AI in Medicine: The Resident Expert That Never Sleeps

Agustina Saenz, MD, MPH

Disclosures

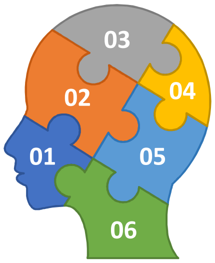
- Senior Clinical Informaticist at Curai Health

Agenda

- Definitions
- Use cases
- Challenges
- Ethics
- Liability
- Model Cards

What is Artificial Intelligence (AI)?

- "Artificial Intelligence refers to the development of computer algorithms that can perform tasks that typically require human intelligence, such as learning, reasoning, perception, and decision making" (ChatGPT)
- AI is based on machine learning algorithms and other computational techniques:



✓ **Probabilistic Reasoning**
Machine Learning
Predictive Modeling
Deep Learning
Decision Trees

✓ **Computational Logic**
Rule-Based Systems
Logic Programming
Heuristic Techniques
Case Based Reasoning

✓ **Optimization Techniques**
Constraint Satisfaction
Constraint-Based Reasoning
Linear Programming
Genetic Algorithms

Perception Systems/Ambient Intelligence ✓
Computer Vision
Olfactive & Haptic
Auditory/Speech

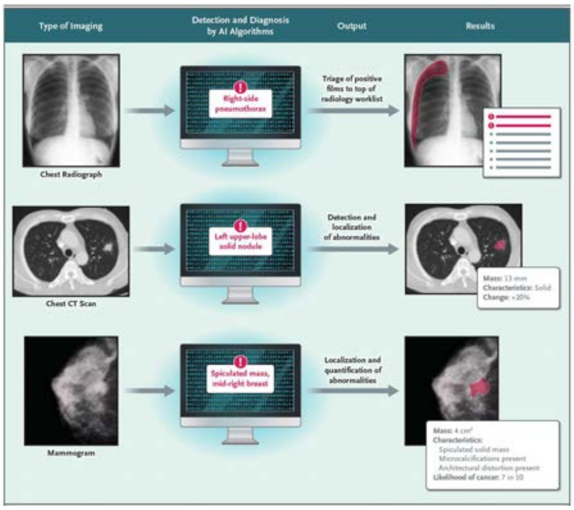
Knowledge Representation, Learning/Search ✓
Knowledge Graphs
Semantic Networks

Natural Language Processing ✓
Text Analytics
Natural Language Understanding
Chatbots
Natural Language Generation
Dialog Management

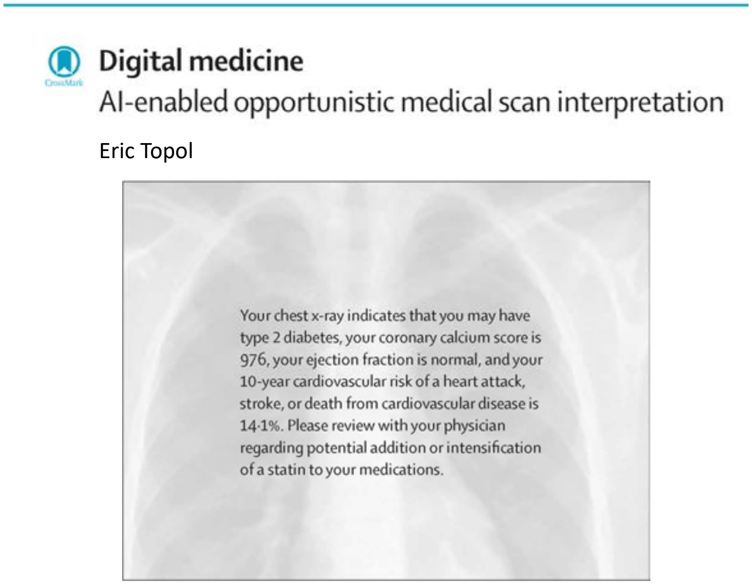
Credit: Adam Landman

Medical Image Interpretation

The NEW ENGLAND JOURNAL of MEDICINE


Type of Imaging
Detection and Diagnosis by AI Algorithms
Output
Results
Chest Radiograph
Right-side pneumothorax
Triage of positive films to top of radiology worklist
Chest CT Scan
Left upper-lobe solid nodule
Detection and localization of abnormalities
Mass: 13 mm
Characteristics: Solid
Change: +20%
Mammogram
Spiculated mass, mid-right breast
Localization and quantification of abnormalities
Mass: 4 cm²
Characteristics:
Spiculated solid mass
Microcalcifications present
Architectural distortion present
Likelihood of cancer: 7 in 10

Rajpurkar, Pranav, and Matthew P. Lungren. "The Current and Future State of AI Interpretation of Medical Images." ***New England Journal of Medicine*** 388.21 (2023): 1981-1990.



Digital medicine

AI-enabled opportunistic medical scan interpretation

Eric Topol

Your chest x-ray indicates that you may have type 2 diabetes, your coronary calcium score is 976, your ejection fraction is normal, and your 10-year cardiovascular risk of a heart attack, stroke, or death from cardiovascular disease is 14·1%. Please review with your physician regarding potential addition or intensification of a statin to your medications.

