AI & Medicine: Hype or Hope?

Matthew Thompson
Professor, UW Department of Family Medicine
Clinical Research Scientist, Google


BLACKFEET INDIAN RESERVATION
Kootenai National Forest
Flathead National Forest
Havre
FORT PECK INDIAN RESERVATION
Glasgow
Willist
Watford City
Coeur d'Alene
Great Falls
Glendive
Missoula
Helena-Lewis and Clark National Forest
MONTANA
Miles City
Butte
Billings
Hardin
Bozeman
Livingston
CROW RESERVATION
Custer Gallatin National Forest
Payette National Forest
Sheridan
Yellowstone National Park
Cody
Buffalo


Kirkwall
Elgin
Inverness
Portree
SCOTLAND
Aberdeen
HEBRIDES
Dundee
St Andrews
Stirling
Edinburgh
Glasgow
United Kingdom
Newcastle upon Tyne
NORTHERN IRELAND


BLACKFEET INDIAN RESERVATION
Kootenai National Forest
Flathead National Forest
Havre
FORT PECK INDIAN RESERVATION
Glasgow
Willist
Watford City
Alene
Great Falls
Glendive
Missoula
Helena-Lewis and Clark National Forest
MONTANA
Miles City
Butte
Billings
Hardin
Bozeman
Livingston
CROW RESERVATION
Custer Gallatin National Forest
Payette National Forest
Sheridan
Yellowstone National Park
Cody
Buffalo
Spearfish


Kootenai
National Forest
Flath
Nationa
'Alene
Missou
CE
ION
Payette
National Forest
Kirkwall
Elgin
Inverness
Portree
SCOTLAND
Aberdeen
HEBRIDES
Dundee
St Andrews
Stirling
Edinburgh
Glasgow
United
Kingdom
Newcastle
upon Tyne
NORTHERN
IRELAND
FORT PECK
INDIAN
RESERVATION
Glasgow
Willist
Watford City
Glendive
Miles City
ON
heridan
Buffalo
Spearfish

Disclosures: Currently employed by Google, with shares and stock options in Alphabet

Today’s presentation represents my personal views as a UW faculty member

AAFP survey of 1,200 FPs

February 24, 2025

AI Is in the Doctor's Bag—And Primary Care Is Ready to Use It

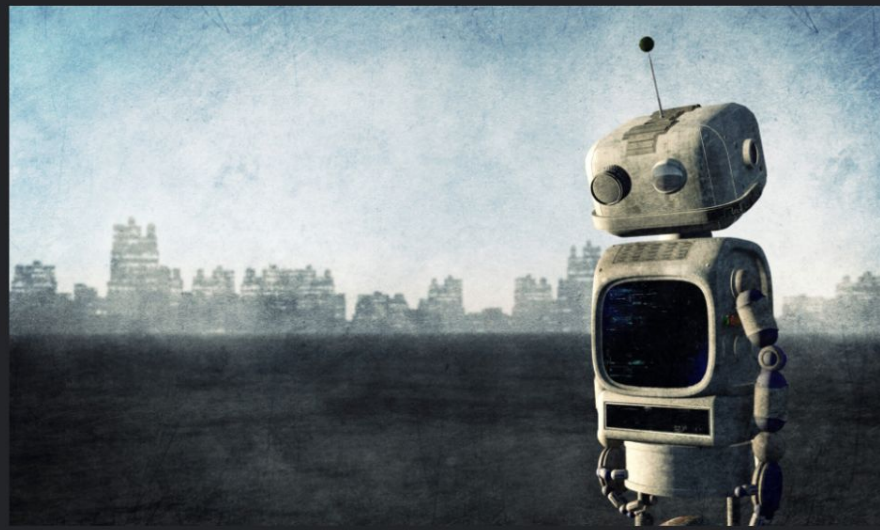
- **PCPs are *actively exploring* AI both personally and professionally:** Half of respondents reported having used AI tools for at least one use case at work. Some 62% said they use generative AI tools like ChatGPT outside of work; among those that haven't, 80% expressed a desire to do so.
- **PCPs are *optimistic about some aspects* of AI but skeptical about others:** ... improving time to diagnosis (73% expect positive impacts), diagnostic accuracy (66%), and appropriateness of treatment plans (66%). Most expect AI will improve their wellbeing (70%) and workload (66%).
- **Family physicians and other PCPs are *concerned* about AI in the workplace:** 18% fear that AI will have negative or very negative impacts on their job security, while another 52% are still unsure. 81% want more training to fully trust AI solutions at work.
- **Most PCPs have *limited influence* over the AI tools available in their practices.**

GLASS HALF FULL OR EMPTY?

Most Americans think AI won't improve their lives, survey says

Rare survey of AI experts exposes deep divide with public opinion.

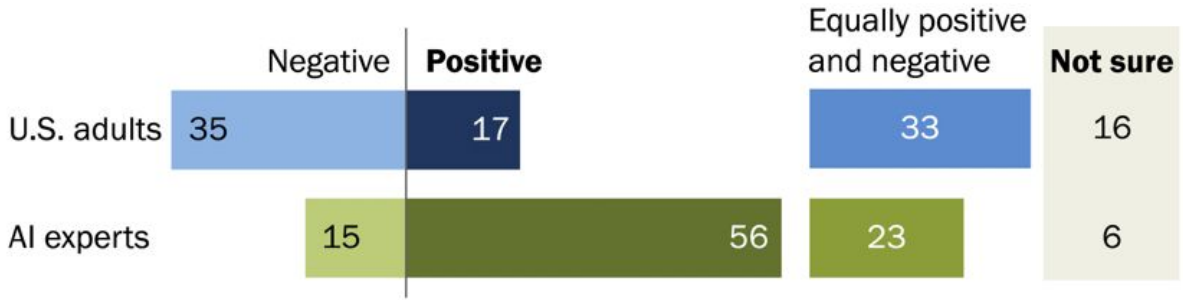
ASHLEY BELANGER – APR 3, 2025 11:18 AM | 238



Credit: piranka | E+

AI experts more likely than the public to say AI will have a positive effect on the U.S. over next 20 years

% who say they think the impact of artificial intelligence (AI) on the U.S. over the next 20 years will be ...



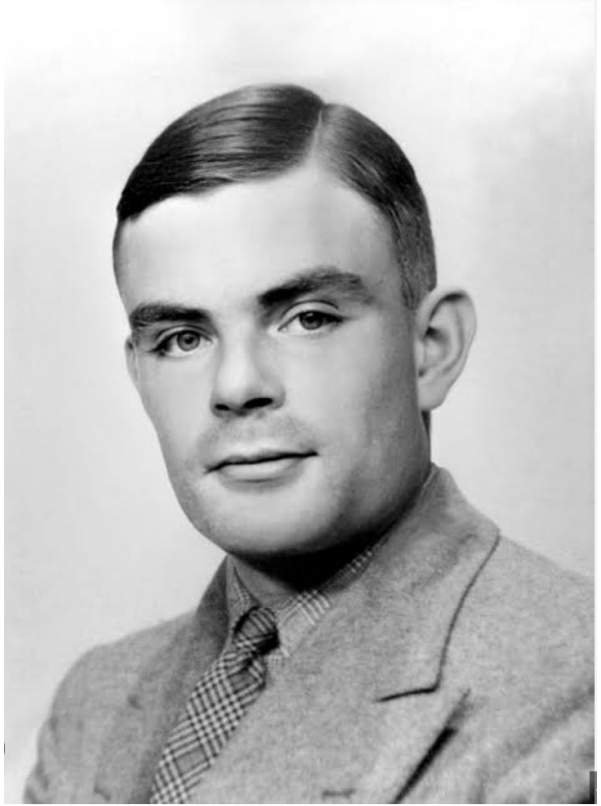
Note: "AI experts" refer to individuals whose work or research relates to AI. The AI experts surveyed are those who were authors or presenters at an AI-related conference in 2023 or 2024 and live in the U.S. Expert views are only representative of those who responded. For more details, refer to the methodology. "Very/somewhat positive" and "very/somewhat negative" are combined. Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Aug. 12-18, 2024. Survey of AI experts conducted Aug. 14-Oct. 31, 2024.
"How the U.S. Public and AI Experts View Artificial Intelligence"

PEW RESEARCH CENTER

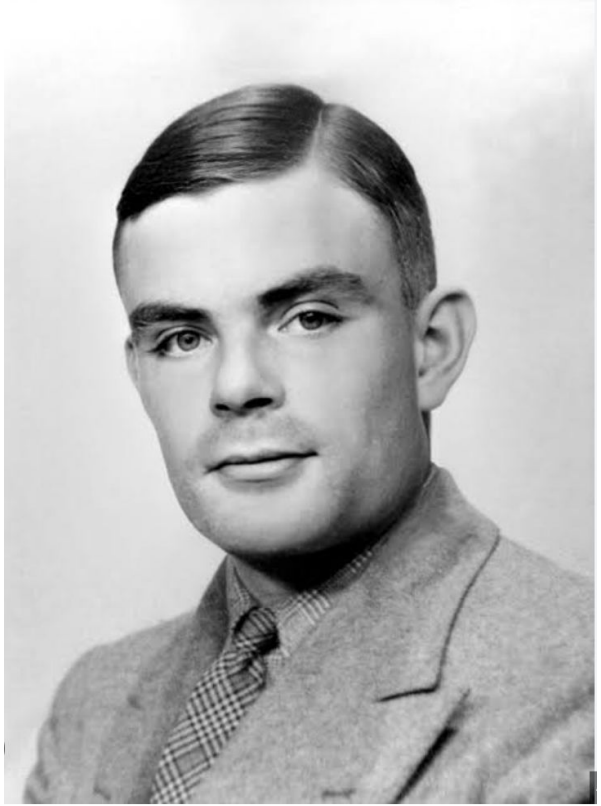
Take home points

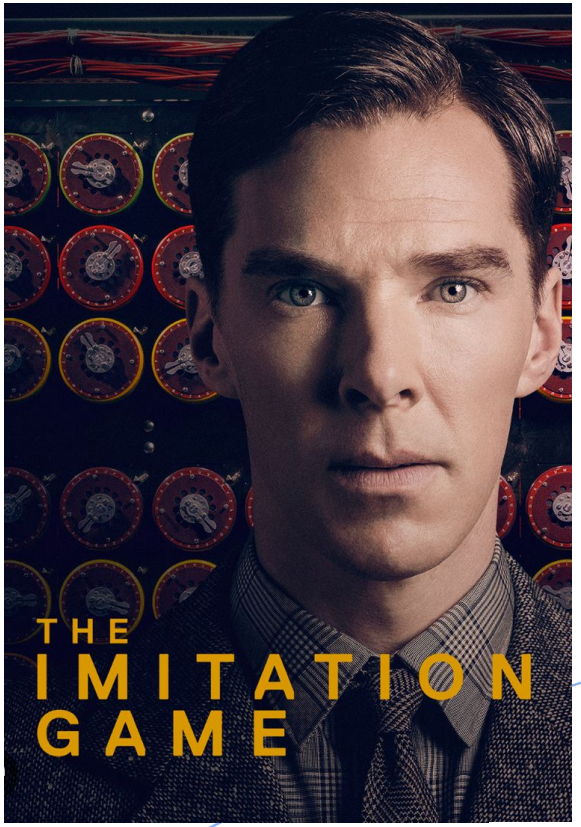
- Understanding AI and Gen AI is ***difficult*** for most of us
- Some tasks can likely be ***safely shifted now***
- ***Real impact*** of Gen AI tools on many areas of health care?
- Clinicians/educators who use AI or Gen AI as ***part of their work*** will be the norm
- ***Gulf*** between AI capabilities & clinician/educator capacity to evaluate, implement
- Multiple ***human attributes*** developed over millions of years are not replaceable (*and nor should we try to*)--- understanding, joy, fear, empathy, connection, touch, kindness, love…

