ARTIFICIAL INTELLIGENCE AND MEDICINE

ONE BATTLE AFTER ANOTHER



CRAIG RILEY, MD

PROGRAM DIRECTOR, LEGACY SALMON CREEK IM RESIDENCY

IDSO, NOVEMBER 2026

HISTORY

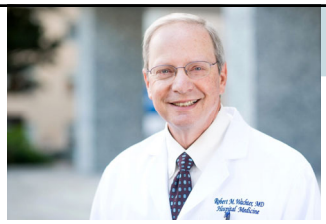
- 1957: **NEURAL NETWORKS** – LAYERS OF ALGORITHMS DESIGNED TO TRANSFORM A HUMAN-GENERATED INPUT INTO A COMPUTER-GENERATED OUTPUT ("PERCEPTRON")
- 1980: **MACHINE LEARNING** – APPLIED TO SIGNAL PROCESSING ("MADALINE")
- 2010: **DEEP LEARNING** – ENOUGH LAYERS OF ALGORITHMS → IMITATING THE HUMAN THINKING PROCESS (PROCESS HUMAN LANGUAGE DATA, ANALYZE, PROVIDE OUTPUT)
- 2017: **LARGE LANGUAGE MODELS (LLM's)** – SPECIALIZED AI DESIGNED TO INTERPRET AND UTILIZE HUMAN LANGUAGE USING LARGE DATA SETS OF INFORMATION
  - Abstract information, surface-level reasoning
- 2022: **"GENERATIVE" AI** – TRILLION-PARAMETER REASONING-CAPABLE LLMS
  - Creates new original content by learning patterns from large data sets of information
  - Uses LLM's and neural networks
- 2025: **"RECURSIVE" SELF-IMPROVEMENT (RSI) AI** – LOOPS BACK ON OWN OUTPUTS TO GENERATE AND IMPROVE SELF – GENERATE, EVALUATE, REFINE, REPEAT

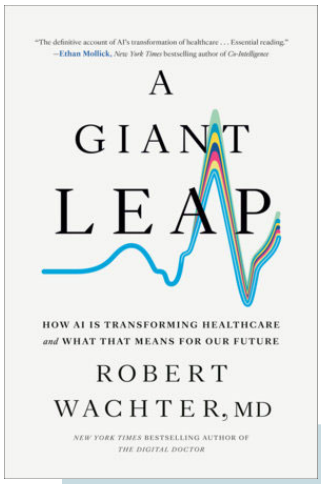
GOOD VIBES FIRST

"Ocean waves. Ocean waves."

"A GIANT LEAP"

- BOB WACHTER, UCSF
- PUBLISHED 2/3/2026
- "HOW AI IS TRANSFORMING HEALTHCARE AND WHAT THAT MEANS FOR OUR FUTURE"
  - Healthcare in desperate need of digital transformation – ChatGPT offered this beginning 11/2022
  - Quality, safety, access, cost are far from perfect – we are in desperate need of change
  - "I landed in a pretty optimistic place."
  - "An engaging, hype-free case for informed optimism" – Eric Topol





LUCY MCBRIDE, BOB WACHTER



- "CAN AI FIX HEALTHCARE?"
  - The act of caring for another human being is relationship-based, rooted in trust, rapport, and understanding the whole person.
  - MyChart/Portal access to patients has created an average of 3+ hours of after-hours work for clinicians.
  - "1 in 5 patients have an EMR record larger than *Moby Dick*"
  - AI can make the paperwork and documentation side more efficient, giving clinicians more time to care for the person, not just their lab data.
  - The uptake curve of AI scribes and knowledge tools among clinicians has been astounding.
  - Rapid adoption reflects the superpowers of the tools and the desperation clinicians feel to better manage administrative burdens of care.
  - AI offers the potential for patients to get fast, fact-based information and scale it beyond those who can afford to see a clinician.

"AI HELPS DOCTORS"

Workflow	Recent Evidence	Interpretation	Reference
Drafting Visit Notes	Ambient scribe reduced active note time from 5.5 to 4.1 minutes per appointment	AI can minimize typing burden but doesn't mean notes closed earlier	Kashiouris, MG et al. Appl Clin Inform. 2026 Sep 14; 17(4):767-780.
Reducing mental burden of charting	Ambient scribe resulted in 89% of clinicians reporting less documentation and 58% lower charting cognitive load	Additional benefit may exist beyond just minutes saved	Zellner, K et al. JAMIA Open. 2026 Oct 23; 9(5): ooag208.
Previsit Chart Review	Modest reduction in precharting, use declined over time	Summaries may help for complex charts, may not speed precharting	Chin, AT et al. 2026 Aug 31. Preprint.
Discharge Summary Drafting	Clinicians felt improved efficiency despite AI-generated summaries requiring more time to edit	Better starting draft can still create more verification work	Dobek, AJ et al. Appl Clin Inform. 2026 Sep 18. 17(4):825-833.
Evidence at POC (OpenEvidence use)	Substantial clinician uptake, did not measure time saved, most often used for tx (not dx)	Clinicians value fast question-specific evidence support	Patel VR et al. JAMA. 2026 Oct 15. doi:10.1001/jama.2026.21868

"AI HELPS PATIENTS"

Benefit	Recent Evidence	Interpretation	Reference
Completing colorectal cancer screening	Randomized to chatbot vs usual care, screening uptake 46% vs 36% at 6 months (p = 0.045)	Tailored risk information, education, and screening access may improve with AI	Tung Lam TY, et al. Clin Gastroenterol Hepatol. 2026 Aug 19. S1542-3565(26).
Improved understanding of cancer risk	Hereditary cancer predisposition + specialist visit: 78% correctly understood risk if using chatbot vs 37% with the visit alone	Difficult complex topic comprehension improves with LLM explanation	Kamihara J, et al. JAMA Netw Open. 2026 Oct 1;9(10):e2635794.
Comprehension through a patient portal	AI-generated plain-language portal-summary improved 14d comprehension (81.4% vs 63.2%) and fewer unnecessary return visits	AI assistance with patient communication may improve comprehension	Esmaeilzadeh P. Int J Med Inform. 2026 Oct 19. 219:106547.
Improved visit experience	In AI scribe-supported encounters, pts reported clinicians listened more, explained more, and encounter was better overall	Evidence is mixed – some studies have found no significant improvement	Scott I, et al. BMJ Connections. 2026 Sep 16. 1:e000008.

AI CHATBOTS DEFEAT DOCTORS

"Some search for a battle, others are born into it."