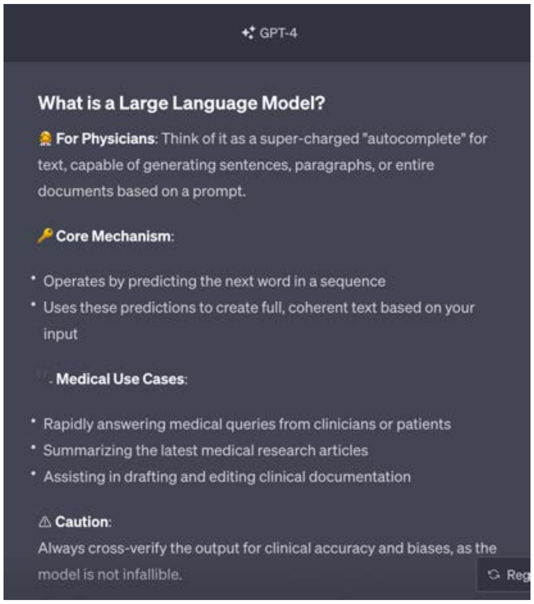
Large Language Models

How frequently do you utilize ChatGPT for work-related activities?*

A) Never used it
B) I used it a few times, but it was not helpful
C) I used it a few times, but I recently started to use it more
D) I use it almost every day, can't work without it!

Mass General Brigham

HARVARD MEDICAL SCHOOL



What is a Large Language Model?

🧑‍⚕️ **For Physicians**: Think of it as a super-charged "autocomplete" for text, capable of generating sentences, paragraphs, or entire documents based on a prompt.

🔑 **Core Mechanism**:

- Operates by predicting the next word in a sequence
- Uses these predictions to create full, coherent text based on your input

🩺 **Medical Use Cases**:

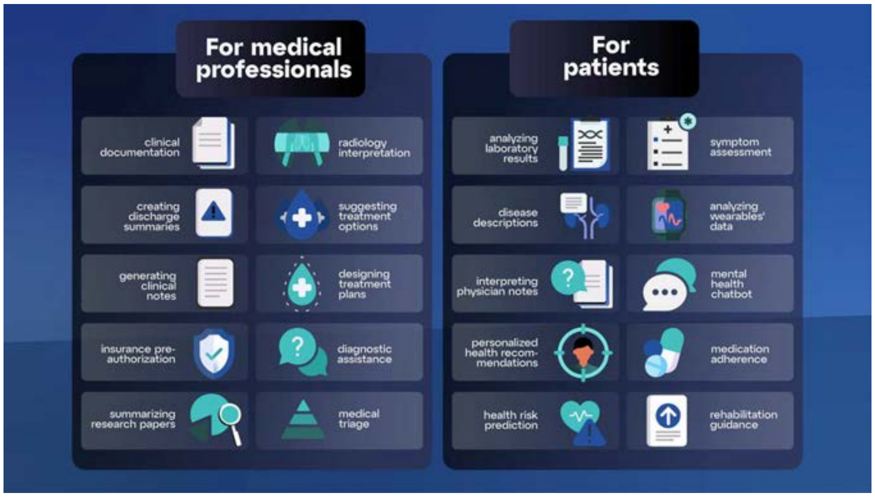
- Rapidly answering medical queries from clinicians or patients
- Summarizing the latest medical research articles
- Assisting in drafting and editing clinical documentation

⚠ **Caution**:
Always cross-verify the output for clinical accuracy and biases, as the model is not infallible.

Mass General Brigham

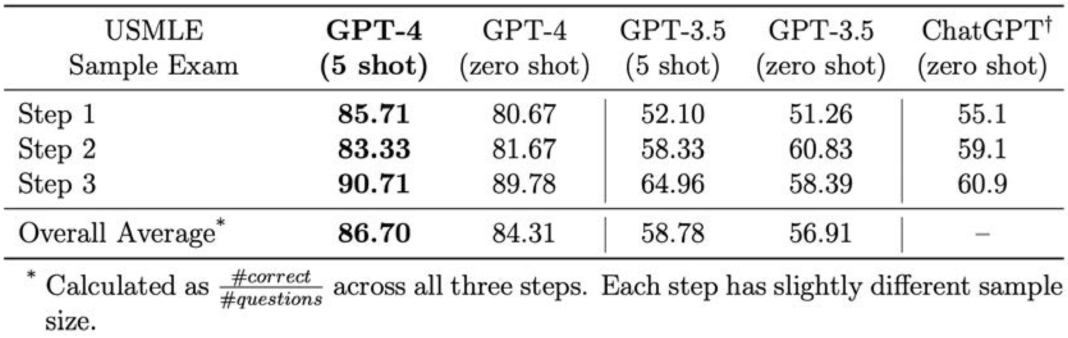
HARVARD MEDICAL SCHOOL

Opportunities for LLMs


For medical professionals
clinical documentation
radiology interpretation
creating discharge summaries
suggesting treatment options
generating clinical notes
designing treatment plans
insurance pre-authorization
diagnostic assistance
summarizing research papers
medical triage
For patients
analyzing laboratory results
symptom assessment
disease descriptions
analyzing wearables' data
interpreting physician notes
mental health chatbot
personalized health recom-mendations
medication adherence
health risk prediction
rehabilitation guidance

Meskó, Bertalan, and Eric J. Topol. "The imperative for regulatory oversight of large language models (or generative AI) in healthcare." ***npj Digital Medicine*** 6.1 (2023): 120.

Medical question answering



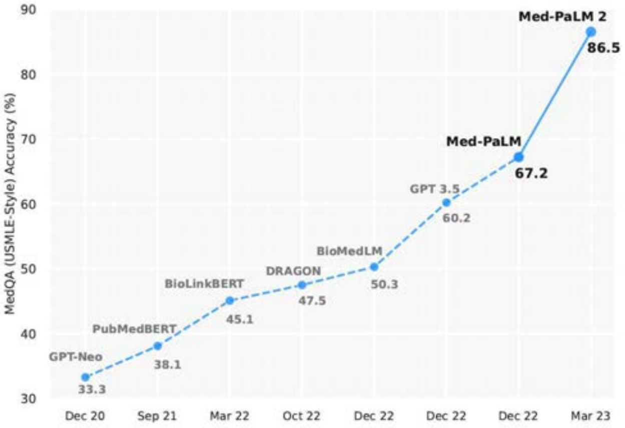
USMLE Sample Exam	**GPT-4 (5 shot)**	GPT-4 (zero shot)	GPT-3.5 (5 shot)	GPT-3.5 (zero shot)	ChatGPT† (zero shot)
Step 1	**85.71**	80.67	52.10	51.26	55.1
Step 2	**83.33**	81.67	58.33	60.83	59.1
Step 3	**90.71**	89.78	64.96	58.39	60.9
Overall Average*	**86.70**	84.31	58.78	56.91	–

* Calculated as $\frac{\#correct}{\#questions}$ across all three steps. Each step has slightly different sample size.

