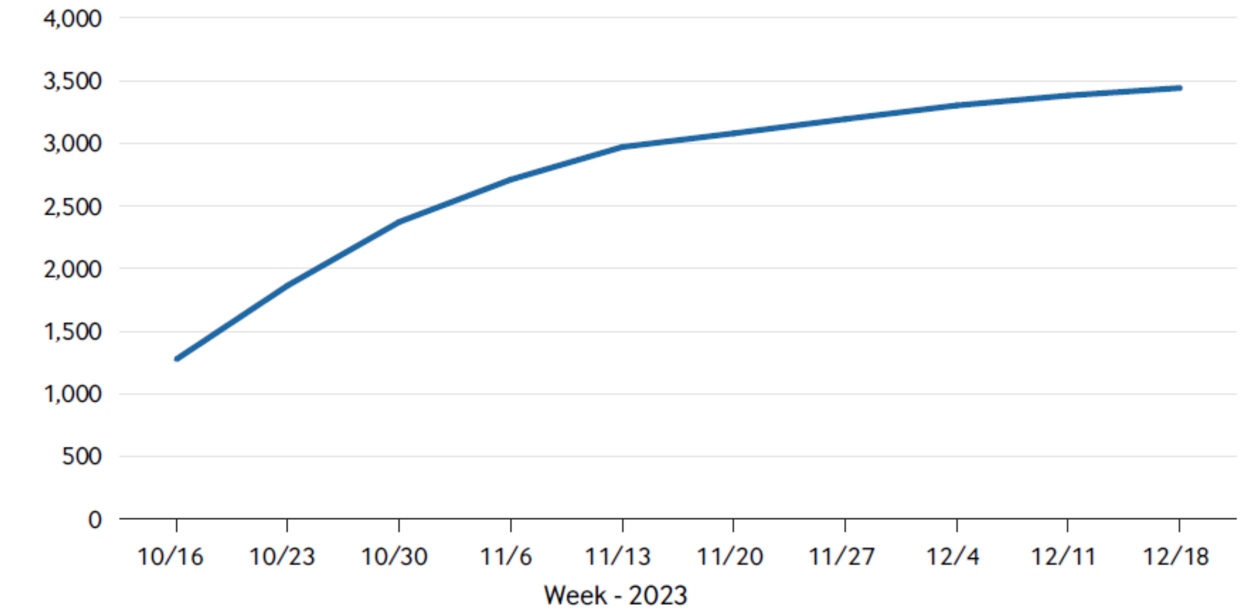
DOI: 10.1056/CAT.23.0404

Pilot in 10000 physicians
Oct 2023 – Dec 2023
Modified PDQI-9, patient surveys, testimonials
Included assessment of confabulations and bias

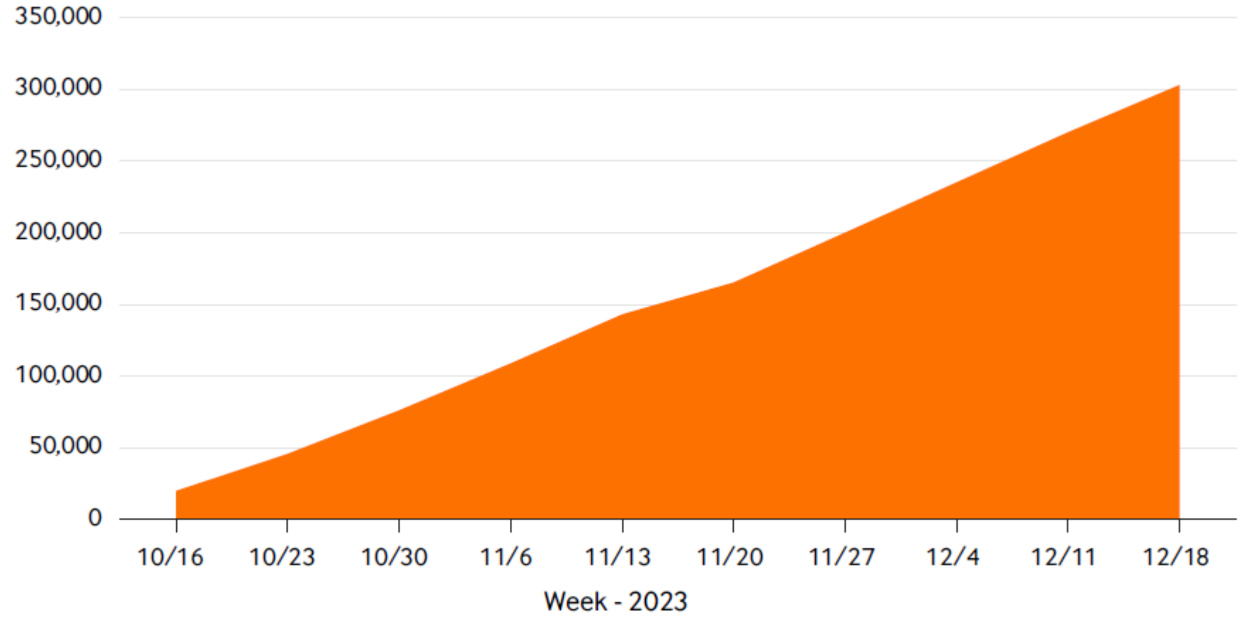
The study had four primary goals:

1. To assess uptake and engagement by both patients and clinicians.
2. To evaluate the effectiveness of the AI scribe in real clinical settings.
3. To determine if the AI scribe enhances the physician-patient relationship.
4. To verify that documentation quality was maintained.

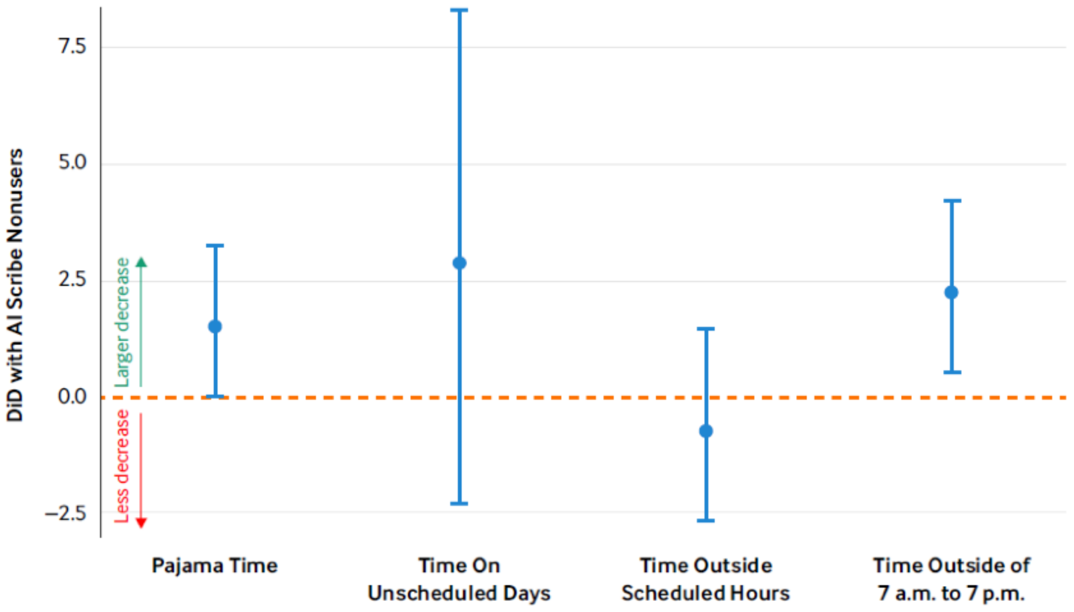
Panel A. Unique Physicians Ever Using AI Scribe