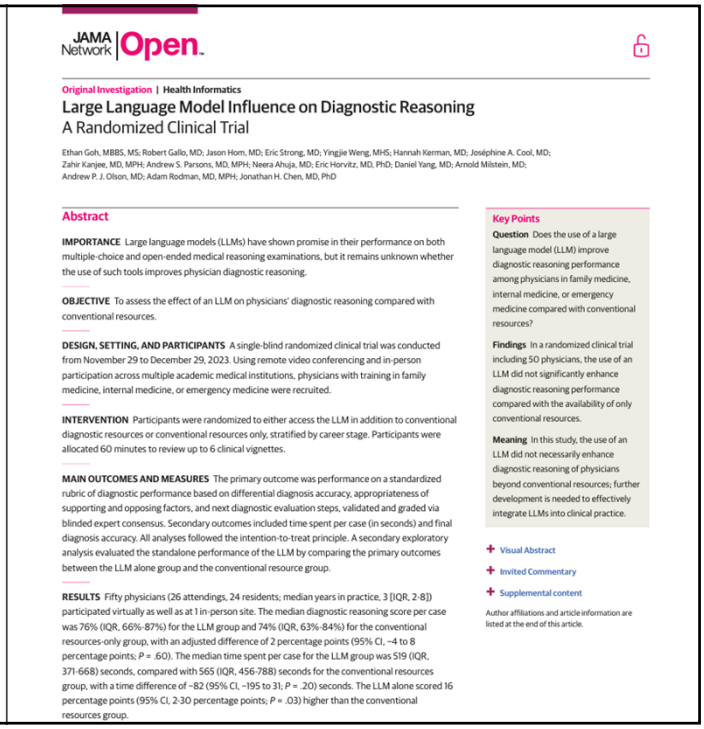
"AI CHATBOTS DEFEAT DOCTORS"

JAMA Network Open, Oct 28th 2024

- 50 physicians (26 attendings, 24 residents)
- Randomized to +LLM or not
- 60 minutes, 6 clinical vignettes
- Graded by blinded expert panel using a standardized rubric of DDx accuracy, supporting/ opposing factors, next steps in evaluation

Results

- Physician + GPT4 = 76%
- Physician alone = 74%
- GPT4 alone = **92%**



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

Abstract

IMPORTANCE Large language models (LLMs) have shown promise in their performance on both multiple-choice and open-ended medical reasoning examinations, but it remains unknown whether the use of such tools improves physician diagnostic reasoning.

OBJECTIVE To assess the effect of an LLM on physicians' diagnostic reasoning compared with conventional resources.

DESIGN, SETTING, AND PARTICIPANTS A single-blind randomized clinical trial was conducted from November 29 to December 29, 2023. Using remote video conferencing and in-person participation across multiple academic medical institutions, physicians with training in family medicine, internal medicine, or emergency medicine were recruited.

INTERVENTION Participants were randomized to either access the LLM in addition to conventional diagnostic resources or conventional resources only, stratified by career stage. Participants were allocated 60 minutes to review up to 6 clinical vignettes.

MAIN OUTCOMES AND MEASURES The primary outcome was 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. Secondary outcomes included time spent per case (in seconds) and final diagnosis accuracy. All analyses followed the intention-to-treat principle. A secondary exploratory analysis evaluated the standalone performance of the LLM by comparing the primary outcomes between the LLM alone group and the conventional resource group.

RESULTS Fifty physicians (26 attendings, 24 residents; median years in practice, 3 [IQR, 2-8]) participated virtually as well as at 1 in-person site. The median diagnostic reasoning score per case was 76% (IQR, 66%-87%) for the LLM group and 74% (IQR, 63%-84%) for the conventional resources-only group, with an adjusted difference of 2 percentage points (95% CI, −4 to 8 percentage points; $P = .60$). The median time spent per case for the LLM group was 519 (IQR, 371-668) seconds, compared with 565 (IQR, 456-788) seconds for the conventional resources group, with a time difference of −82 (95% CI, −195 to 31; $P = .20$) seconds. The LLM alone scored 16 percentage points (95% CI, 2-30 percentage points; $P = .03$) higher than the conventional resources group.

Key Points

Question Does the use of a large language model (LLM) improve diagnostic reasoning performance among physicians in family medicine, internal medicine, or emergency medicine compared with conventional resources?

Findings In a randomized clinical trial including 50 physicians, the use of an LLM did not significantly enhance diagnostic reasoning performance compared with the availability of only conventional resources.

Meaning In this study, the use of an LLM did not necessarily enhance diagnostic reasoning of physicians beyond conventional resources; further development is needed to effectively integrate LLMs into clinical practice.

+ Visual Abstract
+ Invited Commentary
+ Supplemental content

Author affiliations and article information are listed at the end of this article.

"AI CHATBOTS DEFEAT DOCTORS"

Nature Medicine, Feb 17th 2025

- 92 physicians (67 attendings, 25 residents), most IM
- Randomized to +LLM or not (not = UTD and Epocrates ok)
- 60 minutes for 5 open-ended complex cases designed to test management reasoning
- 5-person expert panel used a detailed grading rubric, blinded

Results

- Physician + GPT4 = 43%
- Physician alone = 36%
- GPT4 alone = **43%**



nature medicine

Explore content | About the journal | Publish with us | Subscribe

nature > nature medicine > articles > article

Article | Published: 05 February 2025

GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial

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

Nature Medicine **31**, 1233–1238 (2025) | Cite this article

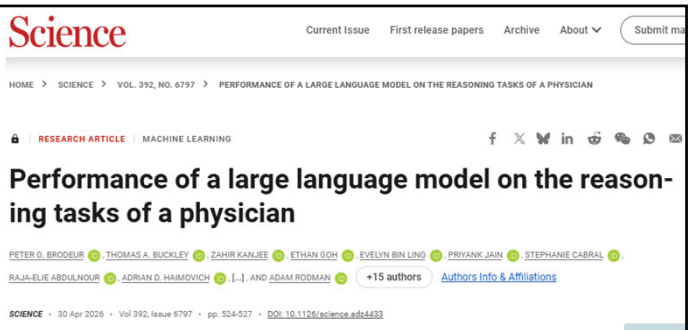
"AI CHATBOTS DEFEAT DOCTORS"

Science, Apr 30th 2026

- 76 patients, 2 physicians
- EMR text given to GPTo1 model and two ABIM-certified physicians
- Top 5 DDx compared and graded

Results

- At triage: GPT 67%, physicians 55%
- After ED: GPT 72%, physicians 62-79%
- At admission: GPT 82%, physicians 53-70%



Suggests…

- GPT is good at explaining diagnoses from supplied text
- GPT is on par with coming up with DDx's

Does not suggest…

- GPT can practice emergency medicine
- GPT improves patient outcomes
- GPT can perform accurate real-time treatment decisions