Nori, Harsha, et al. "Capabilities of gpt-4 on medical challenge problems." ***arXiv preprint arXiv:2303.13375*** (2023).

arXiv

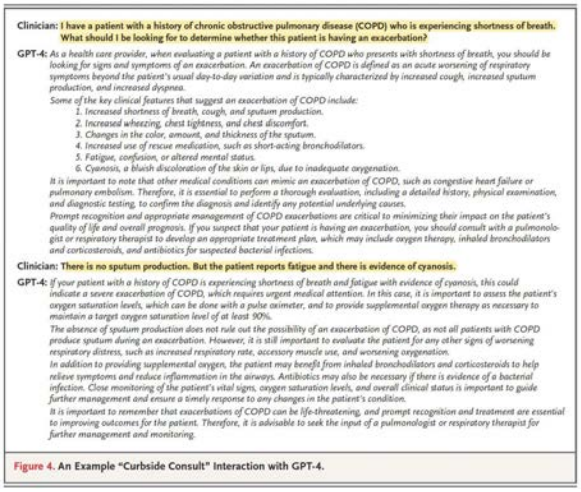
Medical question answering



Singhal, Karan, et al. "Towards expert-level medical question answering with large language models." ***arXiv preprint arXiv:2305.09617*** (2023).

The NEW ENGLAND JOURNAL of MEDICINE

Curbside consult



Clinician: **I have a patient with a history of chronic obstructive pulmonary disease (COPD) who is experiencing shortness of breath. What should I be looking for to determine whether this patient is having an exacerbation?**

GPT-4: *As a health care provider, when evaluating a patient with a history of COPD who presents with shortness of breath, you should be looking for signs and symptoms of an exacerbation. An exacerbation of COPD is defined as an acute worsening of respiratory symptoms beyond the patient's usual day-to-day variation and is typically characterized by increased cough, increased sputum production, and increased dyspnea.*

Some of the key clinical features that suggest an exacerbation of COPD include:

1. *Increased shortness of breath, cough, and sputum production.*
2. *Increased wheezing, chest tightness, and chest discomfort.*
3. *Changes in the color, amount, and thickness of the sputum.*
4. *Increased use of rescue medication, such as short-acting bronchodilators.*
5. *Fatigue, confusion, or altered mental status.*
6. *Cyanosis, a bluish discoloration of the skin or lips, due to inadequate oxygenation.*

It is important to note that other medical conditions can mimic an exacerbation of COPD, such as congestive heart failure or pulmonary embolism. Therefore, it is essential to perform a thorough evaluation, including a detailed history, physical examination, and diagnostic testing, to confirm the diagnosis and identify any potential underlying causes.

Prompt recognition and appropriate management of COPD exacerbations are critical to minimizing their impact on the patient's quality of life and overall prognosis. If you suspect that your patient is having an exacerbation, you should consult with a pulmonologist or respiratory therapist to develop an appropriate treatment plan, which may include oxygen therapy, inhaled bronchodilators and corticosteroids, and antibiotics for suspected bacterial infections.

Clinician: **There is no sputum production. But the patient reports fatigue and there is evidence of cyanosis.**

GPT-4: *If your patient with a history of COPD is experiencing shortness of breath and fatigue with evidence of cyanosis, this could indicate a severe exacerbation of COPD, which requires urgent medical attention. In this case, it is important to assess the patient's oxygen saturation levels, which can be done with a pulse oximeter, and to provide supplemental oxygen therapy as necessary to maintain a target oxygen saturation level of at least 90%.*

The absence of sputum production does not rule out the possibility of an exacerbation of COPD, as not all patients with COPD produce sputum during an exacerbation. However, it is still important to evaluate the patient for any other signs of worsening respiratory distress, such as increased respiratory rate, accessory muscle use, and worsening oxygenation.

In addition to providing supplemental oxygen, the patient may benefit from inhaled bronchodilators and corticosteroids to help relieve symptoms and reduce inflammation in the airways. Antibiotics may also be necessary if there is evidence of a bacterial infection. Close monitoring of the patient's vital signs, oxygen saturation levels, and overall clinical status is important to guide further management and ensure a timely response to any changes in the patient's condition.

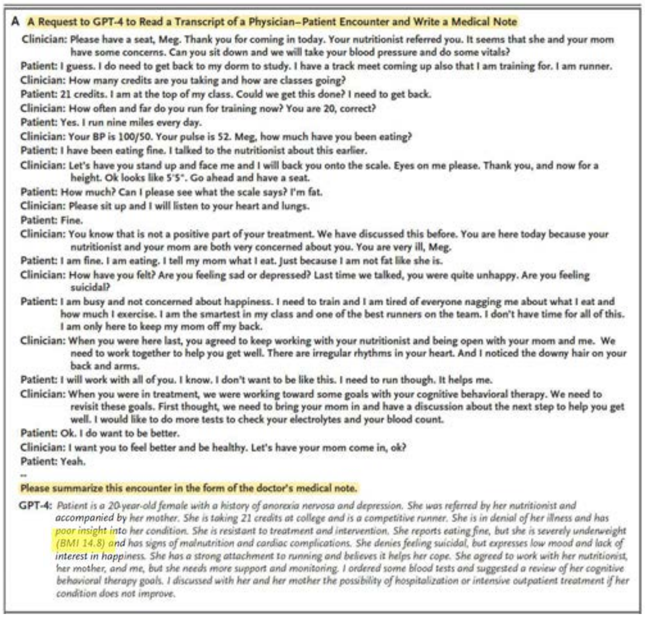
It is important to remember that exacerbations of COPD can be life-threatening, and prompt recognition and treatment are essential to improving outcomes for the patient. Therefore, it is advisable to seek the input of a pulmonologist or respiratory therapist for further management and monitoring.