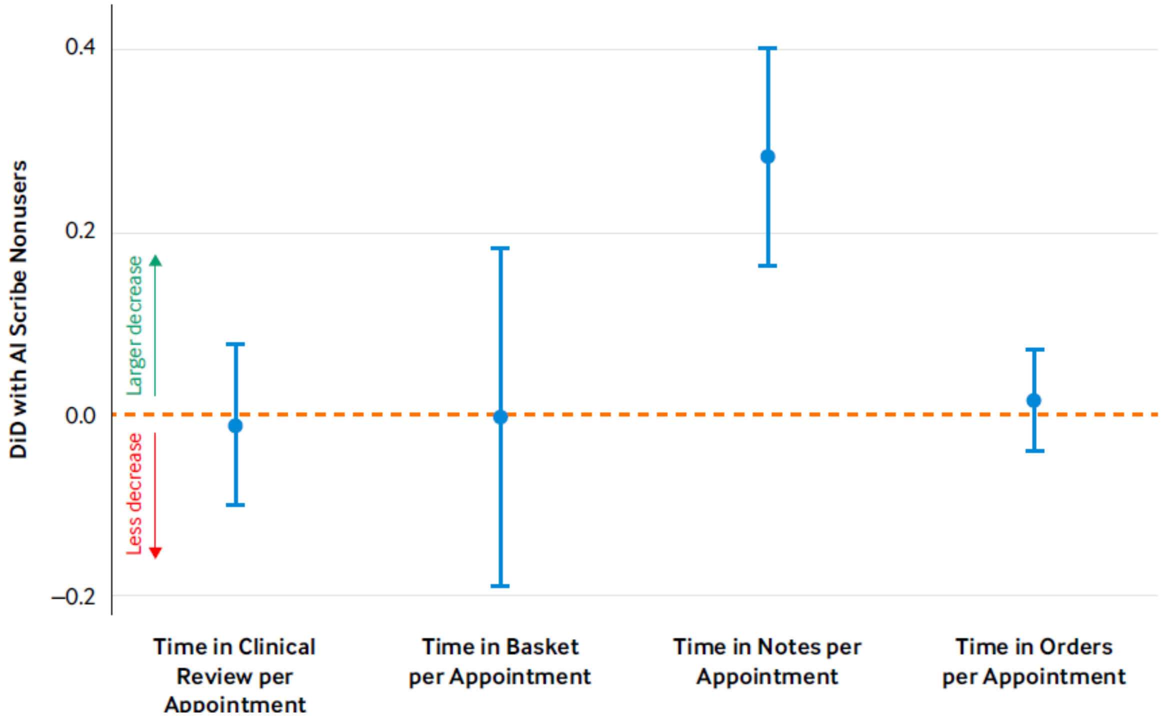
Panel B. Cumulative AI Scribe Visits



Panel A. Primary Care Physician Time Spent in the EHR-Related Activities



Panel B. Primary Care Physician Time Spent in Appointment-Related Activities



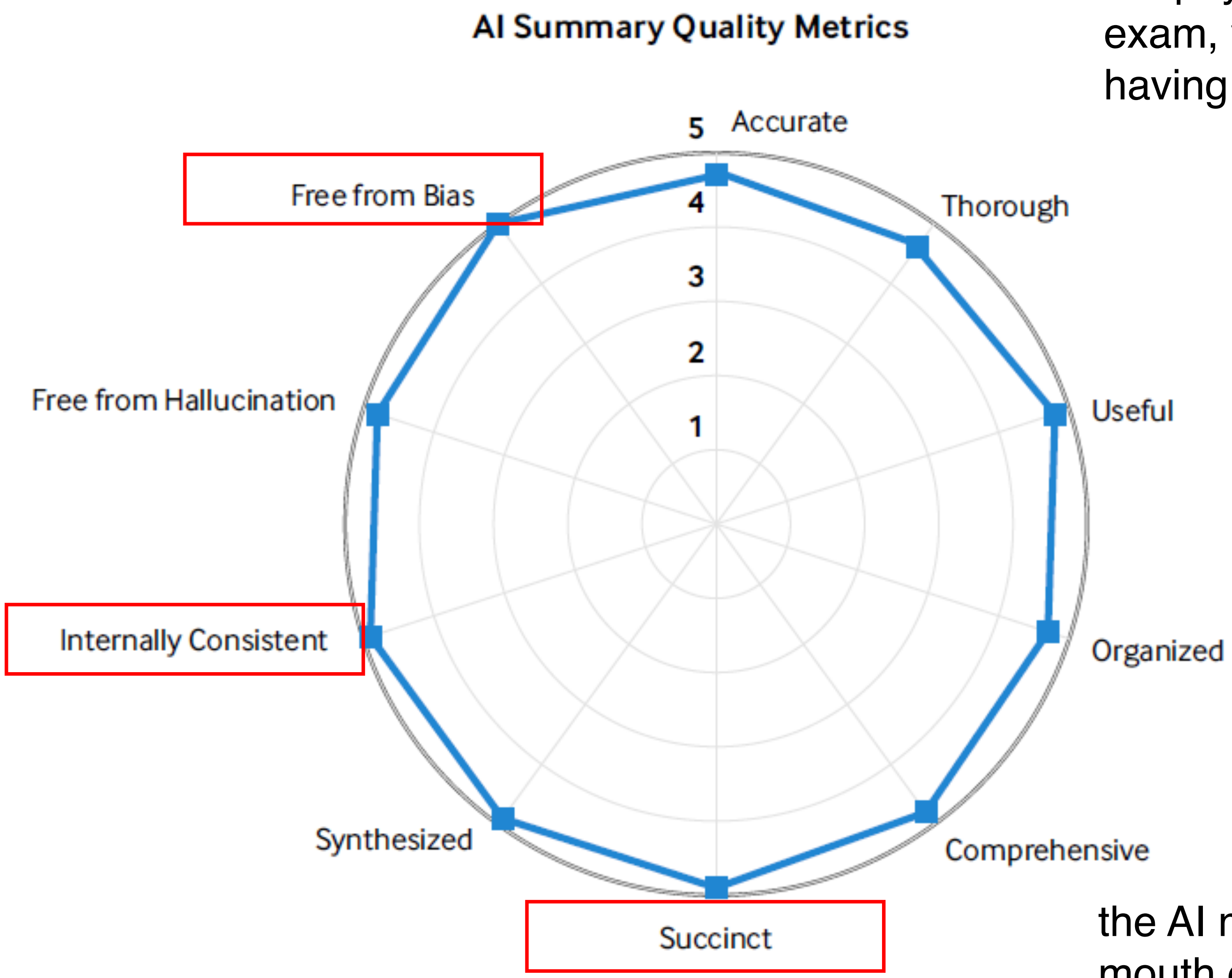
dose-response effect, with higher usage associated with more significant time reductions.