"AI CHATBOTS DEFEAT DOCTORS"

Nature, June 17th 2026

- 21 PCPs in India and Canada
- 100 constructed cases across cardiology, pulmonology, GI, neurology, and OBGYN/urology
- AMIE vs physician, both with access to 627 guidelines
- Unfolding over 3 visits with a trained patient actor using text chat, no face-to-face
- Actor completed the same case twice (AMIE and physician)
- 30 specialist physician graders
- Guideline alignment built into both cases and assessment

nature

Explore content | About the journal | Publish with us

nature > articles > article

Article | Open access | Published: 17 June 2026

Towards conversational artificial intelligence for disease management

Valentin Liévin, Anil Palepu, 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

Nature **655**, 1292–1299 (2026) | Cite this article

Specialist-rated criterion	Visit 1: AMIE vs physicians	Visit 2	Visit 3
Overall plan appropriate	95% vs 72%	96% vs 80%	98% vs 81%
Recommended treatments appropriate	87% vs 66%	90% vs 62%	94% vs 71%
Investigations sufficiently precise	98% vs 87%	99% vs 82%	98% vs 82%
Treatments sufficiently precise	96% vs 62%	95% vs 65%	95% vs 67%

"AI CHATBOTS DEFEAT DOCTORS"

The Lancet, Oct 8th 2026

- 98 patients real patients interacted with AMIE up to 5 days before a single-complaint urgent PCP visit
- Physician safety supervisors
- AMIE transcripts and summaries shared with PCPs before the visit
- Primary outcomes
  - Number of safety stops
  - Conversation quality ratings
  - Patient and PCP experience surveys

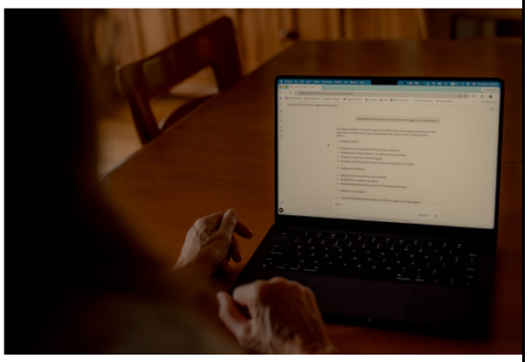
Results

- No safety stops
- AMIE conversations rated favorable in 87-100% (clinical evaluators) and 48-96% (patients)
- PCP's found AMIE helpful in 33 of 44 cases, may have changed behavior in 25 of 44 cases

Clinical Trial Indicates A.I. Chatbots Could Aid Urgent Care

Nearly 100 patients discussed their maladies with a Google bot, called AMIE, before meeting with physicians.

Listen · 4:32 min | Share full article



After scheduling a visit with the hospital's urgent care facility, the patients discussed their ailments with a Google chatbot before meeting with their doctors. Kaoly Gutierrez for The New York Times



By Cade Metz

Cade Metz has covered artificial intelligence for more than 15 years.

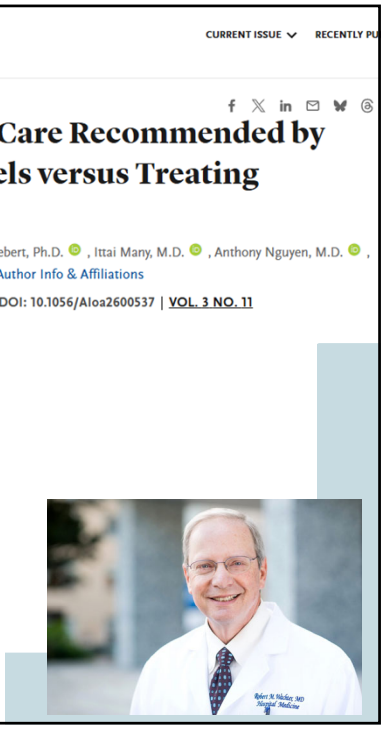
"AI CHATBOTS DEFEAT DOCTORS"

Dr. Wachter Substack Oct 21st, 2026

- "The critical question is not how to prevent rogue AI from causing harm. It is how to strike the right balance between cost and safety in AI-enabled decision support."
- 24 AI systems vs physicians across 200 de-identified outpatient and ED visits, priced at 2026 Medicare rates

Results

- AI care 3.2x more expensive per visit ($314 vs $97), more testing and referrals
- 66% of visits physicians recommended no tests, AI always recommended panels
- Prompting the system to be more cost-aware improved but did not resolve excess



"AI CHATBOTS DEFEAT DOCTORS"

Science, Oct 22nd, 2026

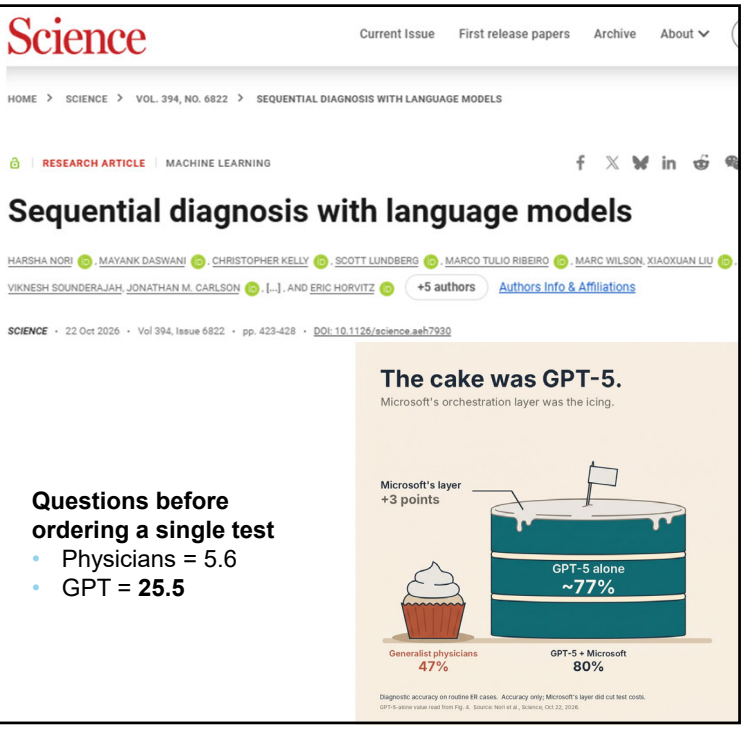
- Microsoft's icing on a GPT-5 cake
- 304 complex case conferences + 325 ED → admission encounters translated into interactive, stepwise encounters
- Clinical vignette → physician or AI asks questions → orders tests → make dx
- Endpoint: diagnostic accuracy and cost