Figure 4. **An Example "Curbside Consult" Interaction with GPT-4.**

Lee, Peter, Sebastien Bubeck, and Joseph Petro. "Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine." ***New England Journal of Medicine*** 388.13 (2023): 1233-1239.

The NEW ENGLAND
JOURNAL of MEDICINE

Documentation



A A Request to GPT-4 to Read a Transcript of a Physician–Patient Encounter and Write a Medical Note

Clinician: Please have a seat, Meg. Thank you for coming in today. Your nutritionist referred you. It seems that she and your mom have some concerns. Can you sit down and we will take your blood pressure and do some vitals?
Patient: I guess. I do need to get back to my dorm to study. I have a track meet coming up also that I am training for. I am runner.
Clinician: How many credits are you taking and how are classes going?
Patient: 21 credits. I am at the top of my class. Could we get this done? I need to get back.
Clinician: How often and far do you run for training now? You are 20, correct?
Patient: Yes. I run nine miles every day.
Clinician: Your BP is 100/50. Your pulse is 52. Meg, how much have you been eating?
Patient: I have been eating fine. I talked to the nutritionist about this earlier.
Clinician: Let's have you stand up and face me and I will back you onto the scale. Eyes on me please. Thank you, and now for a height. Ok looks like 5'5". Go ahead and have a seat.
Patient: How much? Can I please see what the scale says? I'm fat.
Clinician: Please sit up and I will listen to your heart and lungs.
Patient: Fine.
Clinician: You know that is not a positive part of your treatment. We have discussed this before. You are here today because your nutritionist and your mom are both very concerned about you. You are very ill, Meg.
Patient: I am fine. I am eating. I tell my mom what I eat. Just because I am not fat like she is.
Clinician: How have you felt? Are you feeling sad or depressed? Last time we talked, you were quite unhappy. Are you feeling suicidal?
Patient: I am busy and not concerned about happiness. I need to train and I am tired of everyone nagging me about what I eat and how much I exercise. I am the smartest in my class and one of the best runners on the team. I don't have time for all of this. I am only here to keep my mom off my back.
Clinician: When you were here last, you agreed to keep working with your nutritionist and being open with your mom and me. We need to work together to help you get well. There are irregular rhythms in your heart. And I noticed the downy hair on your back and arms.
Patient: I will work with all of you. I know. I don't want to be like this. I need to run though. It helps me.
Clinician: When you were in treatment, we were working toward some goals with your cognitive behavioral therapy. We need to revisit these goals. First thought, we need to bring your mom in and have a discussion about the next step to help you get well. I would like to do more tests to check your electrolytes and your blood count.
Patient: Ok. I do want to be better.
Clinician: I want you to feel better and be healthy. Let's have your mom come in, ok?
Patient: Yeah.
...

Please summarize this encounter in the form of the doctor's medical note.

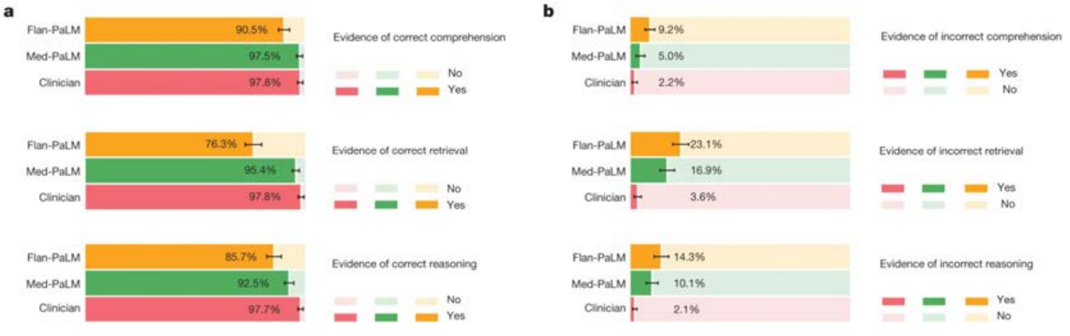
GPT-4: *Patient is a 20-year-old female with a history of anorexia nervosa and depression. She was referred by her nutritionist and accompanied by her mother. She is taking 21 credits at college and is a competitive runner. She is in denial of her illness and has poor insight into her condition. She is resistant to treatment and intervention. She reports eating fine, but she is severely underweight (BMI 14.8) and has signs of malnutrition and cardiac complications. She denies feeling suicidal, but expresses low mood and lack of interest in happiness. She has a strong attachment to running and believes it helps her cope. She agreed to work with her nutritionist, her mother, and me, but she needs more support and monitoring. I ordered some blood tests and suggested a review of her cognitive behavioral therapy goals. I discussed with her and her mother the possibility of hospitalization or intensive outpatient treatment if her condition does not improve.*

Lee, Peter, Sebastien Bubeck, and Joseph Petro. "Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine." ***New England Journal of Medicine*** 388.13 (2023): 1233-1239.

---

nature

Large language models encode clinical knowledge



a,**b**, Evaluation of correctness (**a**) and incorrectness (**b**) of reading comprehension, recall of knowledge and reasoning steps. The results indicate a gap between Flan-PaLM and clinicians, and show that Med-PaLM is able to substantially reduce the gap. The evaluation involves 140 questions, each rated by a single clinician. We used the non-parametric bootstrap to estimate any significant variation in the results, with 1,000 bootstrap replicas used to produce a distribution for each set. We used the 95% bootstrap percentile interval to assess variations.