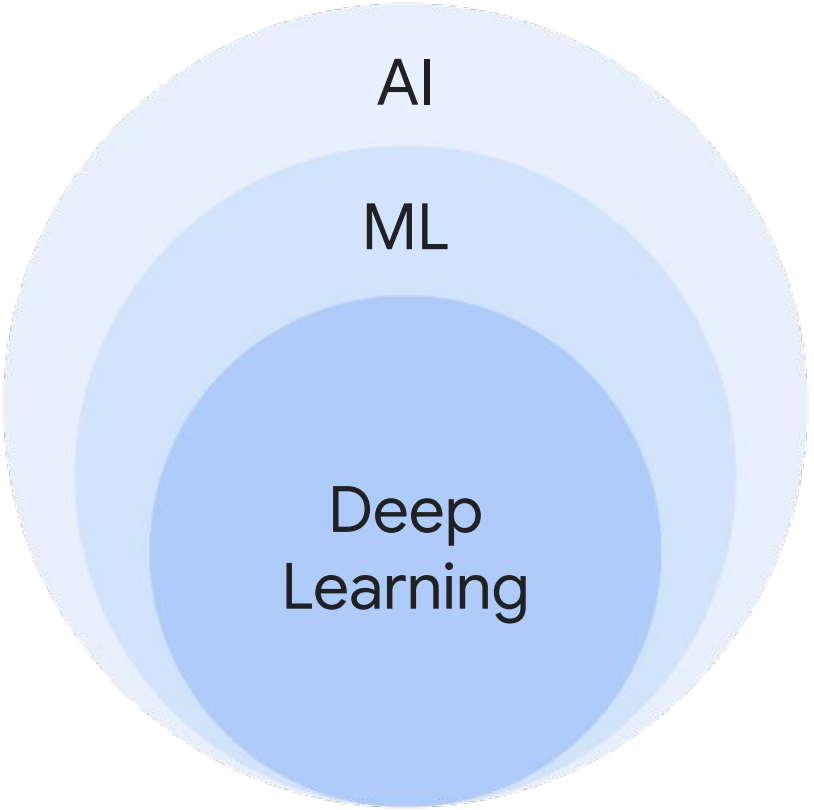
AI & Generative AI



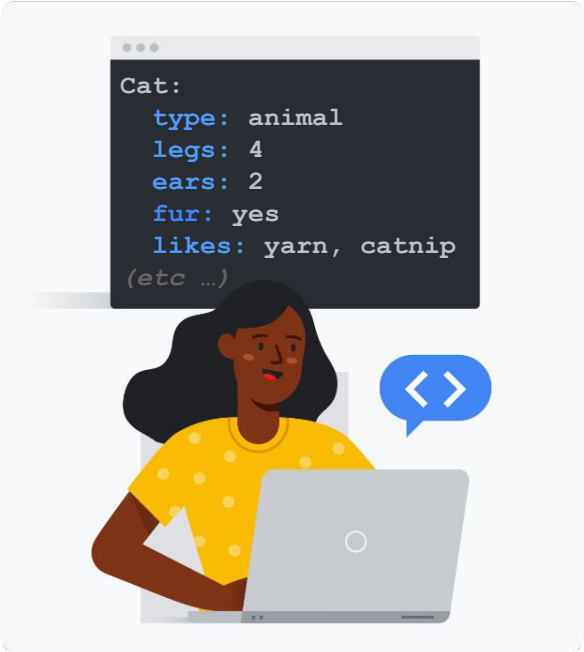
Google Health



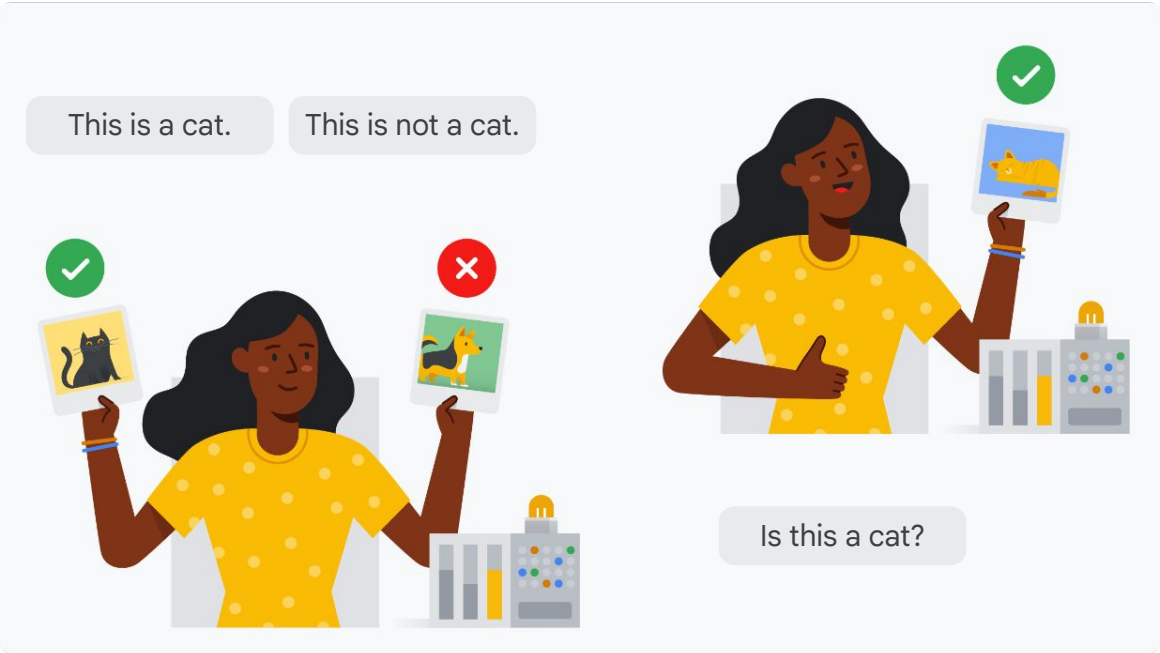

THE
IMITATION
GAME


AI
ML
Deep
Learning

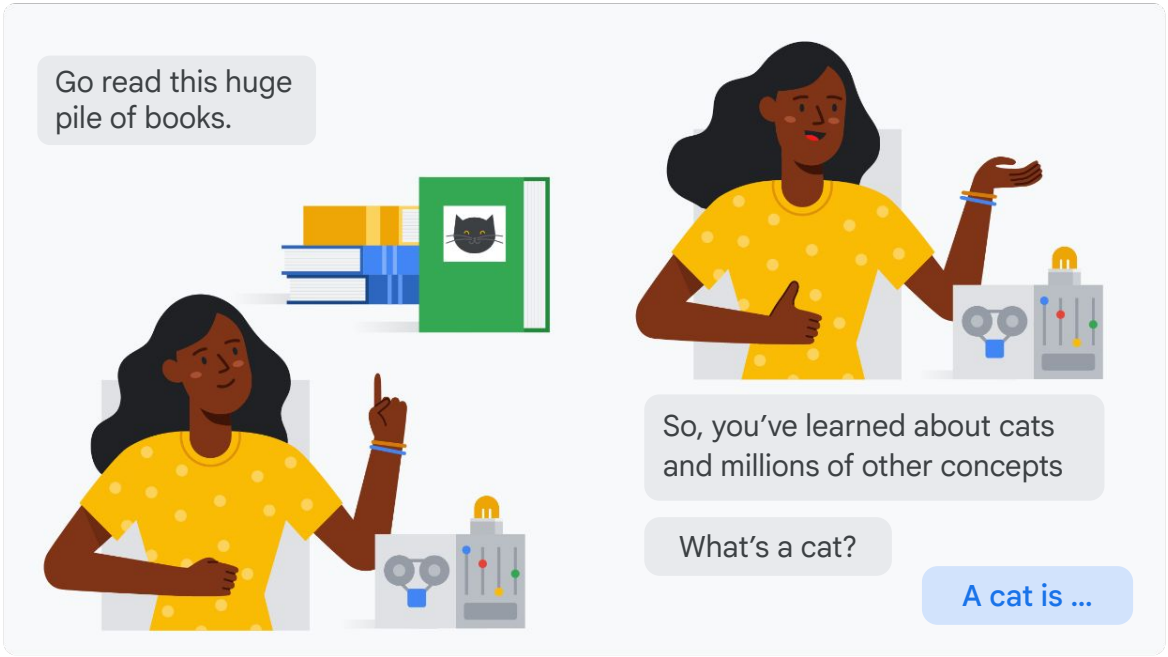
Traditional programming



Wave of neural networks | ~2012



Generative language models | LaMDA, PaLM, GPT, Gemini, etc.



We're in a new era of generative AI

Generalization
Many tasks at the same time
Training using free text
Multimodal and sequential
Rich expressive outputs

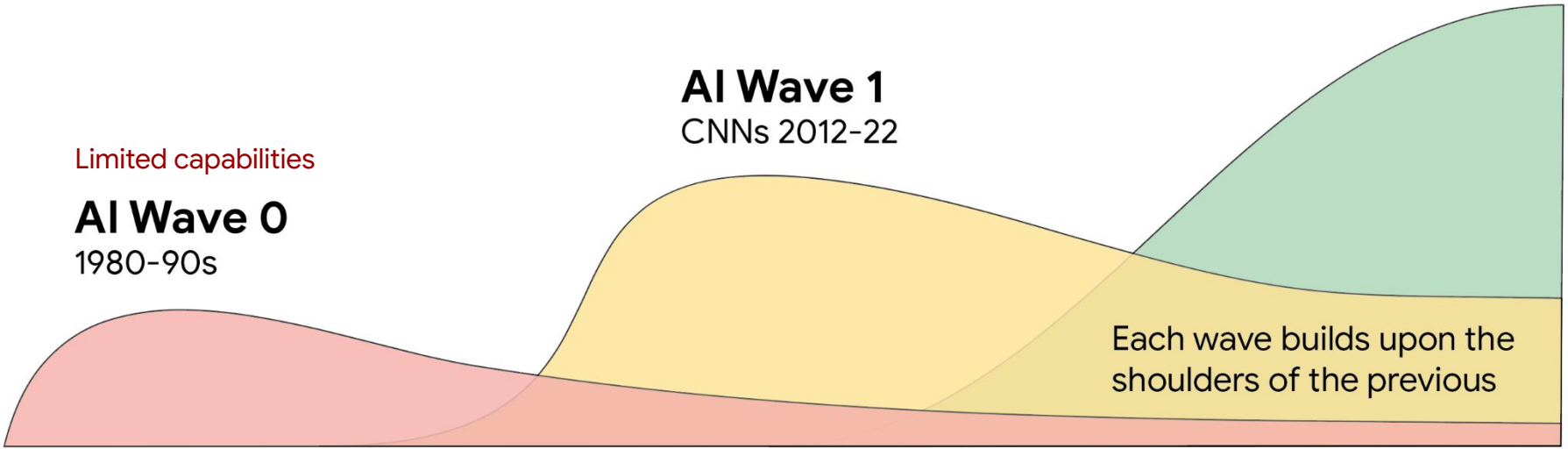
AI Wave 2
GenAI 2022+

Task specific= good at narrow tasks
Predicts or classify things
Expensive training data
Poor multimodal / sequential