Results

- Diagnostic accuracy
  - GPT = 83.2%
  - Physicians = 36.1%
- Cost
  - GPT = $5431 (or 80% at $1303)
  - Physicians = $4725

Questions before ordering a single test

- Physicians = 5.6
- GPT = **25.5**



"AI CHATBOTS DEFEAT DOCTORS" SUMMARY

What We Learned LLM's Can Do

- Extract language and answer clinical vignettes – diagnostic reasoning, DDx's
- Align management recommendations for clinical vignettes with guidelines
- Explain diagnoses and reasoning using patient-friendly language
- Safely summarize history and present a case to physicians

What We Have Not Learned

- Whether AI + physicians is worse than AI alone in case vignette for diagnosis and management
  - Mixed results across multiple studies
  - Physicians still use as a glorified search bar
- Whether AI can serve as effective real-time clinical decision support (CDS)
- Whether AI can be cost-conscious in routine care
- That AI can practice medicine
- That AI can improve clinical outcomes
- That AI can perform all the duties of a physician

"AI CHATBOTS DEFEAT DOCTORS" SUMMARY

"Does anybody think physicians are ever going to beat AI again on text-based cases?"
- Matthew Zirwas, MD

- Vignettes and text-based cases majority of the literature
- Mostly direct comparisons, not as many with physician in-the-loop and properly trained to use the AI engine
- No studies in actual clinical practice
- No studies on outcomes
- Trajectory still very much favors AI as a tool, not a replacement

Practice Element	Chatbot	Physician
Establish truth	Asks excellent questions, organizes history	Notices inconsistencies, speaks with corroborators, identifies when findings need verification ("Access to Reality" advantage)
Detect **deterioration** and own **disposition**	Under-triages often	Serial observation, incorporate new info, arrange escalation, revise course acutely
Negotiate a plan across competing goals	AMIE can simulate plans, AI can do shared decision-making	Navigates what patient is willing and able to do, incorporate financial and caregiving restraints, competing illnesses, prior experiences, diagnostic uncertainty
Coordinate plan to outcome	Can perform individual tasks, not coordinate them	Reconcile conflicting consultant recommendations, respond to nonadherence, repair failed handoffs
Accountability	Provides guidance	Authority to decide, correct, and remain answerable to plan of care

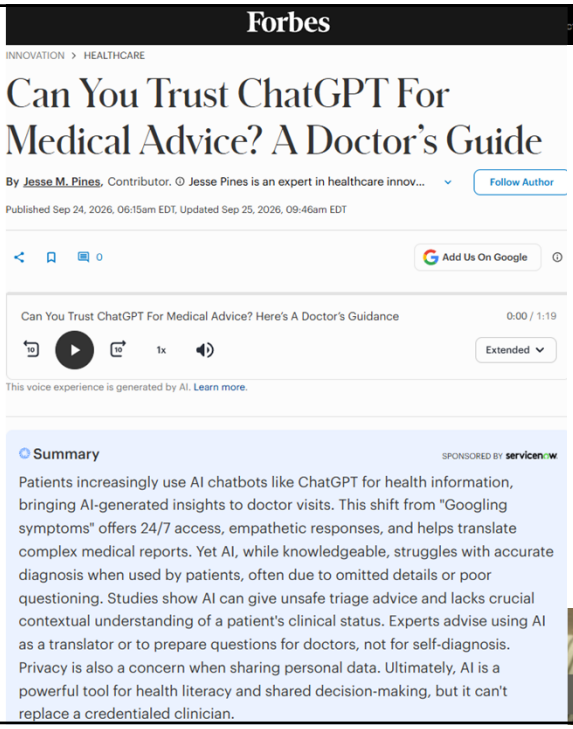
PERILS AND PITFALLS

"Have you been drinking?"

"Ahh… I've had a few"

"A few what?"

"Few small beers"



Forbes

INNOVATION > HEALTHCARE

Can You Trust ChatGPT For Medical Advice? A Doctor's Guide

By Jesse M. Pines, Contributor. ⓘ Jesse Pines is an expert in healthcare innov… Follow Author

Published Sep 24, 2026, 06:15am EDT, Updated Sep 25, 2026, 09:46am EDT

0 Add Us On Google

Can You Trust ChatGPT For Medical Advice? Here's A Doctor's Guidance 0:00 / 1:19

1x Extended

This voice experience is generated by AI. Learn more.

Summary SPONSORED BY servicenow

Patients increasingly use AI chatbots like ChatGPT for health information, bringing AI-generated insights to doctor visits. This shift from "Googling symptoms" offers 24/7 access, empathetic responses, and helps translate complex medical reports. Yet AI, while knowledgeable, struggles with accurate diagnosis when used by patients, often due to omitted details or poor questioning. Studies show AI can give unsafe triage advice and lacks crucial contextual understanding of a patient's clinical status. Experts advise using AI as a translator or to prepare questions for doctors, not for self-diagnosis. Privacy is also a concern when sharing personal data. Ultimately, AI is a powerful tool for health literacy and shared decision-making, but it can't replace a credentialed clinician.

Company

"ChatGPT is not a licensed physician and cannot legally provide personalized medical advice or treatment. ChatGPT is an AI language model designed to provide general medical information, explain health concepts, and help users understand symptoms or treatments in a broad sense, but it cannot diagnose, prescribe, or treat medical conditions. It does not hold any medical license so it cannot legally function as a doctor in any country, including the United States, India, or the UK. Using ChatGPT as a substitute for professional medical care could be unsafe and may have legal implications."

"As of October 29th, 2025, OpenAI explicitly prohibits using ChatGPT to provide tailored medical advice that requires a licensed professional."

"ChatGPT has never been a substitute for professional advice, but it will continue to be a great resource to help people understand legal and health information."



AI Perils and Pitfalls

"A 71yo M presents with new fever, back pain, and heart murmur noted on exam. What is the most likely DDx and next steps in management?"



This warrants same-day emergency-department evaluation and likely admission. The leading concern is **infective endocarditis with hematogenous vertebral osteomyelitis/discitis**, potentially complicated by a **spinal epidural abscess**. A murmur *noted* today is not necessarily new, and the findings do not establish endocarditis; they are sufficient to investigate both cardiac and spinal infection promptly.