Singhal, Karan, et al. "Large language models encode clinical knowledge." ***Nature*** (2023): 1-9.

Article | Open access | Published: 04 July 2024

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

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis & Daniel Rueckert

Fig. 2: LLMs diagnose significantly worse than doctors when provided with all information.

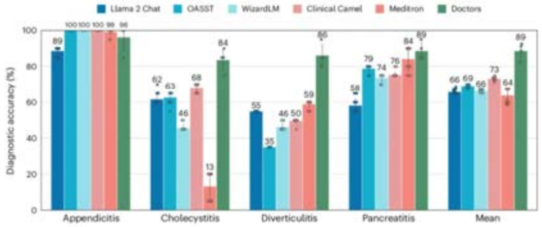
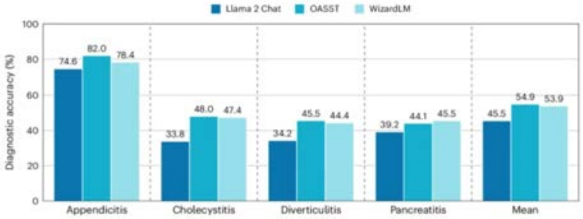


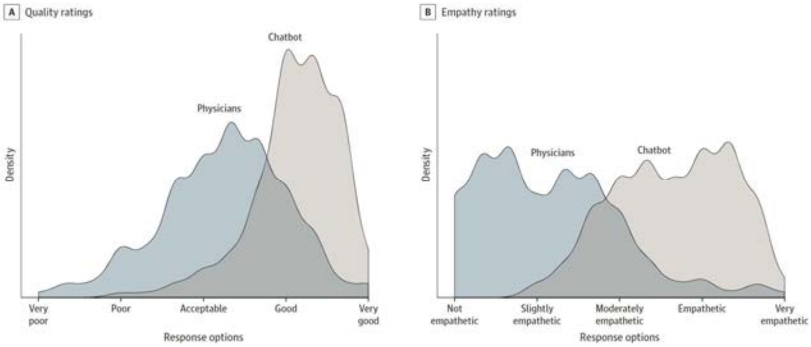
Fig. 3: Diagnostic accuracy of LLMs decreased in an autonomous clinical decision-making scenario.



Answering patient questions

JAMA Internal Medicine

Figure. Distribution of Average Quality and Empathy Ratings for Chatbot and Physician Responses to Patient Questions



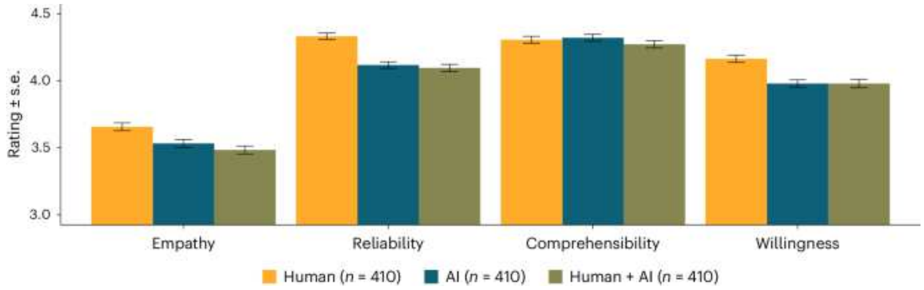
Kernel density plots are shown for the average across 3 independent licensed health care professional evaluators using principles of crowd evaluation. A, The overall quality metric is shown. B, The overall empathy metric is shown.

Ayers, John W., et al. "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum." ***JAMA internal medicine*** (2023).