AI Wave 1
CNNs 2012-22

Limited capabilities

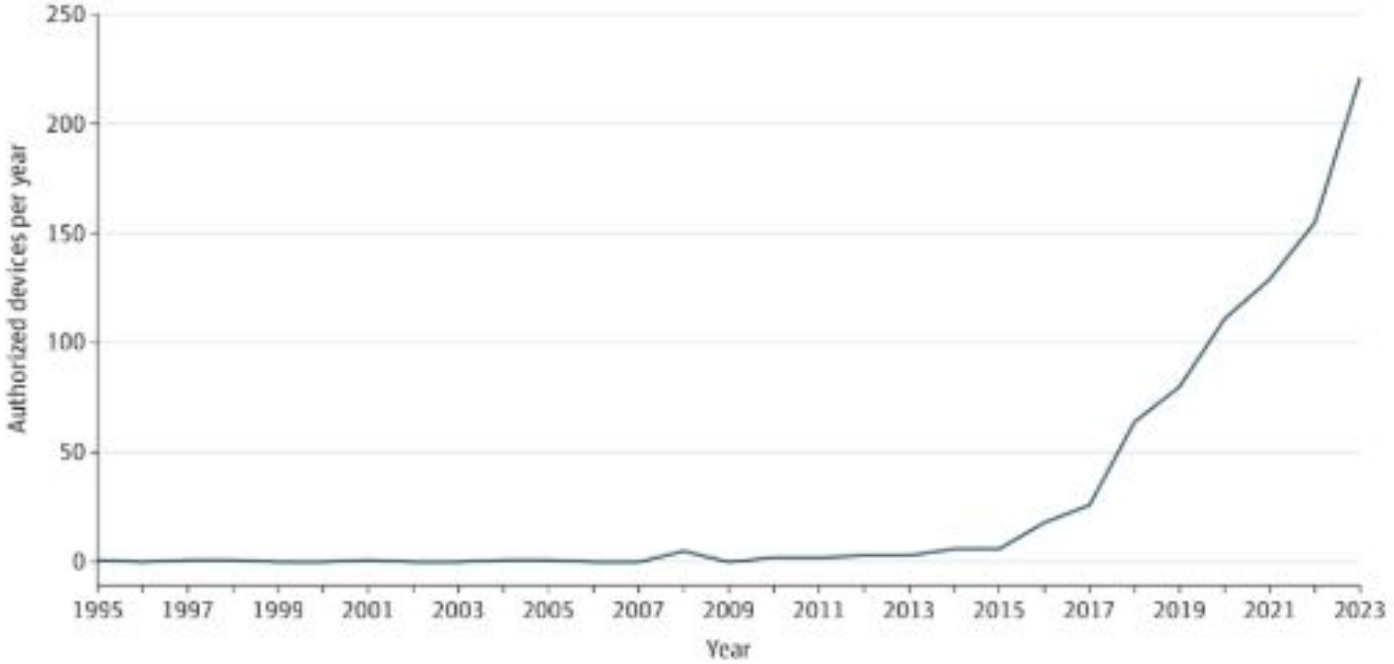
AI Wave 0
1980-90s



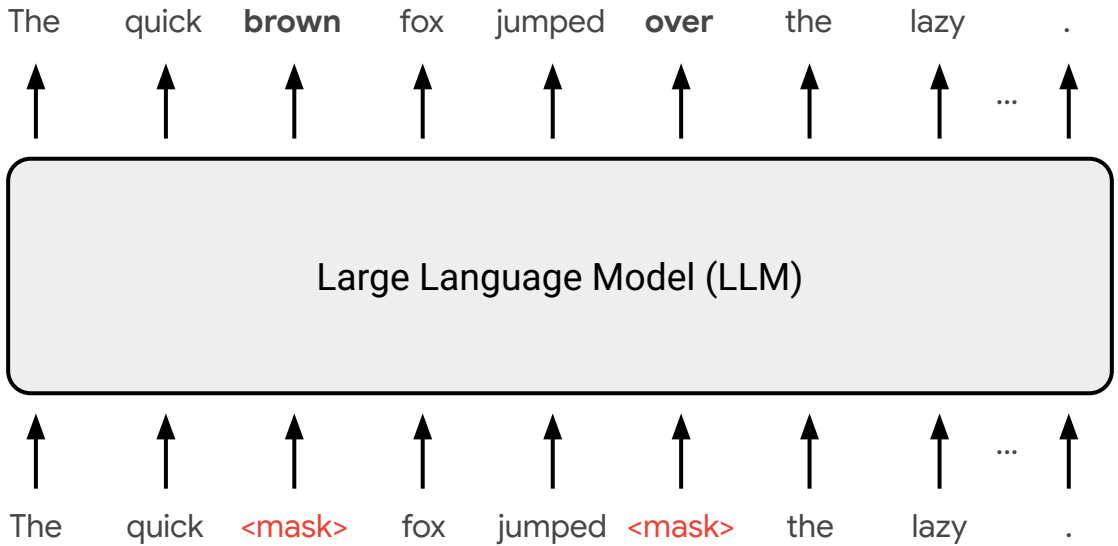
JAMA 2024; 331:242

FDA has authorized approx 1000 AI-enabled medical devices

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



Large Language Models = “next-word prediction engines”

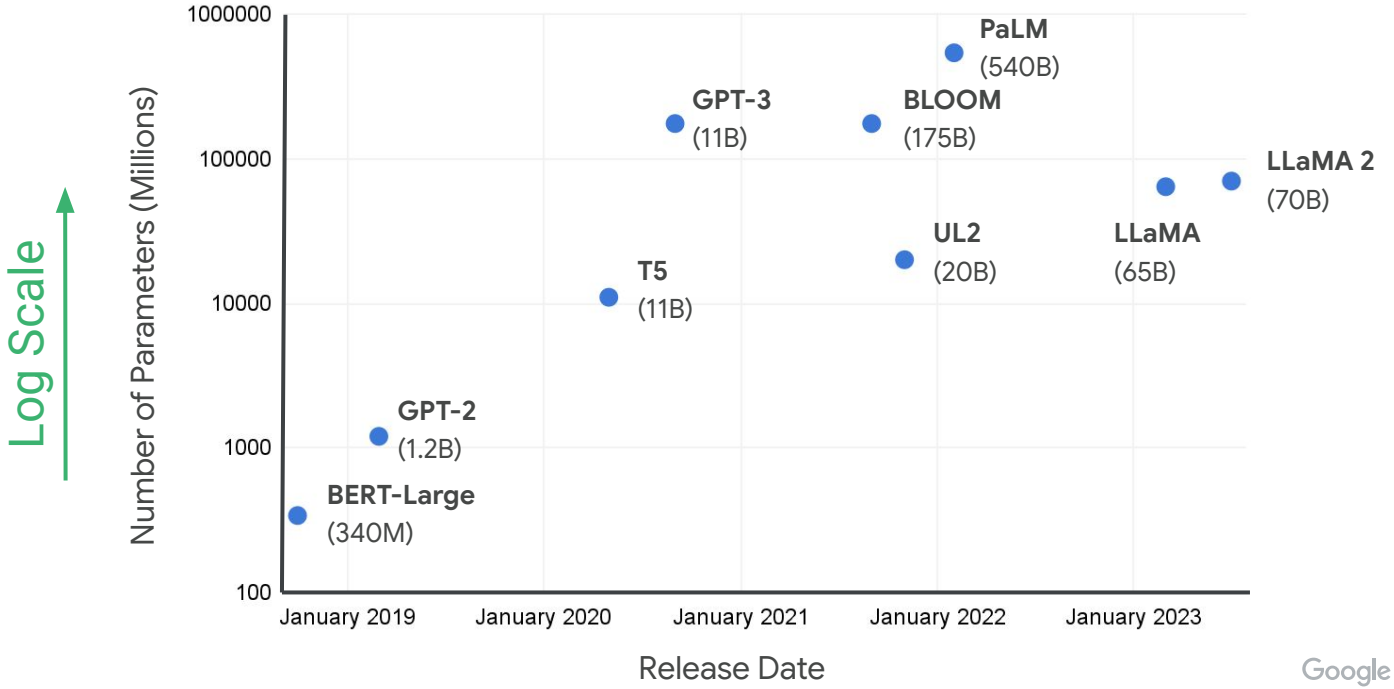


LLM typically created using 3 step process:

1. Pre training with vast web-scale data to create a base model
2. Fine tuning, using question and answer data to improve responses.
3. Reinforcement learning from human feedback to align model with human values and improve response quality

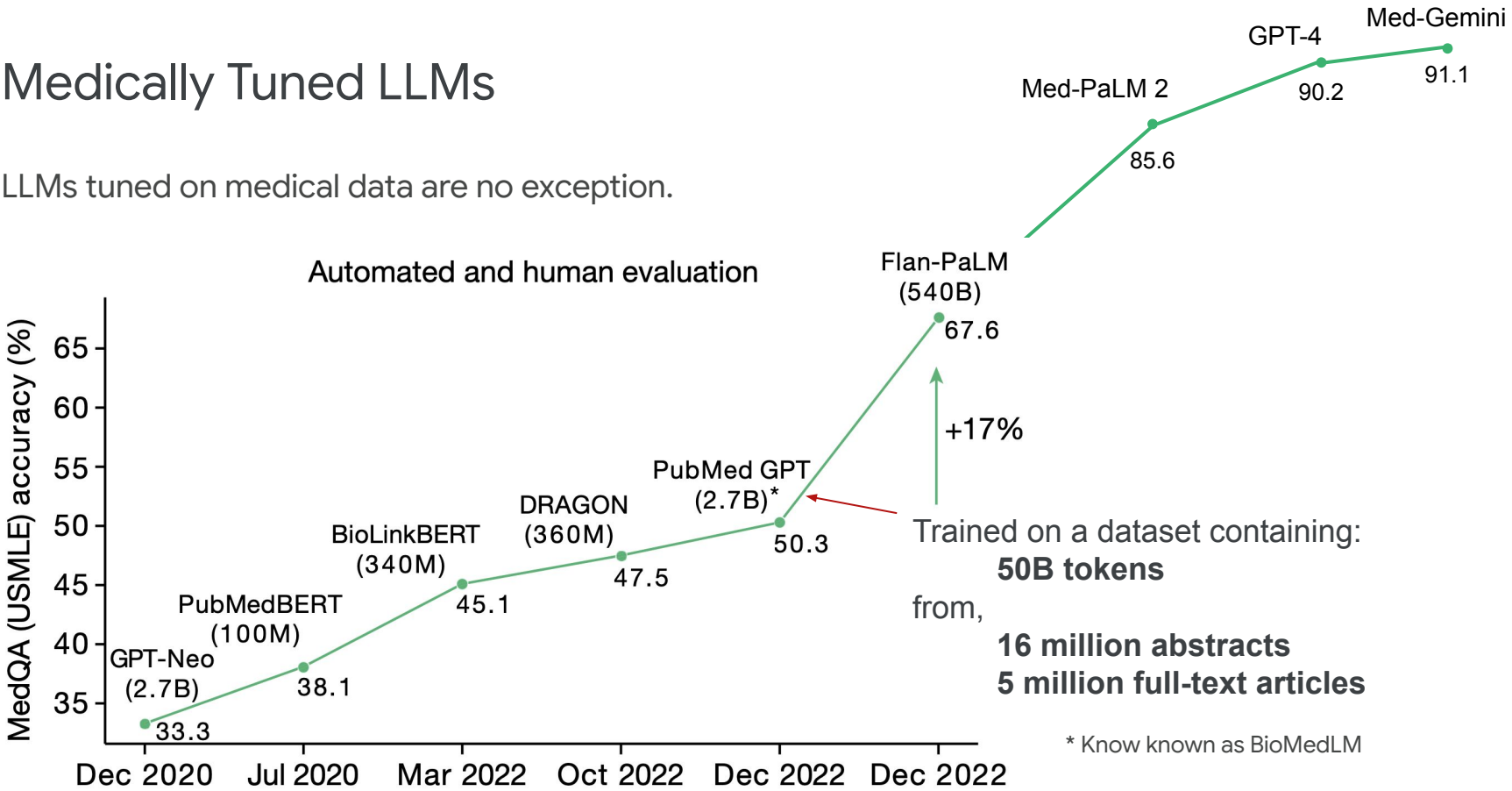
“Bigger is Better”

Language models have been growing in size exponentially in recent years.