“more engaging and focused conversations with patients, enhancing the visit experience”

“that the scribe was especially helpful during lengthy appointments”

“game changer, it made notes more concise and improved the quality of visits”.

Feedback Category	Percentage
Patients who reported spending more time conversing with their physician	71%
Patients who reported spending less time conversing with their physician	1%
Patients who observed that their physician spent less time looking at the computer	81%
Patients who indicated that the AI scribe had no effect or improved the visit experience (negative effect)	100% (0%)
Patients who felt neutral to very comfortable with AI use in their care (felt uncomfortable)	100% (0%)



the physician mentioned the need to schedule a prostate exam, yet the AI summarized this as the exam already having been performed

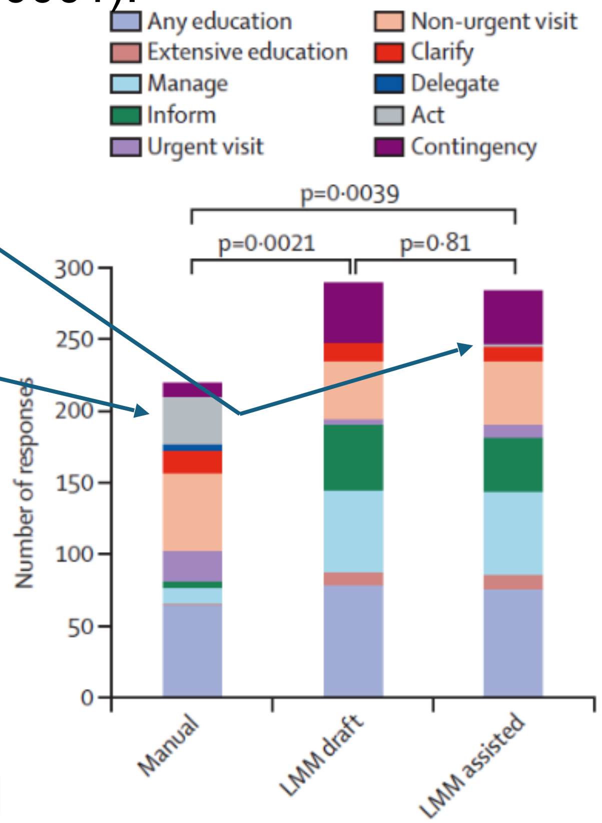
some summaries missed details, such as assessments for chest pain or anxiety.

the AI mistakenly inferred a diagnosis of hand, foot, and mouth disease when the physician had merely listed separate symptoms affecting the hand, feet, and mouth