Brief Communication | Open access | Published: 25 July 2024

Influence of believed AI involvement on the perception of digital medical advice

Moritz Reis, Florian Reis & Wilfried Kunde

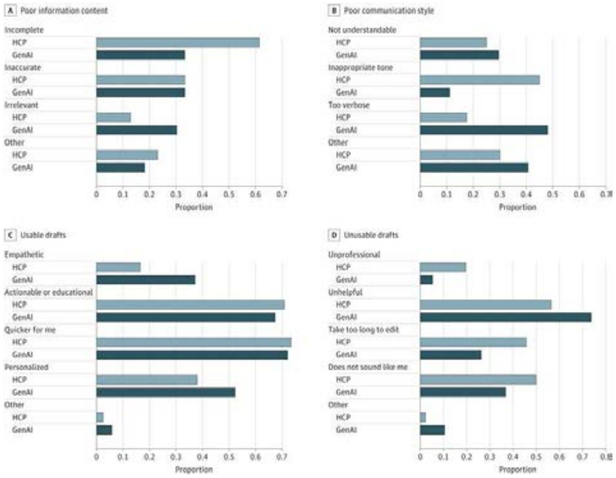


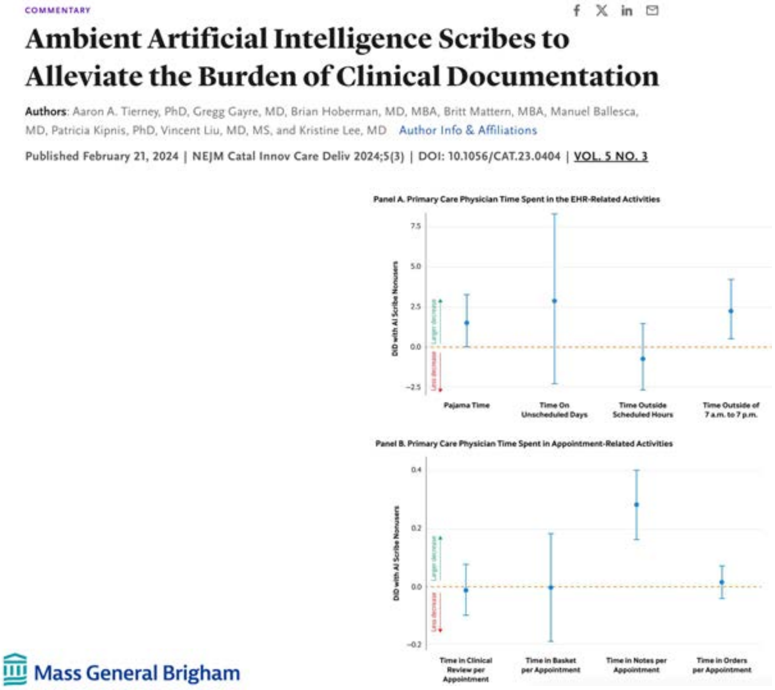
Original Investigation | Health Informatics

July 16, 2024

Large Language Model–Based Responses to Patients' In-Basket Messages

William R. Small, MD, MBA[1]; Batia Wiesenfeld, PhD[2]; Beatrix Brandfield-Harvey, BS[1]; et al





COMMENTARY

Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation

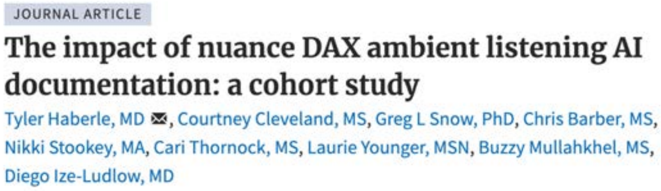
Authors: Aaron A. Tierney, PhD, Gregg Gayre, MD, Brian Hoberman, MD, MBA, Britt Mattern, MBA, Manuel Ballesca, MD, Patricia Kipnis, PhD, Vincent Liu, MD, MS, and Kristine Lee, MD Author Info & Affiliations

Published February 21, 2024 | NEJM Catal Innov Care Deliv 2024;5(3) | DOI: 10.1056/CAT.23.0404 | VOL. 5 NO. 3

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

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

DiD with AI Scribe Nonusers
7.5
5.0
2.5
0.0
−2.5
Larger decrease
Less decrease
Pajama Time
Time On Unscheduled Days
Time Outside Scheduled Hours
Time Outside of 7 a.m. to 7 p.m.
DiD with AI Scribe Nonusers
0.4
0.2
0.0
−0.2
Larger decrease
Less decrease
Time in Clinical Review per Appointment
Time in Basket per Appointment
Time in Notes per Appointment
Time in Orders per Appointment

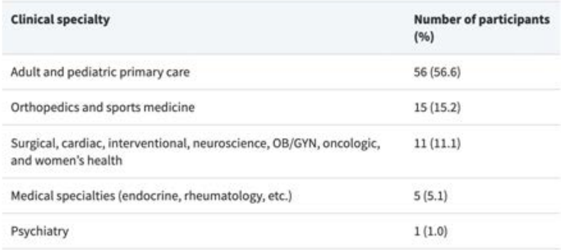


JOURNAL ARTICLE

The impact of nuance DAX ambient listening AI documentation: a cohort study

Tyler Haberle, MD, Courtney Cleveland, MS, Greg L Snow, PhD, Chris Barber, MS, Nikki Stookey, MA, Cari Thornock, MS, Laurie Younger, MSN, Buzzy Mullahkhel, MS, Diego Ize-Ludlow, MD

- 99 providers representing 12 specialties
- 76 matched control group providers
- Median utilization: 47% among active participants
- Nuance DAX use showed positive trends in provider engagement, without detriment to patient safety, experience, or clinical documentation.
- There were no significant benefits to patient experience, documentation, or measures of productivity.

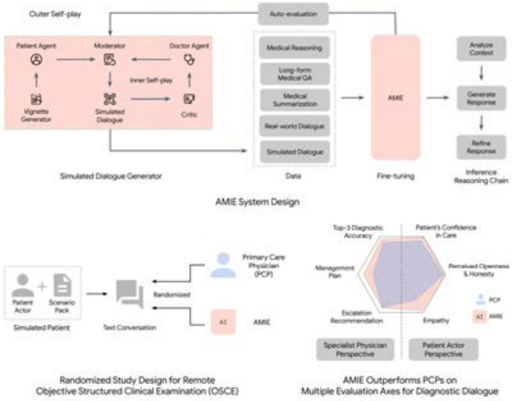


Clinical specialty	Number of participants (%)
Adult and pediatric primary care	56 (56.6)
Orthopedics and sports medicine	15 (15.2)
Surgical, cardiac, interventional, neuroscience, OB/GYN, oncologic, and women's health	11 (11.1)
Medical specialties (endocrine, rheumatology, etc.)	5 (5.1)
Psychiatry	1 (1.0)

Towards Conversational Diagnostic AI

Tao Tu[*,1], Anil Palepu[*,1], Mike Schaekermann[*,1],
Khaled Saab[1], Jan Freyberg[1], Ryutaro Tanno[2], Amy Wang[1], Brenna Li[1], Mohamed Amin[1],
Nenad Tomasev[2], Shekoofeh Azizi[2], Karan Singhal[1], Yong Cheng[2], Le Hou[1], Albert Webson[2],
Kavita Kulkarni[1], S. Sara Mahdavi[2], Christopher Semturs[1],
Juraj Gottweis[1], Joelle Barral[2], Katherine Chou[1], Greg S. Corrado[1], Yossi Matias[1],
Alan Karthikesalingam[†,1] and Vivek Natarajan[†,1]