Medically Tuned LLMs

LLMs tuned on medical data are no exception.

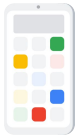


Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.

Gen AI poses novel but addressable risks for the Healthcare and Life Sciences industries...

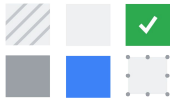
Hallucinations

Gen AI can generate misinformation with high degree of certainty for random questions leading users to make poor decisions



Encoding Bias

Bias in the training data can get baked into the model leading to biased outputs when used in various use cases



Omissions

When summarizing documents such as medical records or literature, Gen AI can leave out important information affecting decisions

Security

Gen AI can generate misinformation with high degree of certainty for random questions leading users to make poor decisions

Public Scrutiny

AI adoption is built on trust amongst all stakeholders, and this trust can quickly erode with a few errors in decision making and safety

Privacy

PHI within the training data can be revealed inadvertently especially when asked very specific questions

… that can be mitigated by guardrails, governance, and processes

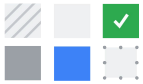
Sample, non-exhaustive mitigation strategies

Hallucinations

- Keep a human in the loop
- Conduct adversarial testing to identify problematic prompts
- Monitor end user behavior to detect consistent issues

Encoding Bias

- Finetune on balanced data sets
- Use adversarial testing and red teaming on end to end products to identify inequities
- Evaluate foundational models for equity considerations before building a solution



Omissions

- Keep a human in the loop
- Provide end user training to raise awareness and teach employees to interrogate results for blinds spots
- Monitor end user behavior to detect common omissions

Security

- Provide all employees with Gen AI awareness training similar to password best practice campaigns
- Build tools to screen end user entered prompts for malicious commands
- Follow SecOps best practices to quickly adapt to emerging threats

Public Scrutiny

- Create a clear communication strategy for any issues
- Require the model to save reasoning for decision making use cases
- Provide clear disclosures
- Proactively monitor and address issues as they arise

Privacy

- Provide clear disclosure on data storage and use
- Work with vendors to understand their data storage practices
- Remove PHI from all training data
- Conduct adversarial testing prior to deploying a solution

Agentic AI

Digital medicine

The rise of agentic AI teammates in medicine

AI = passive, reactive, depends on human users to provide input and context

Agentic AI = can take initiatives, proactive, maintain memory and context, track patient data, identify issues, propose solutions.

Future = AI Clinical manager agent = orchestrate several tools (eg transcribing, searching clinical guidelines, analysing images), clinician would deal with a single manager. Supervised by human clinician

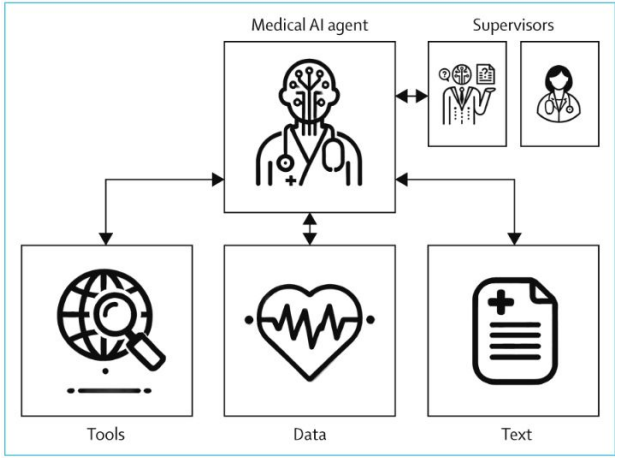


Figure: **Medical AI agent orchestrating tools and data while monitored by human clinician and AI reviewer**

Multimodal

https://www.youtube.com/watch?v=nXVvvRhiGjI

Pain point #1: Documentation

Pajama Time: Working After Work in the Electronic Health Record

Harry S. Saag, MD[1,2,3], Kanan Shah, BS[2], Simon A. Jones, PhD[2,4], Paul A. Testa, MD, JD[5], and Leora I. Horwitz, MD, MHS[1,2,4]

Automatically docu
care with the Drag
Ambient eXperienc
The exam of the future has arrived with clir
documentation that writes itself.

Ambience

Product Specialties EHR Integrations Careers Blog

Book a demo >

A comprehensive ambient AI platform

Streamline every step of patient care with AI features designed to enhance specialty workflows, improve documentation accuracy, and optimize billing—all in real-time and integrated directly inside the EHR.

Real-Time Note Generation
Automatically generates clinical notes during patient encounters, tailored to the needs of each specialty.

Integrated Coding Assistance
Provides real-time coding suggestions (ICD-10, E&M) within the clinical workflow, ensuring accurate billing and compliance.

After-Visit Summaries
Generates clear after-visit summaries, ensuring patients understand their care and follow-up instructions.

Referral Letters
Automatically generates referral letters for seamless handoffs between physicians and specialists.

COMING SOON

Pre-Charting
Provides clinicians with pre-built charts, reducing preparation time for upcoming appointments.

Product overview >

Sunoh.ai

Reclaim Your Time with AI-
Ambient Listening Techr

Reduce Burnout, Enhance Patient Care

Burdened by endless documentation? Sunoh.ai automates and simplifies clinical documentation using breakthrough, AI-powered ambient listening technology. The administrative workload is reduced, staff burnout is minimized, and patient care is improved. Set your practice up for success and allow Sunoh.ai to transform your practice today.

Schedule a Demo

Published on 22.04.2024 in Vol 26 (2024)

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/54419, first published November 09, 2023.

Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study

Annessa Kernberg[1]; Jeffrey A Gold[1]; Vishnu Mohan[1]

Article Authors Cited by (14) Tweetations (9) Metrics

- Abstract

Abstract

Where's the evidence (or, do we *need* any evidence?)

JAMA Network | Open

Research Letter | Health Informatics

AI-Powered Clinical Documentation and Clinicians' Electronic Health Record Experience

A Nonrandomized Clinical Trial

Tsai-Ling Liu, PhD; Timothy C. Hetherington, MS; Casey Stephens, MPH; Andrew McWilliams, MD, MPH; Ajay Dharod, MD; Tracey Carroll, MHA, MBA; Jeffrey A. Cleveland, MD

Recruited FM, IM, Peds from outpatient settings NC, GA within Atrium Health
Non-randomized, stepped wedge design

- Intervention =85 clinicians had 1 hour training on Dragon DAX
- Comparison group = 55 clinicians

140 respondents, 57% female, 39% had 5-15 yr, 37% had 15-25 years experience

Outcome - survey of EHR use, 5 wks before & after implementation

	Intervention	Control	
Decreased time on the EHR at home	40/85 (47.1%)	8/55 (14.5%)	$P < .001$
Decreased weekly time on the EHR outside normal work hours	38/85 (44.7%)	11/55 (20.0%)	$P = .003$
Decreased time on documentation after visit	37/85 (43.5%)	10/55 (18.2%)	$P = .002$
EHR experiences were comparable before and after the intervention	Mean 44.7% (SD 1.7)	Mean 68.7% (SD 2.3)	

Kaiser Permanente - in 2024 implemented **AI based clinical documentation tool (Abridge)** to 40 hospitals and 600 medical offices. Response to major priority from clinicians

The impact of nuance DAX ambient listening AI documentation: a cohort study J Am Med Inform Assoc.

2024 Apr 3;31(4):975-979. doi: 10.1093/jamia/ocae022.

Results: A total of 99 providers representing 12 specialties enrolled in the study; 76 matched control group providers were included for analysis.: Nuance DAX use showed positive trends in provider engagement at no risk to patient safety, experience, or clinical documentation. There were **no significant benefits to patient experience, documentation, or measures of provider productivity.**

The association between use of ambient voice technology documentation during primary care patient encounters, documentation burden, and provider burnout. Owens et al Fam Practi 2024.

Observational study of 83 PCPs **significantly reduced documentation burden** and primary care provider disengagement but not total provider burnout scores.

AI integration in nephrology: evaluating ChatGPT for accurate ICD-10 documentation and coding

Yasir Abdelgadir[1], Charat Thongprayoon[1], Jing Miao[1], Supawadee Suppadungsuk[1,2], Justin H. Pham[3], Michael A. Mao[4], Iasmina M. Craici[1] and Wisit Cheungpasitporn[1]*

[1]Division of Nephrology and Hypertension, Mayo Clinic, Rochester, MN, United States, [2]Chakri Naruebodindra Medical Institute, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Samut Prakan, Thailand, [3]Mayo Clinic College of Medicine and Science, Mayo Clinic, Rochester, MN, United States, [4]Division of Nephrology and Hypertension, Department of Medicine, Mayo Clinic, Jacksonville, FL, United States

Ambient note generation

- Are potential benefits in Dr-patient relationship and reduced keyboard time worth the risks (e.g. errors, patient concerns etc)?
- What is the burden of proof for ‘low but not no risk’ Gen AI tools like this?