Risk	Description	Example
Increased chart clutter	LLM note generation adds text volume, leading to need for summarization and more chart bloat.	Multiple team members create lengthy LLM notes. Covering physician requests LLM summary instead of reading all entries.
Decreased information density	LLMs generate verbose outputs that dilute essential clinical information.	A lengthy LLM-generated cardiology note lacks the focused insights of a concise staff cardiologist's note.
Persuasion and automation bias	LLMs may appear authoritative, causing clinicians to over-rely on their recommendations.	Primary team implements a tentative treatment plan directly from the LLM's confident tone without consulting with the original team.
Increased time to verify	Verifying and editing LLM-generated text adds to clinicians' workload.	Aware of confabulation risk, Physician spends extra time verifying a list of past medications generated by LLM to avoid redundant prescriptions.
Model collapse	LLMs trained on LLM-generated data risk "model collapse," losing insight and diversity in outputs.	An LLM trained on repetitive treatment data struggles to handle complex or rare cases due to limited exposure to varied clinical scenarios.

McCoy, L.G., A.K. Manrai, and A. Rodman, *Large Language Models and the Degradation of the Medical Record. New England Journal of Medicine, 2024.* ***391(17): p. 1561-1564.***

The mean manual response (34 words) was **shorter** than the LLM draft (169 words) and LLM-assisted responses (160 words; $p<0{\cdot}0001$).

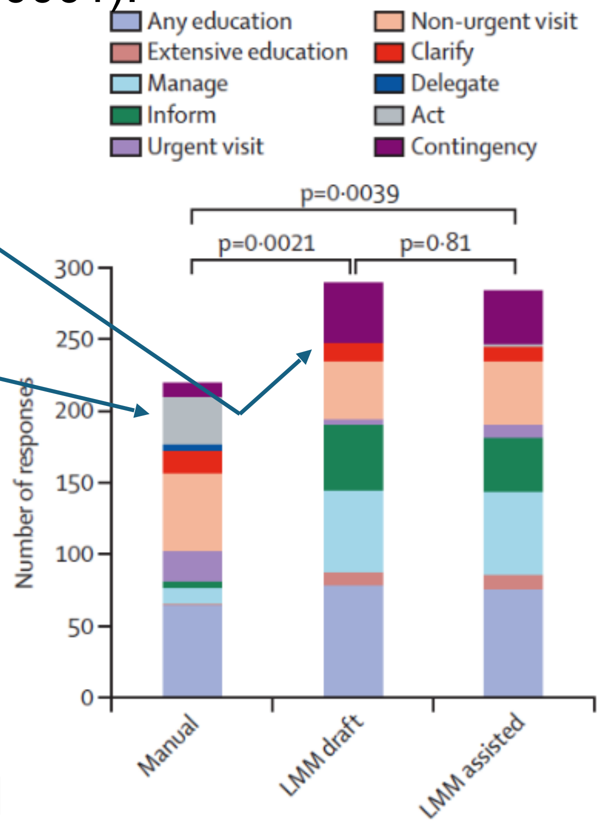


TH ATION SUMMIT

Risk	Description	Example
Increased chart clutter	LLM note generation adds text volume, leading to need for summarization and more chart bloat.	Multiple team members create lengthy LLM notes. Covering physician requests LLM summary instead of reading all entries.
Decreased information density	LLMs generate verbose outputs that dilute essential clinical information.	A lengthy LLM-generated cardiology note lacks the focused insights of a concise staff cardiologist's note.
Persuasion and automation bias	LLMs may appear authoritative, causing clinicians to over-rely on their recommendations.	Primary team implements a tentative treatment plan directly from the LLM's confident tone without consulting with the original team.
Increased time to verify	Verifying and editing LLM-generated text adds to clinicians' workload.	Aware of confabulation risk, Physician spends extra time verifying a list of past medications generated by LLM to avoid redundant prescriptions.
Model collapse	LLMs trained on LLM-generated data risk "model collapse," losing insight and diversity in outputs.	An LLM trained on repetitive treatment data struggles to handle complex or rare cases due to limited exposure to varied clinical scenarios.