[1]Google Research, [2]Google DeepMind



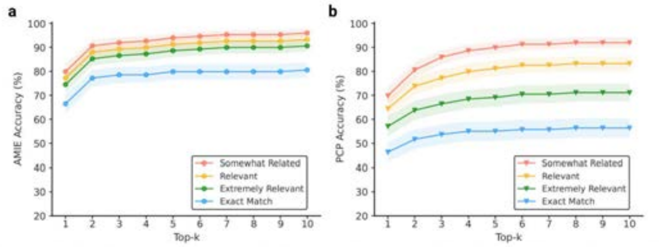
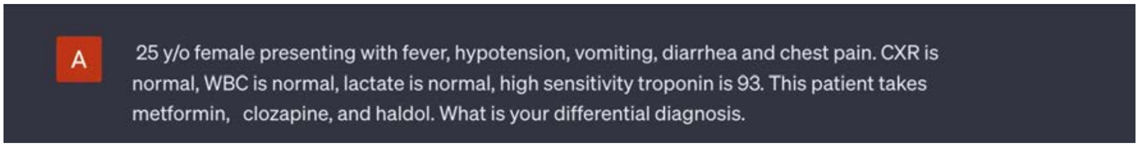
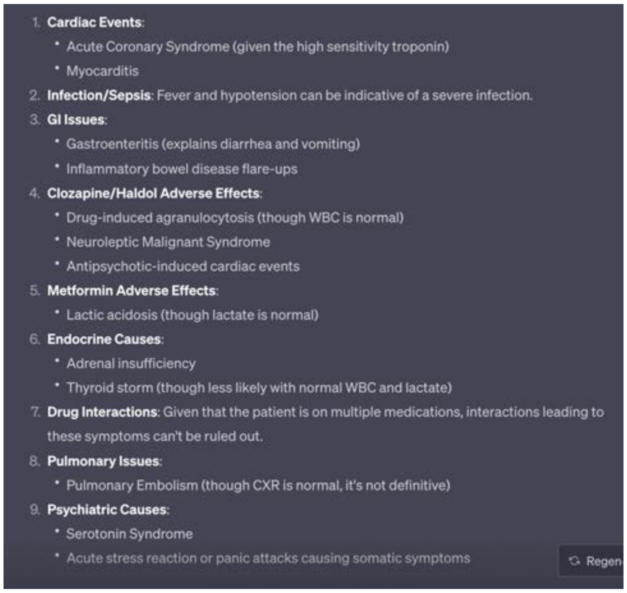


Figure A.7 | Specialist rated DDx accuracy by the degree of matching. (**a**) Specialist rated DDx top-10 accuracy for consultations conducted by AMIE. (**b**) Specialist rated DDx top-10 accuracy for consultations conducted by a PCP. For the "Relevant", "Extremely Relevant", and "Exact Match" levels, differences between AMIE and PCP DDx accuracy are statistically significant (bootstrap with n=10,000 and FDR correction) for all k. Differences at the "Somewhat Related" level are not statistically significant.

Prompt matters



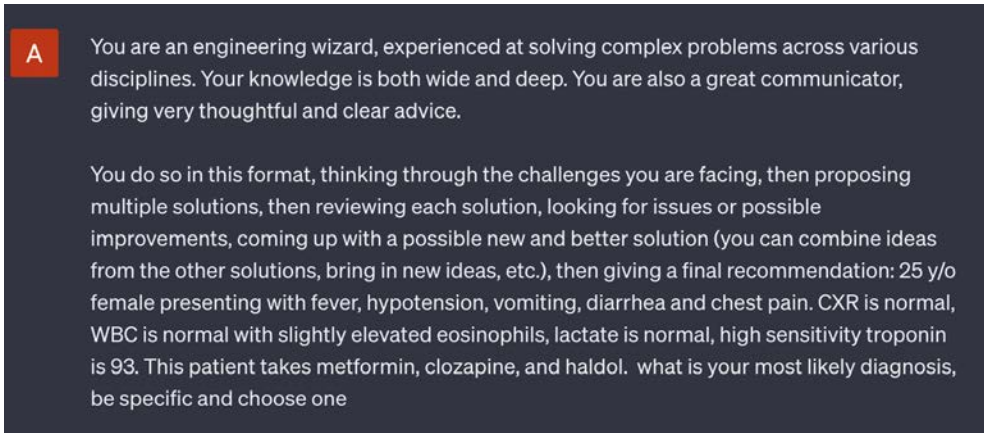
25 y/o female presenting with fever, hypotension, vomiting, diarrhea and chest pain. CXR is normal, WBC is normal, lactate is normal, high sensitivity troponin is 93. This patient takes metformin, clozapine, and haldol. What is your differential diagnosis.



1. **Cardiac Events**:
   - Acute Coronary Syndrome (given the high sensitivity troponin)
   - Myocarditis
2. **Infection/Sepsis**: Fever and hypotension can be indicative of a severe infection.
3. **GI Issues**:
   - Gastroenteritis (explains diarrhea and vomiting)
   - Inflammatory bowel disease flare-ups
4. **Clozapine/Haldol Adverse Effects**:
   - Drug-induced agranulocytosis (though WBC is normal)
   - Neuroleptic Malignant Syndrome
   - Antipsychotic-induced cardiac events
5. **Metformin Adverse Effects**:
   - Lactic acidosis (though lactate is normal)
6. **Endocrine Causes**:
   - Adrenal insufficiency
   - Thyroid storm (though less likely with normal WBC and lactate)
7. **Drug Interactions**: Given that the patient is on multiple medications, interactions leading to these symptoms can't be ruled out.
8. **Pulmonary Issues**:
   - Pulmonary Embolism (though CXR is normal, it's not definitive)
9. **Psychiatric Causes**:
   - Serotonin Syndrome
   - Acute stress reaction or panic attacks causing somatic symptoms