Pain point #2: After visits - lab results, responding to eCare messages

Original Investigation FREE

April 28, 2023

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD, MA[1,2]; Adam Poliak, PhD[3]; Mark Dredze, PhD[4]; et al

Author Affiliations | Article Information

JAMA Intern Med. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

Compared the actual physician response vs Chat GPT chatbot to 195 questions posted on a public social media forum (Reddit)

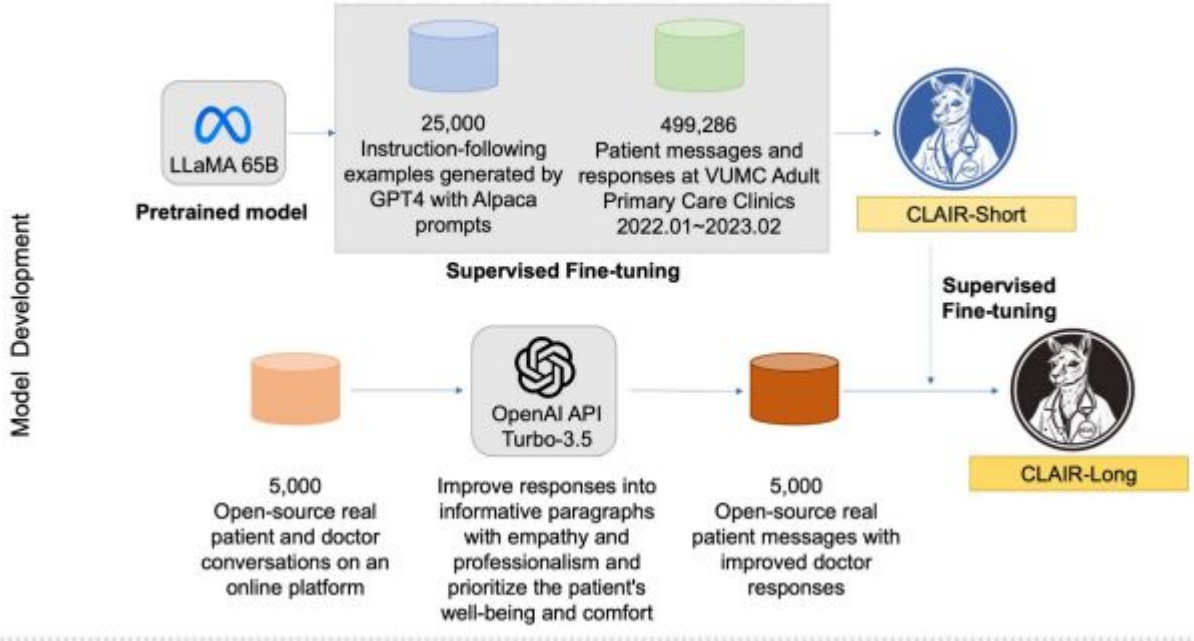
Responses of physicians & chatbot evaluated blindly by 3 health care professionals

Results

- Evaluators ***preferred chatbot responses*** over physicians for 78%
- Chatbot responses ***higher quality, more empathetic, longer***

Leveraging large language models for generating responses to patient messages—a subjective analysis

Siru Liu, PhD*[,1], Allison B. McCoy, PhD[1], Aileen P. Wright, MD, MS[1,2], Babatunde Carew, MD[3], Julian Z. Genkins, MD[4], Sean S. Huang, MD[1,2], Josh F. Peterson, MD, MPH[1,2], Bryan Steitz, PhD[1], Adam Wright, PhD[1]



Evaluation

Patient messages and responses at VUMC Adult Primary Care Clinics 2023.03.01~2023.03.07

10 Representative Questions in Primary Care (Rephrased)

CLAIR-Short

CLAIR-Long

ChatGPT (GPT-3.5)

ChatGPT (GPT-4)

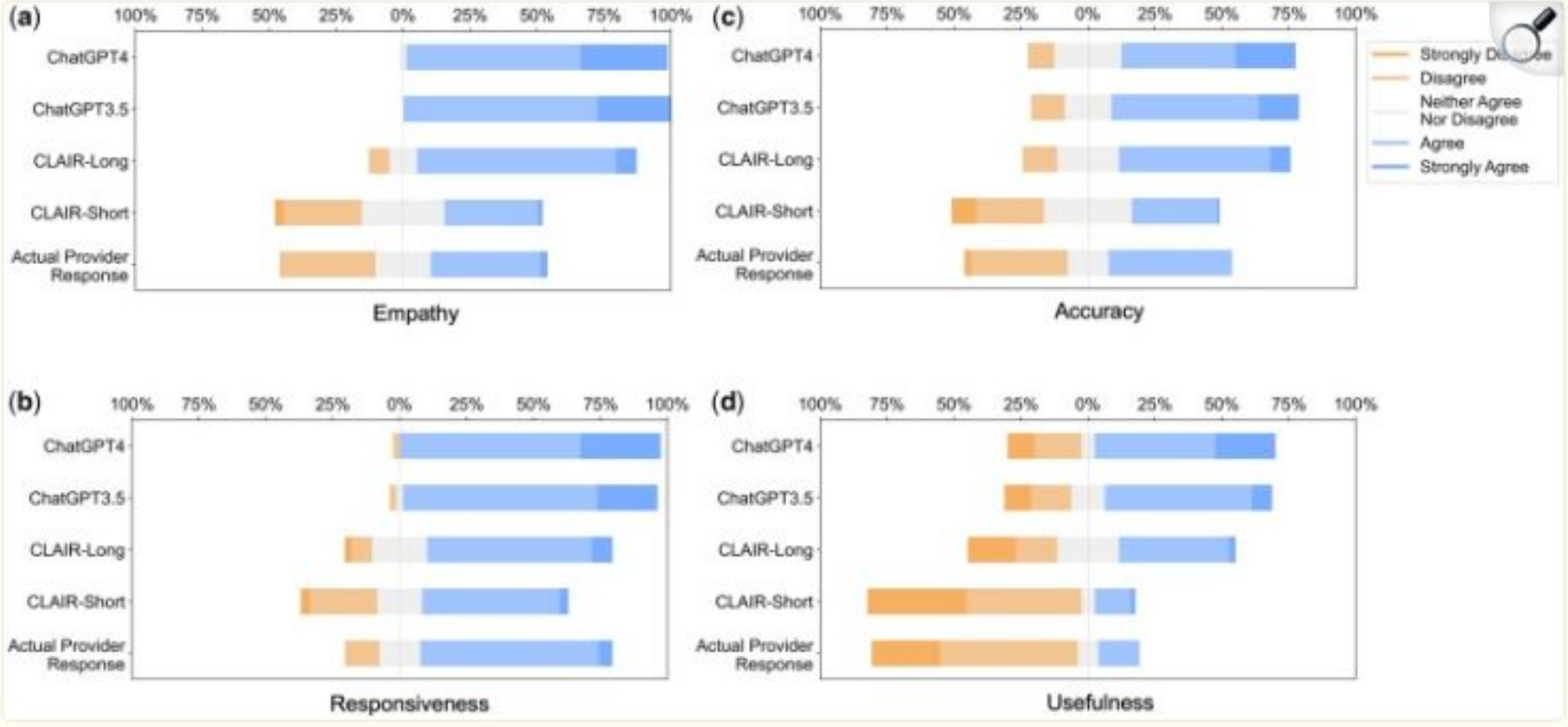
Generated Responses + Healthcare Provider Response

4 primary care physicians rated a mix of generated responses with the actual provider responses

- Empathy
- Responsiveness
- Accuracy
- Usefulness

Prompt:
"Imagine that you are a primary care doctor, and you have received a message from your patient. Your task is to reply the patient's message with polite and informative paragraphs, providing helpful guidance or next steps for the patient to take, offering patient education. Be sure to approach the message with empathy and professionalism, prioritizing the patient's well-being and comfort throughout your response. Remember that you are this patient's primary care doctor, and your goal is to provide your patient with the best possible care and support."

Figure 3.


(a)
100% 75% 50% 25% 0% 25% 50% 75% 100%
ChatGPT4
ChatGPT3.5
CLAIR-Long
CLAIR-Short
Actual Provider Response
Empathy
(c)
100% 75% 50% 25% 0% 25% 50% 75% 100%
ChatGPT4
ChatGPT3.5
CLAIR-Long
CLAIR-Short
Actual Provider Response
Accuracy
Strongly Disagree
Disagree
Neither Agree Nor Disagree
Agree
Strongly Agree
(b)
100% 75% 50% 25% 0% 25% 50% 75% 100%
ChatGPT4
ChatGPT3.5
CLAIR-Long
CLAIR-Short
Actual Provider Response
Responsiveness
(d)
100% 75% 50% 25% 0% 25% 50% 75% 100%
ChatGPT4
ChatGPT3.5
CLAIR-Long
CLAIR-Short
Actual Provider Response
Usefulness

Responding to patient messages etc:

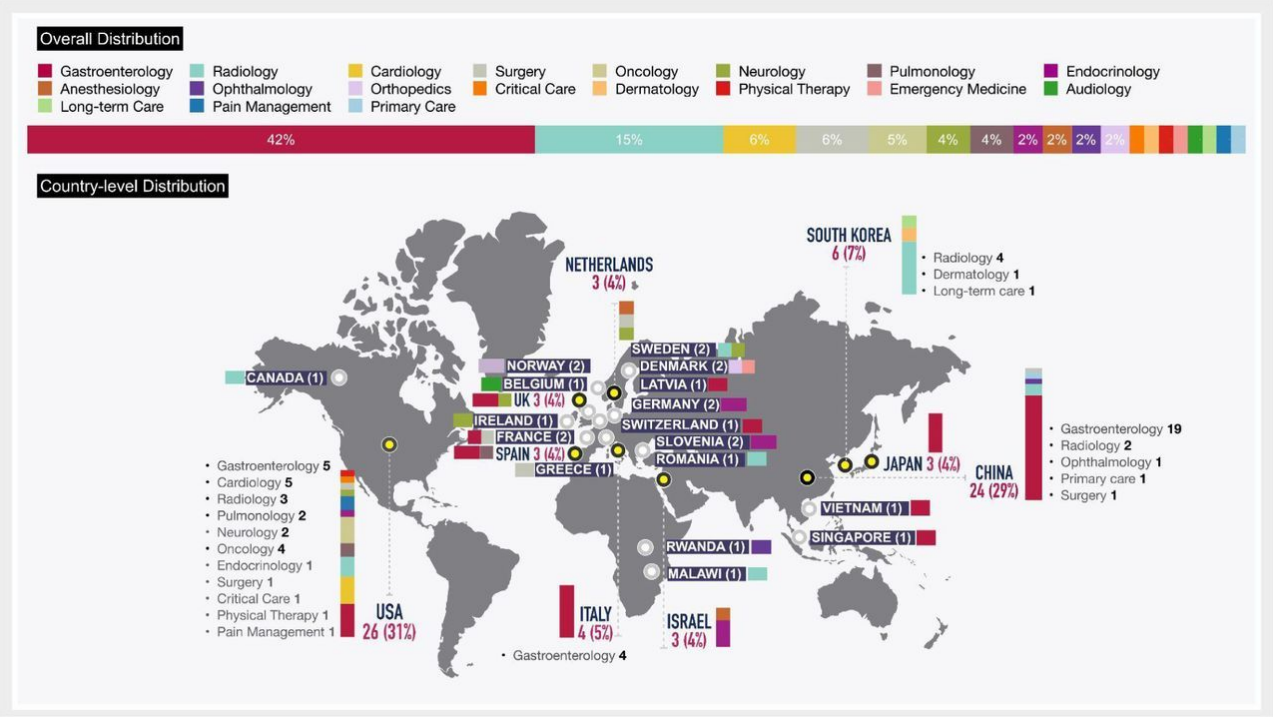
- *Apparent* superior performance across many attributes clinicians are *meant* to be good at
  - Are these studies a fair comparison (ie AI vs a stressed/busy Dr)?
- Potential value of using chatbot to draft responses that the clinician reviews…learns the clinician's style, usual recommendations etc.
- BUT, need to determine IF these AI assistants improve response time/quality/ satisfaction/free up clinician and staff (MA, Nurse, Front Desk etc) time….***or***… do they just drive more eCare messages?