McCoy, L.G., A.K. Manrai, and A. Rodman, *Large Language Models and the Degradation of the Medical Record. New England Journal of Medicine*, 2024. ***391(17): p. 1561-1564.***

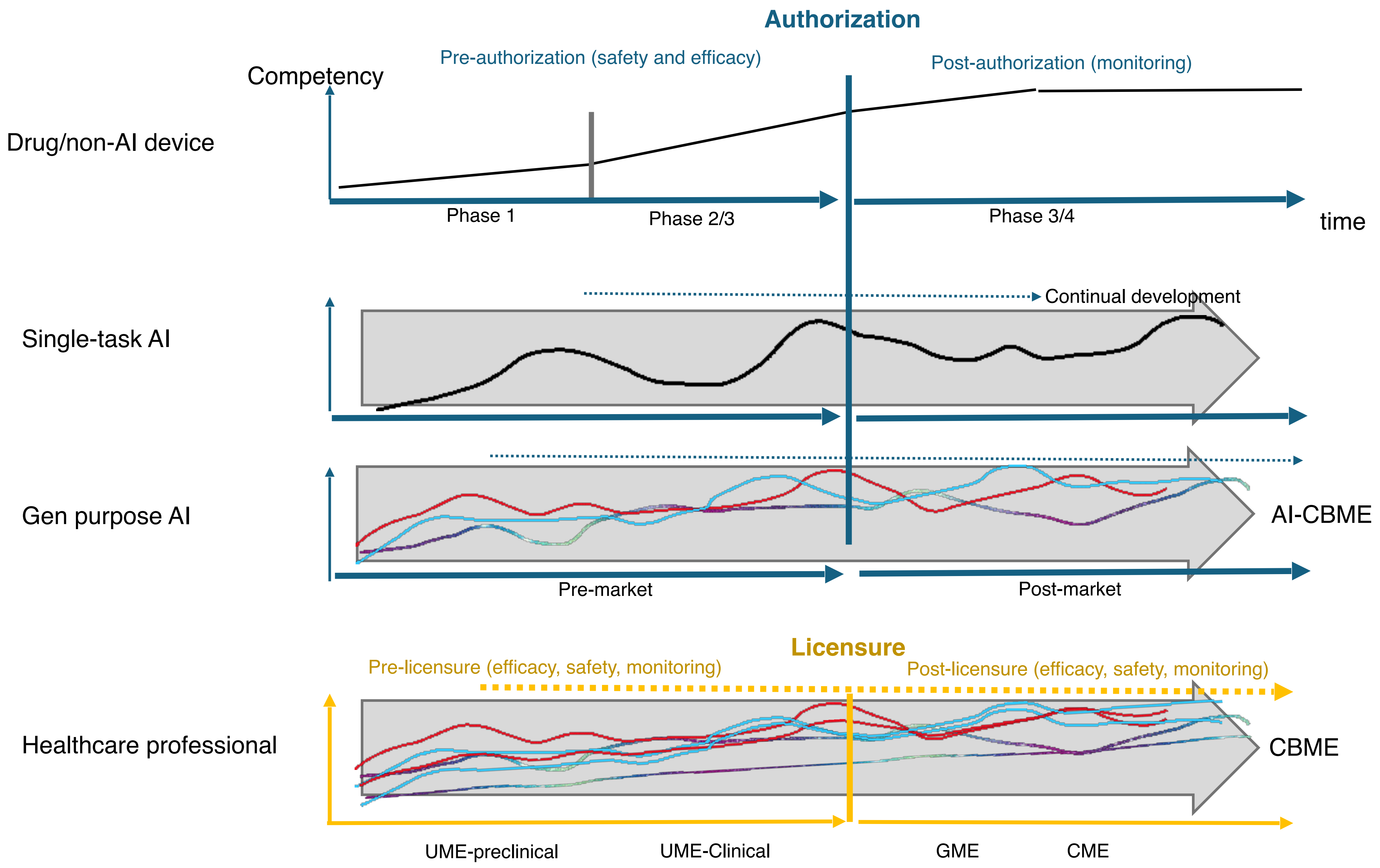
The mean manual response (34 words) was **shorter** than the LLM draft (169 words) and LLM-assisted responses (160 words; p<0·0001).