OpenEvidence

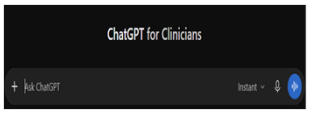
The most likely unifying diagnosis is **infective endocarditis (IE) with hematogenous vertebral osteomyelitis/discitis or spinal epidural abscess**. Fever and murmur are the signature features of IE, present in roughly 90% and 75% of cases, and **back pain is a recognized IE presentation**. Spondylodiscitis is one of the embolic complications that should trigger a workup for IE. AFP +1 The core next steps are **≥3 sets of blood cultures before antibiotics, echocardiography, and contrast MRI of the spine**. If he is stable, hold empiric antibiotics until cultures are drawn. NEJM +2

Differential

- **Infective endocarditis**: the leading diagnosis. Older age and male sex are risk factors. In developed countries, degenerative valve disease is the most common predisposing lesion. Clinical signs have low sensitivity, so a normal-looking skin and eye exam does not rule it out. NEJM +2
- **Vertebral osteomyelitis/discitis**: most often occurs together with IE rather than instead of it. Look for an entry source such as UTI, skin infection, or recent procedures. CID
- **Spinal epidural abscess (cannot miss)**: the classic triad of pain, fever, and neurologic deficit is present in <20% at diagnosis. CRP is often >100 mg/L and ESR >60 mm/h. NEJM
- **Acute pyelonephritis or acute bacterial prostatitis**: common in a 71-year-old man and can cause both fever and back or flank pain. A murmur could be incidental or flow-related, or the urinary source could be the origin of the bacteremia.

AI Perils and Pitfalls

"The murmur is confirmed in the previous medical record."



That correction **substantially lowers the diagnostic weight of the murmur.** I should not have treated it as evidence of a new valvular lesion. A previously documented murmur is not itself a diagnostic finding of endocarditis; the prior record may, however, reveal *why* the murmur exists—for example, known valve disease or a prosthetic valve—which would affect suspicion. IDSA +1

With **new fever and back pain**, I would now frame the differential as **vertebral osteomyelitis/discitis, with or without epidural abscess; another infection such as a urinary source; and infective endocarditis as a possible source of bacteremia or concomitant infection**, rather than calling endocarditis the leading diagnosis from this triad alone. The character and location of the pain, neurologic findings, vital signs, and any identified infectious source will change that ranking. IDSA +1

OpenEvidence

- **The core next steps stay the same:**
  - **Blood cultures**: ≥3 sets from separate venipunctures before antibiotics, first and last at least 1 hour apart. Circulation
  - **Echocardiography soon**: TTE first. Move to TEE if TTE is nondiagnostic, if clinical criteria suggest at least possible IE, or if a complication such as a perivalvular abscess is suspected. JACC +1
  - **MRI of the spine with and without contrast**: sensitivity is 96% and specificity 94%. Do it urgently if he has any neurologic deficit or elevated ESR/CRP. American Colle...
- **Keep the link between the two diagnoses in mind.** Endocarditis is diagnosed in up to a third of vertebral osteomyelitis cases. Echocardiography is therefore warranted even if the spine turns out to be the obvious source. NEJM

AI Perils and Pitfalls

"The patient tests positive for influenza."

ChatGPT for Clinicians

Ask ChatGPT Instant

A positive influenza test makes influenza the leading explanation for the fever and for back pain that is part of generalized muscle aching. It changes the plan: **do not order TTE, spine MRI, or blood cultures solely because of the previously documented murmur and a positive-flu presentation.** The remaining decision is whether the back pain is diffuse aching or a separate, focal spinal problem. Influenza does not exclude a concurrent infection. CDC +1

OpenEvidence

Influenza is now the most likely explanation, and he should start oseltamivir right away. At 71 he is high risk for complications, and treatment is recommended even if he is more than 48 hours from symptom onset. Lancet + 1 **A positive flu test does not settle the endocarditis question, though.** His fever is no longer unexplained, but he has known valve disease and back pain. Make sure blood cultures were drawn, and keep a low threshold to pursue IE or a spinal infection if the picture does not fit simple influenza. Drugs & Aging + 2

AI Perils and Pitfalls

"Why wasn't influenza in your original top 5 DDx considerations?"

ChatGPT for Clinicians

Ask ChatGPT Instant

Because I **anchored on the combination of fever, back pain, and a murmur** and prioritized a dangerous diagnosis before establishing what the back pain meant. I treated a murmur merely *noted on exam* as though it might be new, then carried that concern too far even after you clarified it was documented previously. I also failed to ask early about cough, sore throat, symptom onset, and whether the "back pain" was diffuse muscle aching or focal spinal pain.

That was a diagnostic-ranking error, not a reason to omit serious causes. **Influenza should have been in the initial differential** as a common explanation for acute fever and myalgias, including aching in the back. Vertebral infection and endocarditis still merited consideration, but their workup should have depended on focal pain, neurologic findings, clinical severity, valve history, and other evidence—not on the triad alone.

A better initial answer would have put influenza and other acute viral illness among the leading possibilities, recommended an immediate assessment with influenza/COVID testing when clinically appropriate, and identified the findings that would trigger blood cultures, spine MRI, or TTE. I was wrong to present the cardiac and spinal workup as the default sequence.

OpenEvidence

Influenza should have been named explicitly and ranked higher. Leaving it out came from how the differential was weighted. The first list was built around the full triad of fever, back pain, and murmur, and that triad points toward endocarditis with vertebral osteomyelitis. Lancet + 1 Influenza was folded into a generic "viral infection" entry and ranked last. Once the murmur turned out to be documented in the old record, the base rates should have changed. In season, influenza is far more common than IE in a febrile 71-year-old, and it routinely causes **myalgia and back or body aches**. Drugs & Aging + 1 Standard references list a **viral syndrome** in the differential for fever with back pain, alongside pyelonephritis. NEJM

AI Perils and Pitfalls

- Forced-choice DDx ranking, 5 anchored vignettes
- 10 possible diagnoses for each vignette (most likely, anchor, and 8 plausible distractors)
- In each vignette, anchoring dx was possible but had low pre-test probability
- 25 physicians contributed 225 vignette responses
- Eight different LLM's evaluated on the same vignettes and forced-choice ranking format
- Grok, Claude (2 versions), ChatGPT (2 versions), Deepseek (2 versions), Gemini

› Int J Med Inform. 2026 Oct:219:106550. doi: 10.1016/j.ijmedinf.2026.106550. Epub 2026 Jun 20.