Pain point #3: Chronic disease management

Randomised controlled trials evaluating AI in clinical practice: A scoping review (Han R et al. Lancet Digit Health 2024)

Scoping review of RCTs of AI interventions integrated into *actual* patient management done by clinical teams

- **86 trials**, most in Radiology, Cardiology, GI (....very few Primary Care)
- Most reported positive primary endpoints
  - Mostly diagnostic performance eg detection rate
  - Others reporting care management outcomes, clinical decision making
- Mostly single-center trials, varying quality



Evidence for AI powered **chatbot interventions** for chronic disease management

ANNALS OF MEDICINE
2024, VOL. 56, NO. 1, 2302980
https://doi.org/10.1080/07853890.2024.2302980

RESEARCH ARTICLE

OPEN ACCESS Check for updates

A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness

Moh Heri Kurniawan[a,b], Hanny Handiyani[c], Tuti Nuraini[c], Rr Tutik Sri Hariyati[c] and Sutrisno Sutrisno[b]

8 RCTs

- Breast cancer x3, hypertension, Type 2DM, chronic pain, post-cancer treatment, irritable bowel syndrome.
- USA x3, Egypt, Australia, Germany & Switzerland, France, Spain
- Various theoretical approaches (eg behavior change theory)
- Duration 4 wk to 2 yr
- Outcomes - user satisfaction, health-related measures, patient safety.

Tools to support chronic disease management:

- Mismatch between the clinical need to improve chronic disease outcomes, vs the size and quality of evidence.
- AI tools should be no different to other clinical interventions in the burden of proof (and regulatory approval) e.g safety, effectiveness, cost effectiveness, patient outcomes etc)
- Long way to go…

Pain point #4: Diagnostic decision making

- Clinicians already use many different ‘decision aids’ - eg UptoDate, prescribing tools, cardiovascular risk calculators, ECG interpretation, internet searches
- Volume and complexity of medical information and change overwhelming
- Diagnostic errors do occur and impact our patients. Many signals in the EHR are hidden (messy, disparate data) - would ‘more eyes’ not help us?

JAMA Network | Open

Original Investigation | Health Informatics

Large Language Model Influence on Diagnostic Reasoning
A Randomized Clinical Trial

Ethan Goh, MBBS, MS; Robert Gallo, MD; Jason Hom, MD; Eric Strong, MD; Yingjie Weng, MHS; Hannah Kerman, MD; Joséphine A. Cool, MD; Zahir Kanjee, MD, MPH; Andrew S. Parsons, MD, MPH; Neera Ahuja, MD; Eric Horvitz, MD, PhD; Daniel Yang, MD; Arnold Milstein, MD; Andrew P. J. Olson, MD; Adam Rodman, MD, MPH; Jonathan H. Chen, MD, PhD

Reviewed 6 clinical vignettes over 60 minutes

Compared diagnostic performance of physicians randomised to:

- Physician plus LLM (GPT4)
- Physicians + conventional resources
- (***AND*** the LLM on its own, without a physician)

Outcome: Performance on a standardized rubric of diagnostic performance (based on differential diagnosis accuracy, appropriateness of supporting and opposing factors, and next diagnostic evaluation steps, validated and graded via blinded expert consensus)

Table 2. Diagnostic Performance Outcomes

Group	Median (IQR), %		Difference (95% CI), percentage points[a]	*P* value
	Physicians plus LLM	Physicians plus conventional resources		
All participants	76 (66 to 87)	74 (63 to 84)	2 (−4 to 8)	.60
Level of training				
Attending	79 (63 to 87)	75 (61 to 87)	0.5 (−9 to 1)	.92
Resident	76 (68 to 84)	74 (63 to 84)	3 (−6 to 11)	.50
LLM experience				
Less than monthly	76 (63 to 84)	76 (63 to 87)	−0.5 (−8 to 7)	.90
More than monthly	79 (68 to 90)	74 (63 to 84)	5 (−7 to 16)	.40

LLM alone = **better than physicians,** median score per case of **92%**.

Table 2. Diagnostic Performance Outcomes

Group	Median (IQR), %		Difference (95% CI), percentage points[a]	*P* value
	Physicians plus LLM	Physicians plus conventional resources		
All participants	76 (66 to 87)	74 (63 to 84)	2 (−4 to 8)	.60
Level of training				
Attending	79 (63 to 87)	75 (61 to 87)	0.5 (−9 to 1)	.92
Resident	76 (68 to 84)	74 (63 to 84)	3 (−6 to 11)	.50
LLM experience				
Less than monthly	76 (63 to 84)	76 (63 to 87)	−0.5 (−8 to 7)	.90
More than monthly	79 (68 to 90)	74 (63 to 84)	5 (−7 to 16)	.40

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

Compared diagnostic performance of several LLMs for diagnosis of 4 common conditions:

- Appendicitis
- Diverticulitis
- Pancreatitis
- Cholecystitis

Using data from MIMIC database of 2400 real patients (hospital, ICU)

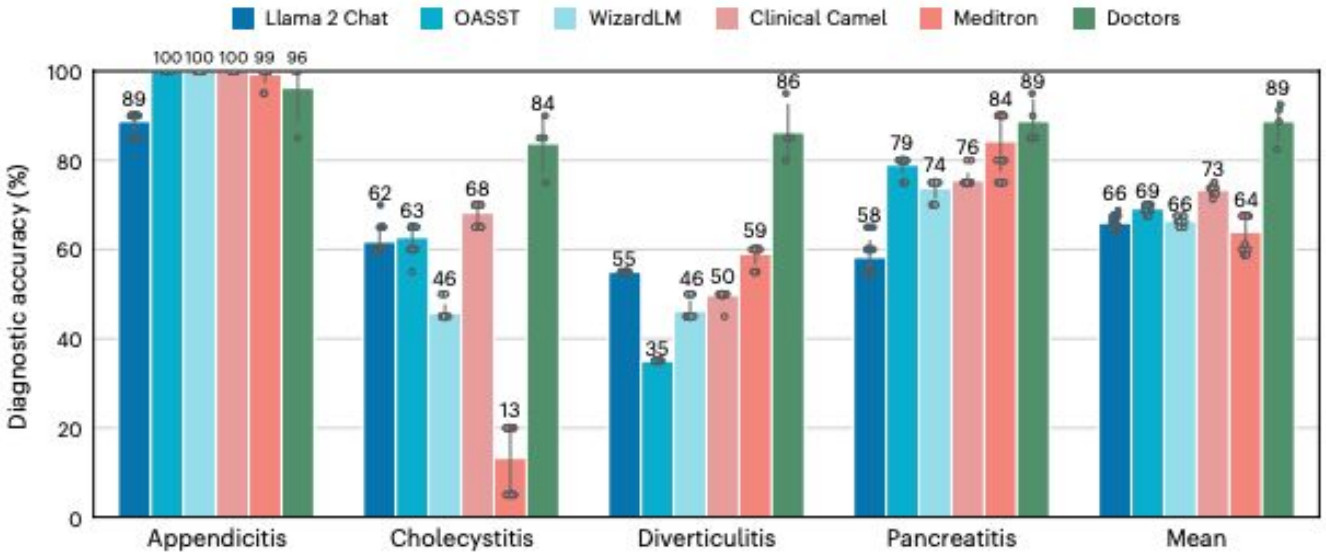


Fig. 2 | LLMs diagnose significantly worse than doctors when provided with all information. On a subset ($n = 80$) of MIMIC-CDM-FI, we compared the mean diagnostic accuracy of LLMs over multiple seeds ($n = 20$) with clinicians ($n = 4$) and found that LLMs perform significantly worse on average ($P < 0.001$) and especially on cholecystitis ($P < 0.001$) and diverticulitis ($P < 0.001$). The mean diagnostic accuracy is shown above each bar. Vertical lines indicate the standard deviation. The individual data points are shown.

LLM as part of decision making: Reasons to be sceptical

Most current research

- Vignettes presenting information in orderly fashion - relevant history, exam, labs, imaging etc.
- Probability of serious illness higher than many primary care settings
- Often MCQ type answers, ie one right answer
- Representativeness of cases, data, settings, SDOH etc?

VS

Real world diagnostic reasoning:

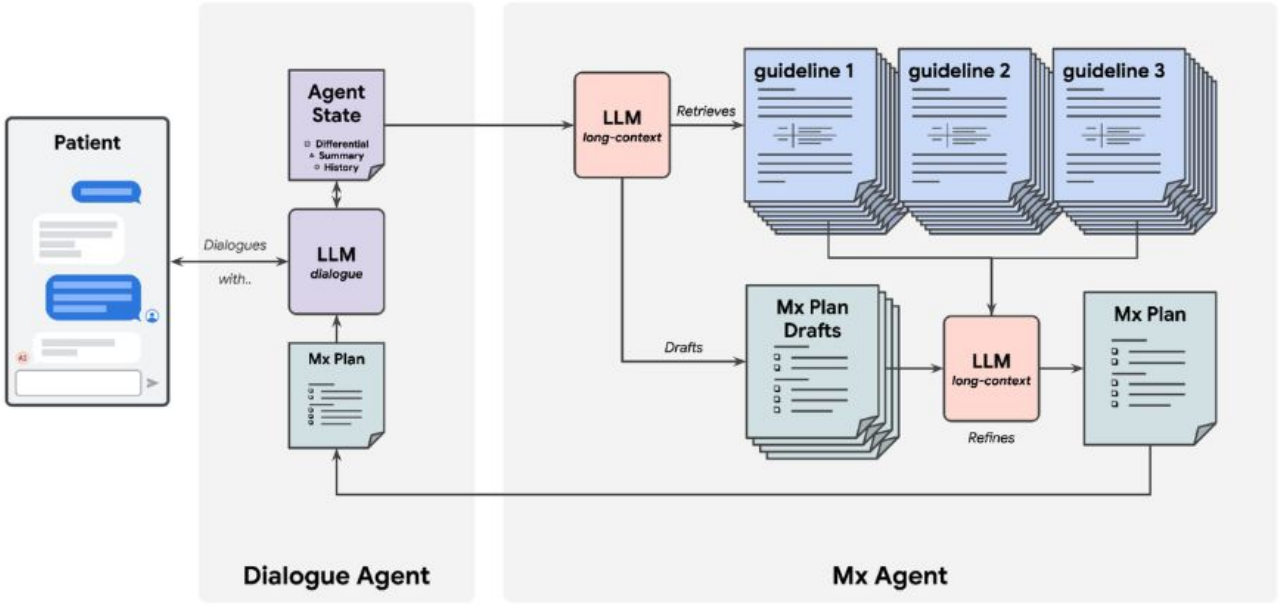
- Iterative over time, non-linear, messy, multiple inputs from patient, caregivers, other clinicians.
- Often no ‘one’ diagnosis
- ‘rule out’ in primary care (low probability of serious conditions)
- Use resources (referrals, labs, imaging), judiciously.
- Balancing a ‘new’ problem among a **patient’s myriad other concerns, priorities, preferences, beliefs**

Computer Science > Computation and Language

[Submitted on 8 Mar 2025]

Towards Conversational AI for Disease Management

Anil Palepu, Valentin Liévin, Wei-Hung Weng, Khaled Saab, David Stutz, Yong Cheng, Kavita Kulkarni, S. Sara Mahdavi, Joëlle Barral, Dale R. Webster, Katherine Chou, Avinatan Hassidim, Yossi Matias, James Manyika, Ryutaro Tanno, Vivek Natarajan, Adam Rodman,



AMIE's two-agent architecture: The Dialogue Agent interacts with the patient, while the Mx Agent creates structured management plans based on clinical guidelines. Management plans define the sequence of investigations and treatments recommended for that patient.

arXiv > cs > arXiv:2503.06074

Search...