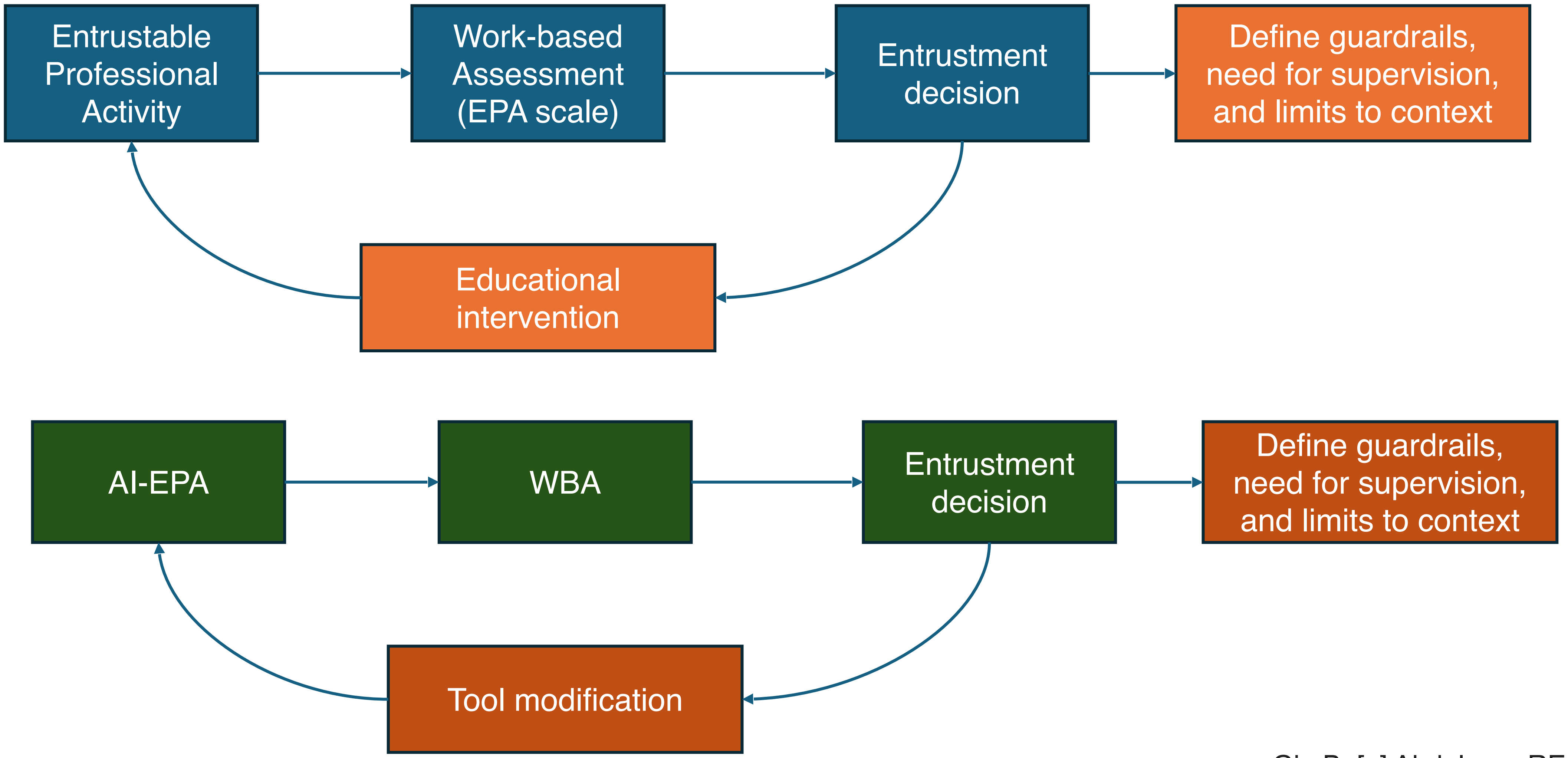
Benevolence: The Need For Regulation

- Societal, non-commercial entities must be involved in the development life cycle and regulation
- FDA
- Non-profit coalitions (e.g., CHAI)
  - FDA stepped away
- The case for AI-CBME: Continuous assessment of multimodal AI by stakeholder “educators”
  - GenAIs are not deterministic but stochastic
  - Real-world GenAIs can drift
  - GenAIs are general purpose and can be used for a wide variety of often unanticipated tasks
  - Large-scale GenAIs, such as large language models (LLMs), are typically kept opaque
  - GenAIs are not fully describable – “dark complexity” make them unpredictable