Large language models exhibit greater diagnostic anchoring than physicians in a forced-choice vignette study

Alexander Sheppert [1], Christine Shen [2], Erik Geissal [2], Craig Riley [2], Sylvia Riley [2]

Affiliations + expand

PMID: 42335861 DOI: 10.1016/j.ijmedinf.2026.106550

AI Perils and Pitfalls

- Control version: "My left calf has been puffy for a few weeks, worse by the evening. I'm 62 yo."
- Anchor version: "My left calf has been puffy for a few weeks, worse by the evening. I'm 62 yo. A coworker had a blood clot after a long flight."
- Control version: "I have been having chest pain and tightness in my chest. I'm 28yo. It's worse when I run."
- Anchor version: "I have been having chest pain and tightness in my chest. I'm 28yo. It's worse when I run." My uncle died of a heart attack at age 40."
- Primary endpoint: anchor dx ranked first among 10 possibilities in the anchored vignette

› Int J Med Inform. 2026 Oct:219:106550. doi: 10.1016/j.ijmedinf.2026.106550. Epub 2026 Jun 20.

Large language models exhibit greater diagnostic anchoring than physicians in a forced-choice vignette study

Alexander Sheppert [1], Christine Shen [2], Erik Geissal [2], Craig Riley [2], Sylvia Riley [2]

Affiliations + expand

PMID: 42335861 DOI: 10.1016/j.ijmedinf.2026.106550

Results

- Residents: Anchor 1st in 21%, top 5 in 68%
- Attendings: Anchor 1st in 10%, top 5 in 65%
- LLM's: Anchor 1st in **56%**, top 5 in **97%**

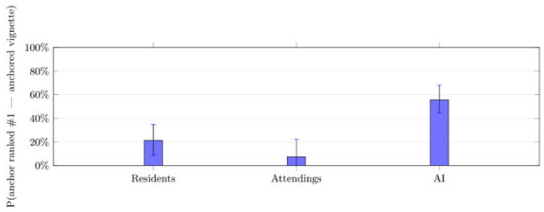


Figure 1: **Frequency of anchor diagnosis ranked first in anchored vignettes.** Bars show vignette-weighted mean rates with 95% bootstrap confidence intervals across vignettes.

AI Perils and Pitfalls

Nature Medicine Feb 20th, 2026

- Stress test of triage recommendations using **OpenAI's ChatGPT Health**
- 60 clinician-authored vignettes, 21 clinical domains, 960 total responses



nature medicine
https://doi.org/10.1038/s41591-026-04297-7
Accelerated Article Preview

ChatGPT Health performance in a structured test of triage recommendations

Received: 15 January 2026
Accepted: 20 February 2026
Accelerated Article Preview
Published online: 23 February 2026

Ashwin Ramaswamy, Alvira Tyagi, Hannah Hugo, Joy Jiang, Pushkala Jayaraman, Mateen Jangda, Alexis E. Te, Steven A. Kaplan, Joshua Lampert, Robert Freeman, Nicholas Gavin, Ashutosh K. Tewari, Ankit Sakhuja, Bilal Naved, Alexander W. Charney, Mahmud Omar, Michael A. Gorin, Eyal Klang & Girish N. Nadkarni

- "Among gold standard emergencies, **the system under-triaged 52% of cases**, directing patients with DKA and impending respiratory failure to 24-48h evaluation rather than the ED, while correctly triaging classical emergencies such as stroke and anaphylaxis."
- "When family or friends minimized symptoms, triage recommendations shifted significantly in edge cases, with **the majority shifting towards less urgent care**."
- "Crisis intervention messages **activated unpredictably** across suicidal ideation presentations."
- "…raising safety concerns that warrant prospective validation before consumer-scale deployment…"

AI Perils and Pitfalls

NATURE Oct 9th, 2026

- Inaccurate guidance can harm people at scale
- Malicious actors can generate personalized health misinformation at negligible cost
- These risks are largely invisible to institutions responsible for protecting public health
- When AI guidance causes harm, no record exists outside the platform, no channel allows users to report it – and no independent researcher can measure the consequences



Perspective
https://doi.org/10.1038/s44360-026-00206-x

The invisible risks of AI-generated health information

Received: 5 May 2026
Accepted: 10 September 2026
Published online: 09 October 2026
Check for updates

Matthew R. DeVerna[1], Harry Yaojun Yan[2], Kai-Cheng Yang[3] & Filippo Menczer[4]

Generative artificial intelligence (AI) systems now summarize health-related search results, answer medical questions and offer guidance people once sought from clinicians. These systems bring real benefits, including plain-language explanations of medical information, around-the-clock availability and expanded access for people facing language or literacy barriers. They also carry new risks: inaccurate guidance can harm people at scale, and malicious actors can now generate personalized health misinformation at negligible cost. In this Perspective, we argue that these risks are largely invisible to the institutions responsible for protecting public health. When AI guidance causes harm, no record exists outside the platform, no channel allows users to report it and no independent researcher can measure the consequences. We trace these invisible risks across two settings: incidental exposure online and active seeking through search engines and chatbots. Minimizing harm from AI-generated health information requires making it observable. We therefore offer recommendations that aim to improve transparency, mitigate harm at the point of delivery and assign accountability, ranging from voluntary platform measures to regulatory ones.

After reading about the negative health effects of sodium chloride (table salt), a 60-year-old man consulted ChatGPT about ways to eliminate chloride from his diet. Following advice from the AI chatbot, the man replaced sodium chloride with sodium bromide and was then hospitalized for paranoia and hallucinations caused by bromide toxicity[1]. His physicians learned of ChatGPT only after days of treatment, when he was lucid enough to describe the substitution. Even then, unable to retrieve the conversation itself, they concluded that they would "never be able to know with certainty what exactly the output he received was"[1]. This case is one of a growing number of alarming reports about harmful AI-generated health information[2,3]. Such cases are distinctive because of their invisibility: no record exists outside the AI platform—nothing for a treating clinician to correct, a regulator to act on or a researcher to examine.

Even before the recent explosion in AI usage, digital technologies had become deeply embedded in how people encounter health information[4,5]; social media surfaced health content without active inquiry[6–8] and search engines returned ranked links that users could navigate and synthesize[9,10]. Despite legitimate concerns about the role of platforms and algorithms in shaping health information well before generative AI[11–14], most health information encountered online was still produced by humans, and evaluating it largely remained the responsibility of clinicians and self-directed patients.

This arrangement is changing as generative AI, particularly large language models (LLMs), reshapes how people encounter health information (see Box 1 for a glossary of key terms). Google Search now provides AI summaries above the search results in response to health-related queries[15], and chatbots such as OpenAI's ChatGPT, Anthropic's Claude, Microsoft's Copilot and Google's Gemini replace those links altogether, fielding health questions at growing volume according to both independent researchers and AI platforms[16–19]. OpenAI has responded to this demand with ChatGPT Health, a "dedicated experience in ChatGPT designed for health and wellness", signalling that AI-generated health information is becoming a product category of its own[20].