Help | Adva

Computer Science > Computation and Language

[Submitted on 8 Mar 2025]

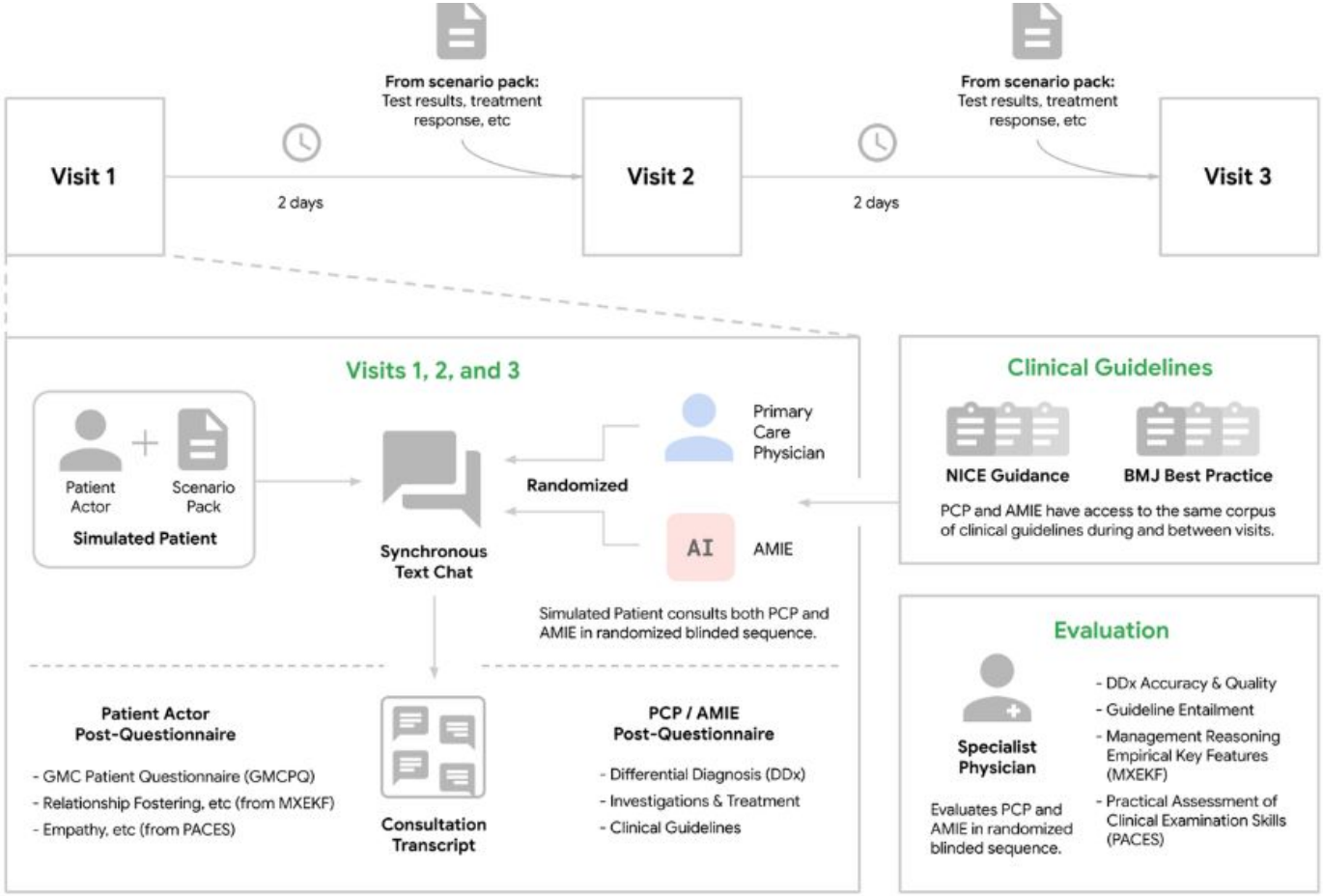
Towards Conversational AI for Disease Management

Anil Palepu, Valentin Liévin, Wei-Hung Weng, Khaled Saab, David Stutz, Yong Cheng, Kavita Kulkarni, S. Sara Mahdavi, Joëlle Barral, Dale R. Webster, Katherine Chou, Avinatan Hassidim, Yossi Matias, James Manyika, Ryutaro Tanno, Vivek Natarajan, Adam Rodman, Tao Tu, Alan Karthikesalingam, Mike Schaekermann

A randomized, blinded virtual OSCE study, comparing AMIE to 21 PCPs across 100 multi-visit case scenarios and five medical specialties

.

Outcome: Panel of specialist physicians and patient actors assessed AMIE and PCPs on multiple domains of management reasoning (including appropriateness, completeness, the use of clinical guidelines, and patient-centeredness)



Overview of randomized multi-visit OSCE study.

Findings:

- AMIE's management reasoning capabilities overall were ***non-inferior to PCPs***, and ***scored better*** in preciseness of treatments and investigations, and in its alignment with management plans in clinical guidelines.
- For medication reasoning ***AMIE outperformed PCPs*** on the subset rated as higher-difficulty by pharmacists.

Pain point #5: Missing bad things: could AI help us spot serious disease sooner?

Hb 4.2 g/dL. Diagnosis = Transient Erythroblastopenia of Childhood. Treatment = transfusion. Full recovery

Lung cancer -

- leading cause of cancer death in the US
- A study of over 43 million patients using Medicare claims data identified a median time from symptom onset to diagnosis of approximately 6 months.
- About half present at Stage 3 or 4