GPT-4 Performance, Nondeterminism, and Drift. NEJM AI. 2024

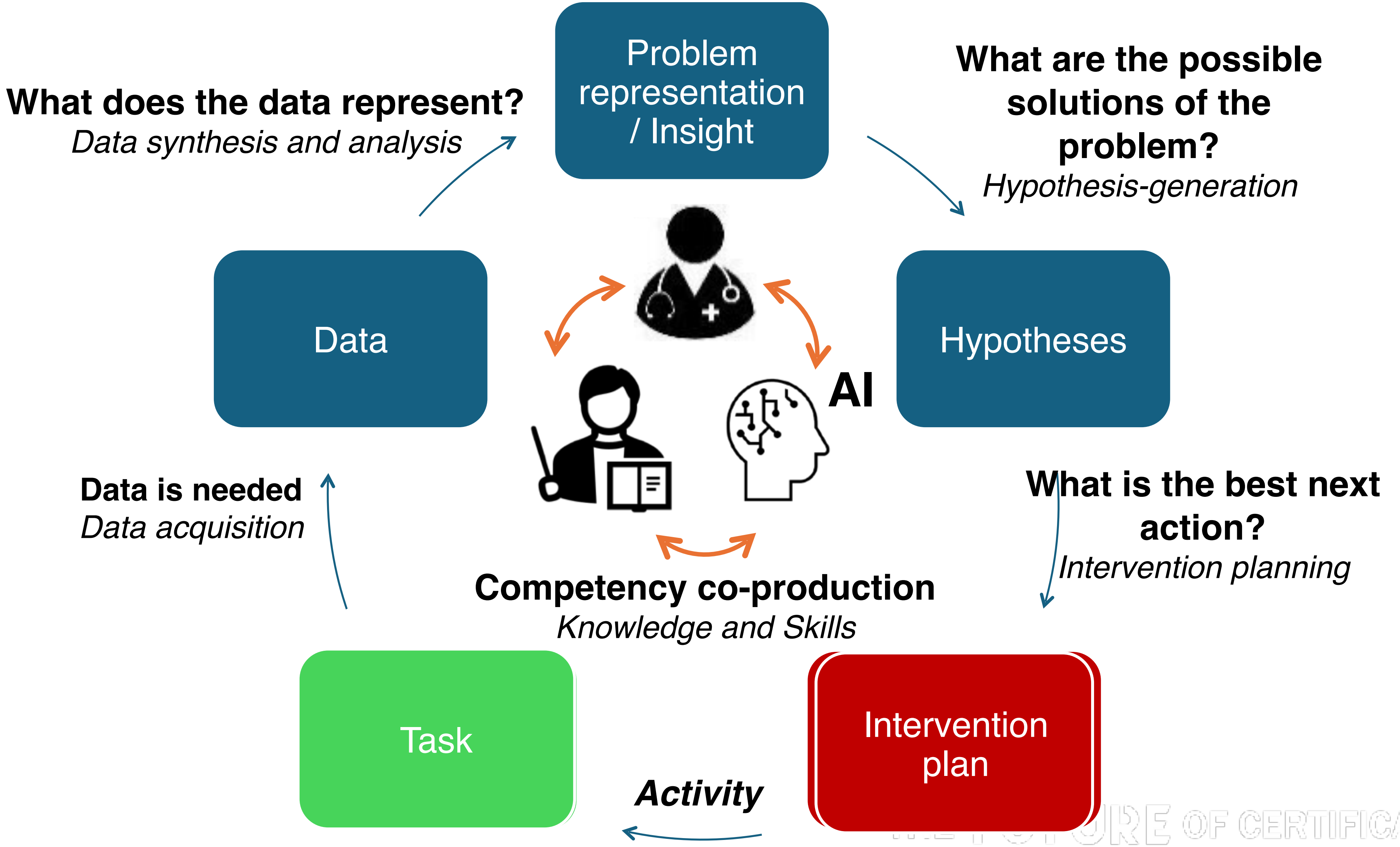

Authorization
Pre-authorization (safety and efficacy)
Post-authorization (monitoring)
Competency
Drug/non-AI device
Phase 1
Phase 2/3
Phase 3/4
time
Continual development
Single-task AI
Gen purpose AI
AI-CBME
Pre-market
Post-market
Licensure
Pre-licensure (efficacy, safety, monitoring)
Post-licensure (efficacy, safety, monitoring)
Healthcare professional
CBME
UME-preclinical
UME-Clinical
GME
CME

Entrustment Framework To Safeguard Use Of AI In Health Professions Education



Gin B, [..] Abdulnour RE, et al.
Acad Med, in press

Conclusion: Co-Production



Thank you!