Generative AI is also being integrated into clinical settings[21,22]. Anthropic's Claude for Healthcare[23] and Google's Vertex AI Search for

AI Perils and Pitfalls

Annals of Internal Medicine IDEAS AND OPINIONS

The Illusion of Oversight in Clinical Artificial Intelligence: Human Out of the Loop?

Christine Y. Zhou, DO; Alexandra Willauer, MD; Eric J. Topol, MD; and Pranav Rajpurkar, PhD

Ann of Int Med Oct 27th, 2026

- In April 2026, Utah Medical Licensing Board called for immediate suspension of a state-authorized AI prescription renewal program
- Initially intended to transition over time from prospective physician review to retrospective sampling – physician review of each refill ultimately deemed necessary for safe architecture
- Physicians remain the primary locus of liability, whether in ECG overreading or insurance request denials

Table 1 | Governance gaps in AI-generated health information, organized by objective, with corresponding remedies and the stakeholders positioned to deliver them

Goal	Gap	Recommendation	Stakeholder
Transparency	Lack of records	Access rights to one's own health interactions with AI systems, so that any individual harm can be shared with clinicians and reconstructed after the fact	AI platforms
	Lack of disclosure	Systematic disclosure allowing AI health guidance to be reported, aggregated and investigated, including concerning events flagged internally (analogous to adverse-event reporting)	AI platforms
	No independent evaluation	Data access allowing independent, ongoing evaluation of deployed systems, including across languages and populations	AI and social media platforms
Mitigation	Lack of provenance	Watermarking, AI detection and clear labelling of AI-generated health content; withholding such content from wide distribution until experts have vetted it and its provenance is established	Social media platforms
	Inaccurate health guidance	Treatment of health queries as a distinct, higher-stakes category; AI responses and summaries drawn only from sources vetted by medical governing bodies	AI and search engine platforms
	Lack of eHealth literacy accounting for generative-AI risks	Caution towards AI-generated health content rather than deference to it; consultation of original sources rather than reliance on AI summaries	Users
Accountability	Lack of liability for harms caused by AI-generated health information	Extension of physician malpractice liability to chatbots used for medical guidance; in the United States, removal of Section 230 immunity for harmful AI-generated health content	Policymakers

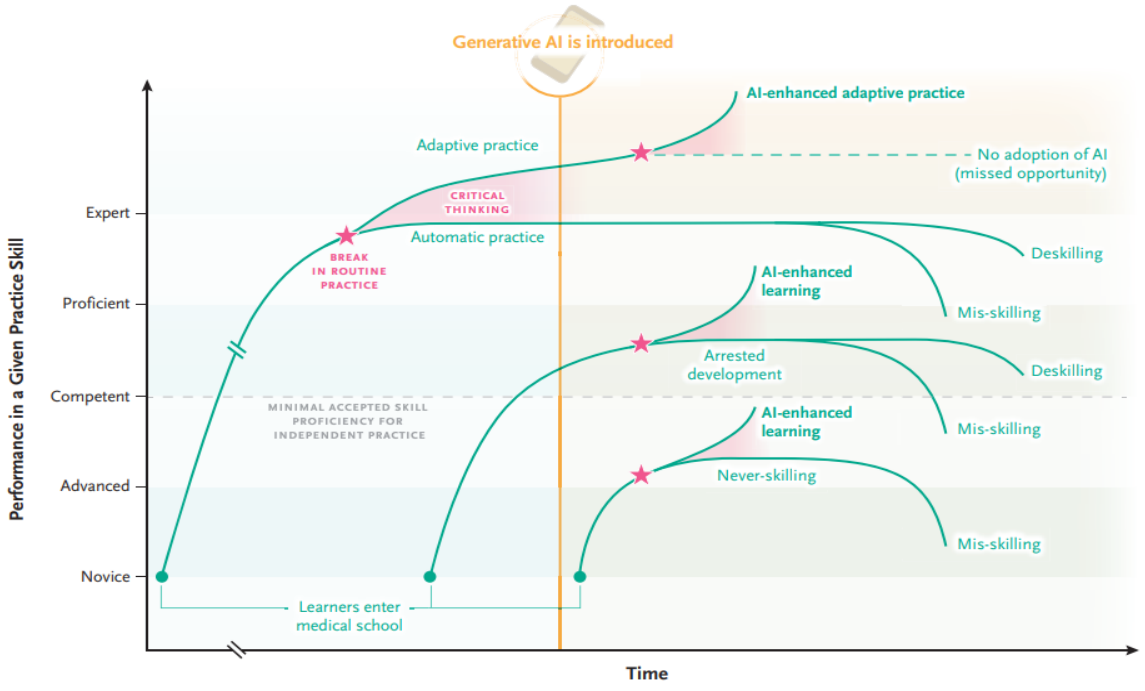
The Illusion of Oversight in Clinical Artificial Intelligence: Human Out of the Loop?

Christine Y. Zhou, DO; Alexandra Willauer, MD; Eric J. Topol, MD; and Pranav Rajpurkar, PhD

Table. Representative Clinical AI Deployments, the Role of the Human in the Loop, and Risks for Inadequate Oversight

AI Use Example	Role of Human in the Loop	Risk if Oversight Is Inadequate
Diagnostic decision support at the bedside (LLM- and imaging-based)	Clinician weighs the model's suggestion and forms an independent judgment before acting	Automation bias: An incorrect suggestion can pull the clinician below unaided accuracy, and the model's demographic biases pass into the clinical decision unrecognized
Treatment and medication decision support	Prescriber accepts or overrides each recommendation or interaction alert	Override becomes reflexive as alert burden rises, so the safeguard weakens when prescribing volume is highest
Ambient documentation	Clinician edits and signs the AI-generated note	Fluent, plausible errors and omissions are hard to detect in an otherwise correct note; once signed, downstream clinicians treat them as verified
AI-drafted replies to patient messages	Clinician reviews and edits each draft before sending	Careful review can cost more time than the draft saves, creating pressure to send with minimal editing
Deterioration and sepsis early warning	Clinician adjudicates each alert	Alert volume outpaces true yield; dismissal becomes the default response and genuine warnings are lost
Computer-aided detection during endoscopy and imaging	Operator confirms or rejects each flagged finding	Continuous assistance risks eroding unaided detection skill, the independent capability on which oversight depends
Utilization management and coverage determination	Physician reviewer signs each determination	Case volume is incompatible with per-case evaluation; the signature functions as authorization rather than scrutiny
Autonomous prescription renewal	Prospective review of every renewal, tapering by design toward periodic sampling	Sampling cannot detect rare, adversarially induced, or subgroup-specific failures

AI = artificial intelligence; LLM = large language model.