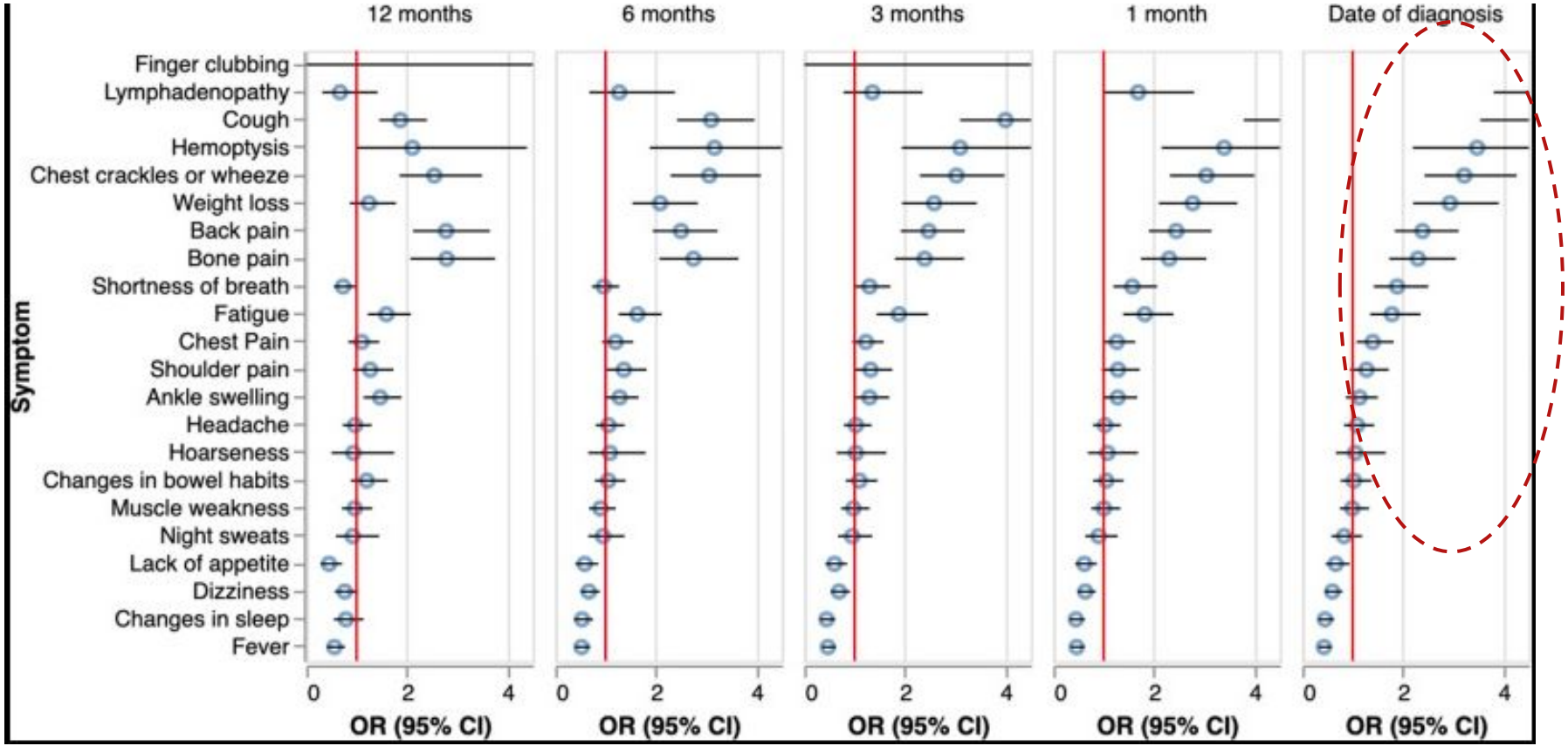
BMJ Open

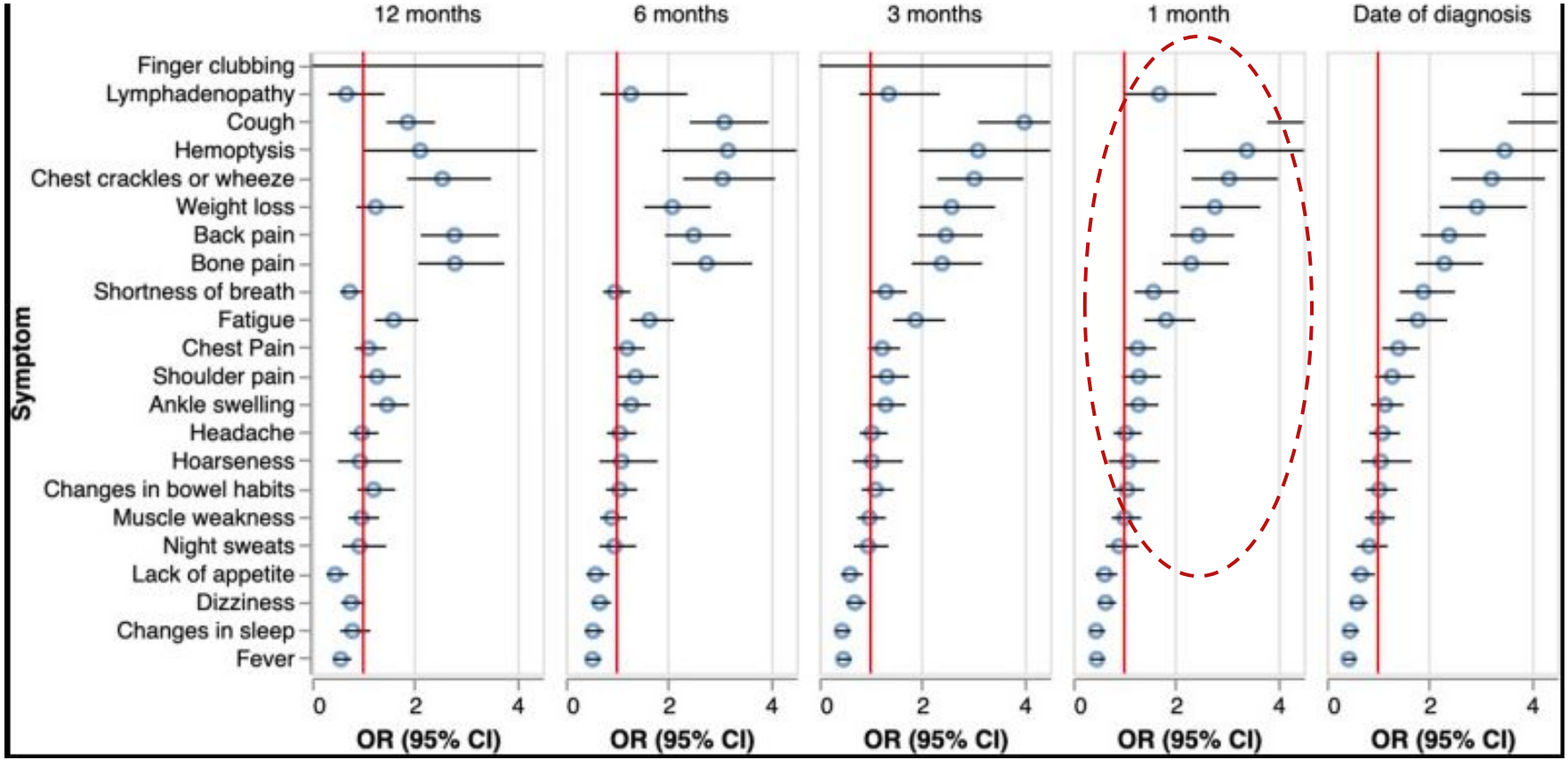
Symptoms and signs of lung cancer prior to diagnosis: case–control study using electronic health records from ambulatory care within a large US-based tertiary care centre

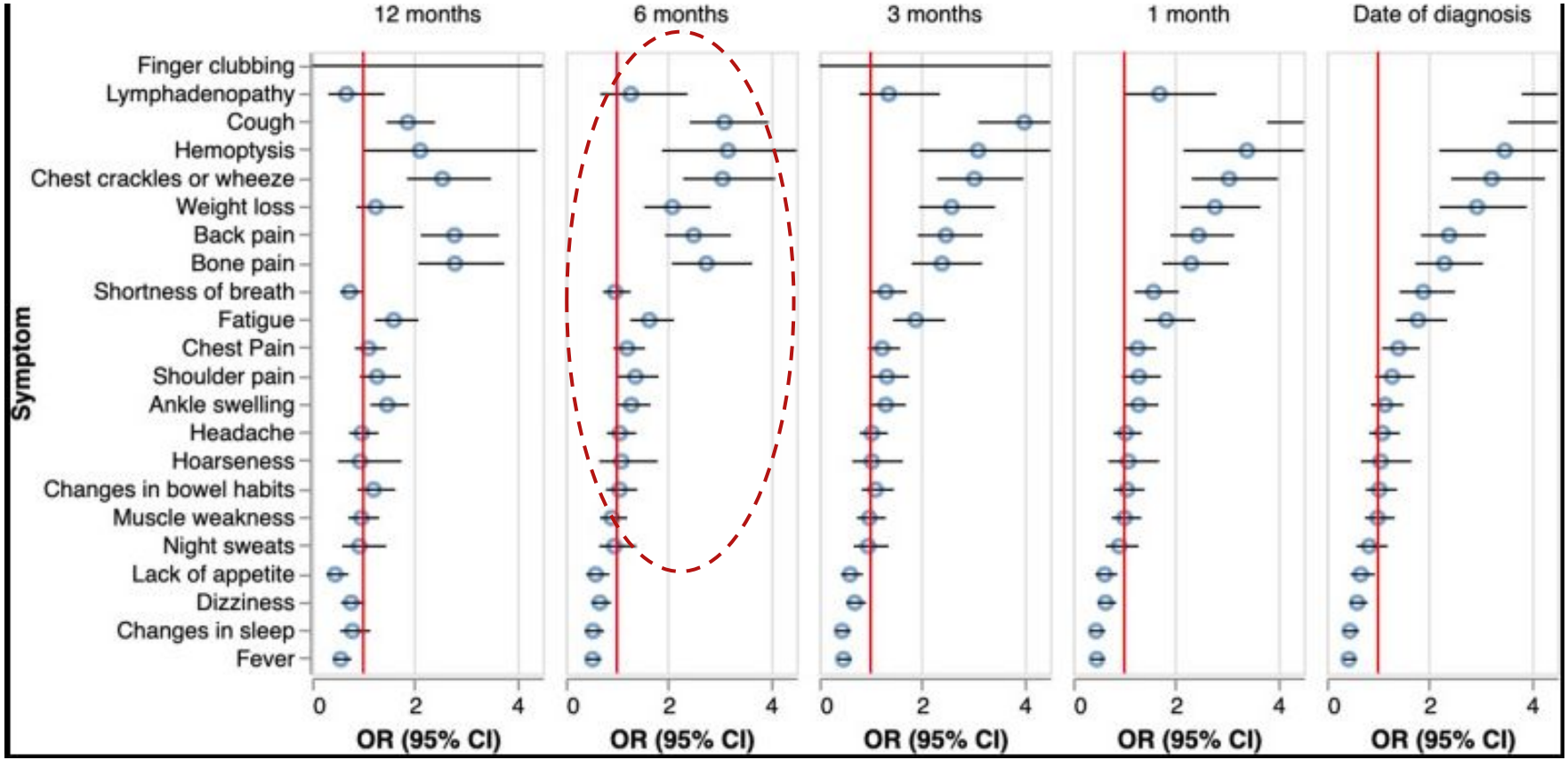
Maria G Prado,[1] Larry G Kessler,[2] Margaret A Au,[1] Hannah A Burkhardt,[3] Monica Zigman Suchsland ,[1] Lesleigh Kowalski,[1] Kari A Stephens,[1] Meliha Yetisgen,[3] Fiona M Walter,[4,5] Richard D Neal,[6] Kevin Lybarger,[7] Caroline A Thompson,[8,9] Morhaf Al Achkar ,[1] Elizabeth A Sarma,[10] Grace Turner,[3] Farhood Farjah,[11] Matthew J Thompson [1]

Case control study from UW Medicine:
600 people with lung cancer *seen in ambulatory care*
6000 controls matched by age, gender, type of clinic

Obtained all their clinical records. Extracted all the coded symptoms AND all the free text comments (e.g in the history) from 12 months before diagnosis using an AI tool called NLP (Natural Language Processing)







Decision support tools:

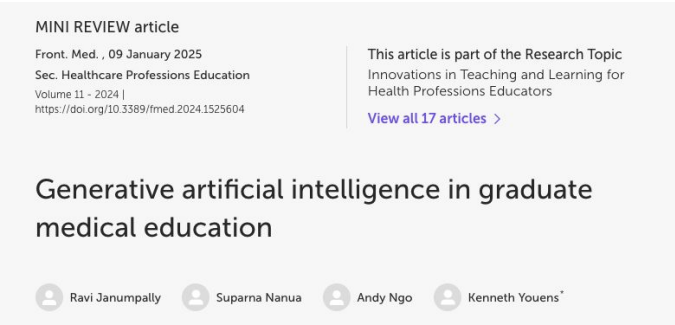
- ***Early evidence*** for complex decision making around diagnoses and management of conditions
- How will clinicians work with these systems?
- High bar for regulatory approval, safety and effectiveness in clinical settings (*and in hands of patients/consumers?*)

AI and medical education

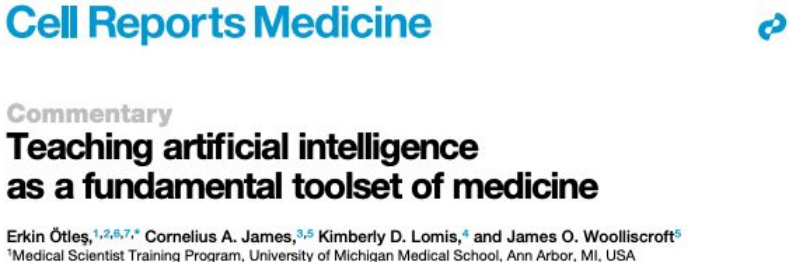
Google Health

Tools

- Content creation using AI - videos, blogs
- Individualised learning plans and content
- Reducing bureaucratic work
- Clinical
  - Real time evaluation of student/resident
  - Feedback on patient interactions, knowledge gaps etc
  - Clinical simulation - conversational, cognitive skills, procedural skills
- Several proposed core curricula (Competencies for the use of AI-based tools by health care professionals (Russell RG et al. Acad Med, 2023; Proposed competencies for use of AI-based tools in primary care (Liaw W et al. Ann Fam Med 2022)



AI for Medical Educators



- How to advance learning that ***keeps pace*** with AI development?
- What are ***core AI competencies*** for clinicians, health care teams, educators, researchers?
- **Question: Is AI (yet another) *addition* to the crowded curriculum, or, a *fundamental* component of medical practice and thus should be deeply integrated?**

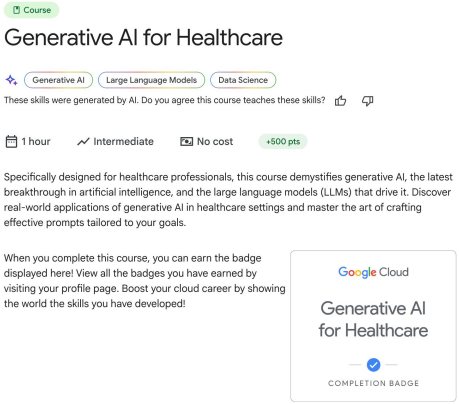
Learn more

STFM
[Artificial Intelligence and Machine Learning for Primary Care Curriculum]

Stanford
https://www.youtube.com/c/StanfordCME

https://hai.stanford.edu/education/professional-education

Google - 1hr course on Generative AI for Healthcare
https://www.cloudskillsboost.google/course_templates/1081



Resources for YouTube videos
https://support.google.com/youtube/answer/14867414?hl=en

Hype or Hope.....?

Take home points

- Understanding AI and Gen AI is ***difficult*** for most of us *(check out the resources)*
- Some tasks can likely be ***safely shifted now*** *(and could make our lives better)*
- ***Real impact*** of Gen AI tools on many areas of health care…*remains to be seen*.
- Clinicians/educators who use AI or Gen AI as ***part of their work*** will be the norm
- ***Gulf*** between AI capabilities & clinician/educator capacity to evaluate, implement
- Multiple ***human attributes*** developed over millions of years are not replaceable (*and nor should we try to*)--- understanding, joy, empathy, connection, touch, kindness, love…