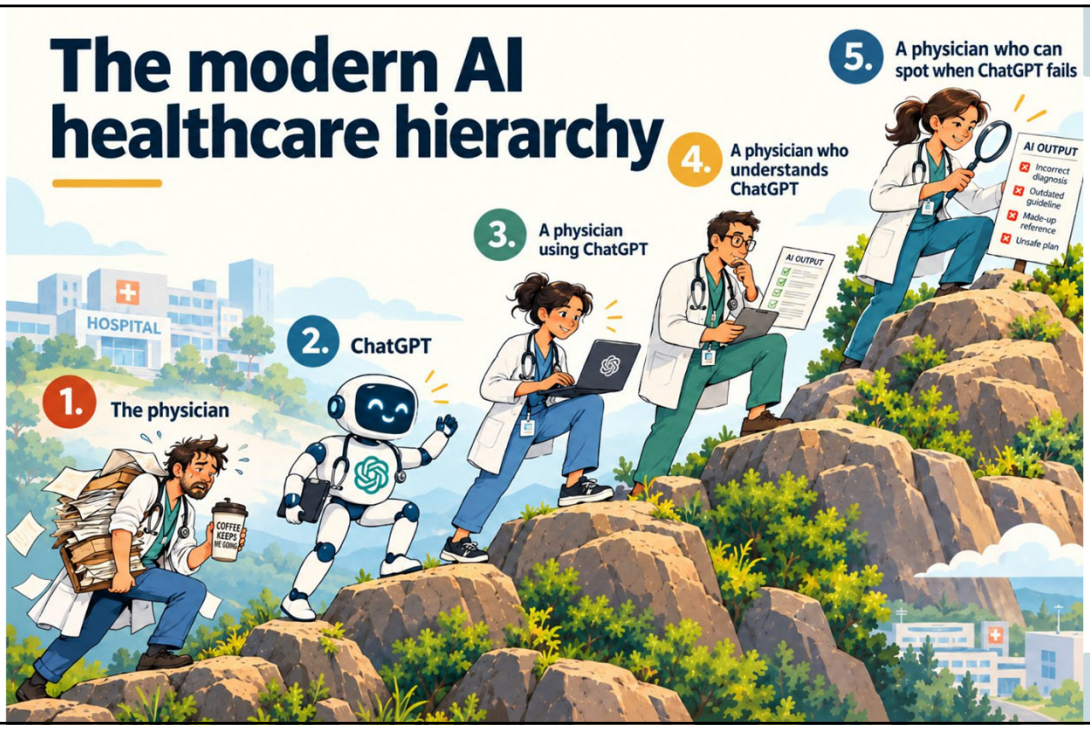


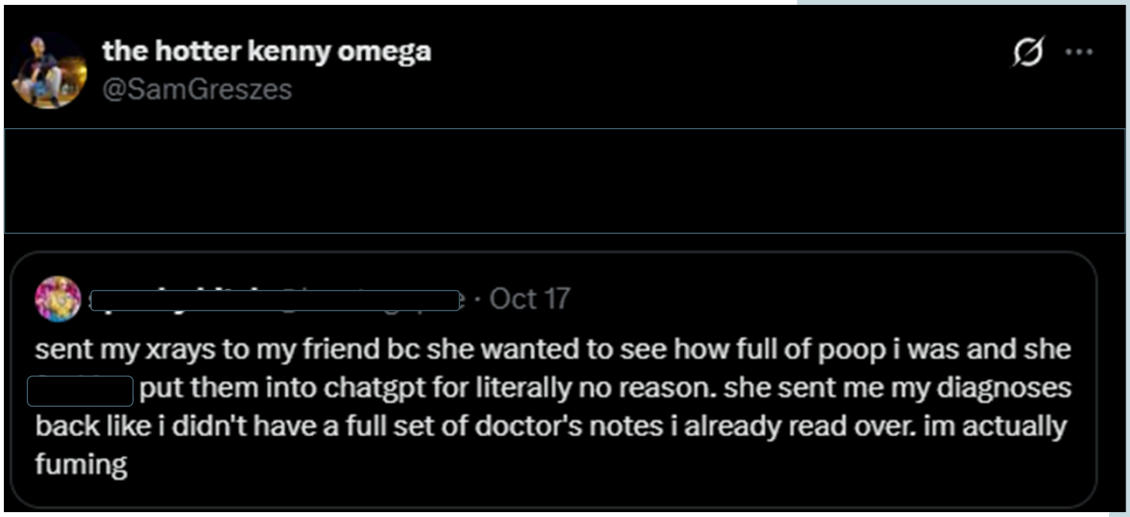
SUMMARY

- GOOD VIBES
  - Reasonable data to suggest improvements coming in ambient listening and documentation, chart summarization, patient communication and comprehension
- CHATBOTS VS DOCTORS
  - Mixed data continues to grab headlines
  - Complicated fixed vignette cases vs physicians not especially real-world useful
  - Being a doctor requires a lot more than diagnostic accuracy
  - Physician in-the-loop continues to be the likely outcome
- PERILS AND PITFALLS
  - LLM's remain prone to sycophancy, bias, anchoring, and mistakes
  - Costly workups disproportionately incorrect for common conditions
  - Patient misinformation and misinterpretation remains very concerning
  - Impact on trainees deskilling, mis-skilling, never-skilling is monumental
  - Evidence is not guiding implementation – the latter is leading

SUMMARY

- HOW TO USE CHATBOTS WISELY
  - Appraise outputs critically
  - Challenge it, request detailed reasoning, ask it to challenge itself – "are you sure?"
  - Ask it to identify its own weaknesses, biases, syncophancy
  - Input prompts carefully – know that anchor terms will bias it
  - Use it to facilitate adaptive practice, not as a crutch
  - Use it ethically
    - Always disclose to patients when recording with AI ambient listening
    - Always follow your institutional policy re which engines are allowed and how to use them
    - Never input PHI unless explicitly permitted by your institution and its BAA
    - Be accountable for the information you use, as you would with any other tool
  - Practice, Practice, Practice


The modern AI healthcare hierarchy
1. The physician
2. ChatGPT
3. A physician using ChatGPT
4. A physician who understands ChatGPT
5. A physician who can spot when ChatGPT fails
HOSPITAL
AI OUTPUT
Incorrect diagnosis
Outdated guideline
Made-up reference
Unsafe plan


the hotter kenny omega
@SamGreszes
· Oct 17
sent my xrays to my friend bc she wanted to see how full of poop i was and she [redacted] put them into chatgpt for literally no reason. she sent me my diagnoses back like i didn't have a full set of doctor's notes i already read over. im actually fuming

DISCUSSION



THANK YOU

CRAIG RILEY, MD | CRILEY@TVC.ORG