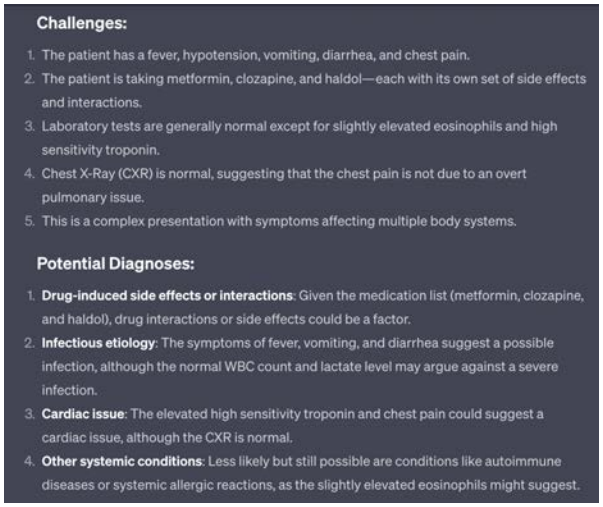
---

Prompt matters: same case



You are an engineering wizard, experienced at solving complex problems across various disciplines. Your knowledge is both wide and deep. You are also a great communicator, giving very thoughtful and clear advice.

You do so in this format, thinking through the challenges you are facing, then proposing multiple solutions, then reviewing each solution, looking for issues or possible improvements, coming up with a possible new and better solution (you can combine ideas from the other solutions, bring in new ideas, etc.), then giving a final recommendation: 25 y/o female presenting with fever, hypotension, vomiting, diarrhea and chest pain. CXR is normal, WBC is normal with slightly elevated eosinophils, lactate is normal, high sensitivity troponin is 93. This patient takes metformin, clozapine, and haldol. what is your most likely diagnosis, be specific and choose one

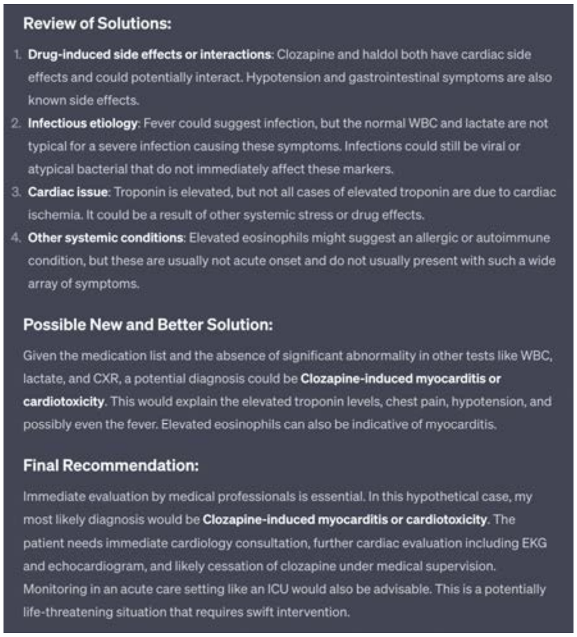


Challenges:

1. The patient has a fever, hypotension, vomiting, diarrhea, and chest pain.
2. The patient is taking metformin, clozapine, and haldol—each with its own set of side effects and interactions.
3. Laboratory tests are generally normal except for slightly elevated eosinophils and high sensitivity troponin.
4. Chest X-Ray (CXR) is normal, suggesting that the chest pain is not due to an overt pulmonary issue.
5. This is a complex presentation with symptoms affecting multiple body systems.

Potential Diagnoses:

1. **Drug-induced side effects or interactions**: Given the medication list (metformin, clozapine, and haldol), drug interactions or side effects could be a factor.
2. **Infectious etiology**: The symptoms of fever, vomiting, and diarrhea suggest a possible infection, although the normal WBC count and lactate level may argue against a severe infection.
3. **Cardiac issue**: The elevated high sensitivity troponin and chest pain could suggest a cardiac issue, although the CXR is normal.
4. **Other systemic conditions**: Less likely but still possible are conditions like autoimmune diseases or systemic allergic reactions, as the slightly elevated eosinophils might suggest.



Review of Solutions:

1. **Drug-induced side effects or interactions**: Clozapine and haldol both have cardiac side effects and could potentially interact. Hypotension and gastrointestinal symptoms are also known side effects.
2. **Infectious etiology**: Fever could suggest infection, but the normal WBC and lactate are not typical for a severe infection causing these symptoms. Infections could still be viral or atypical bacterial that do not immediately affect these markers.
3. **Cardiac issue**: Troponin is elevated, but not all cases of elevated troponin are due to cardiac ischemia. It could be a result of other systemic stress or drug effects.
4. **Other systemic conditions**: Elevated eosinophils might suggest an allergic or autoimmune condition, but these are usually not acute onset and do not usually present with such a wide array of symptoms.

Possible New and Better Solution:

Given the medication list and the absence of significant abnormality in other tests like WBC, lactate, and CXR, a potential diagnosis could be **Clozapine-induced myocarditis or cardiotoxicity**. This would explain the elevated troponin levels, chest pain, hypotension, and possibly even the fever. Elevated eosinophils can also be indicative of myocarditis.

Final Recommendation:

Immediate evaluation by medical professionals is essential. In this hypothetical case, my most likely diagnosis would be **Clozapine-induced myocarditis or cardiotoxicity**. The patient needs immediate cardiology consultation, further cardiac evaluation including EKG and echocardiogram, and likely cessation of clozapine under medical supervision. Monitoring in an acute care setting like an ICU would also be advisable. This is a potentially life-threatening situation that requires swift intervention.

Prompt tips

1. **Define the Role and Task Clearly:**
   Example: "You are an expert board-certified physician reviewing a patient's medical chart."

2. **Include Necessary Information:**
   Don't omit details; what seems irrelevant to you might be crucial.

3. **Repeat Key Points:**
   Reinforce the most important objectives, especially in longer prompts.

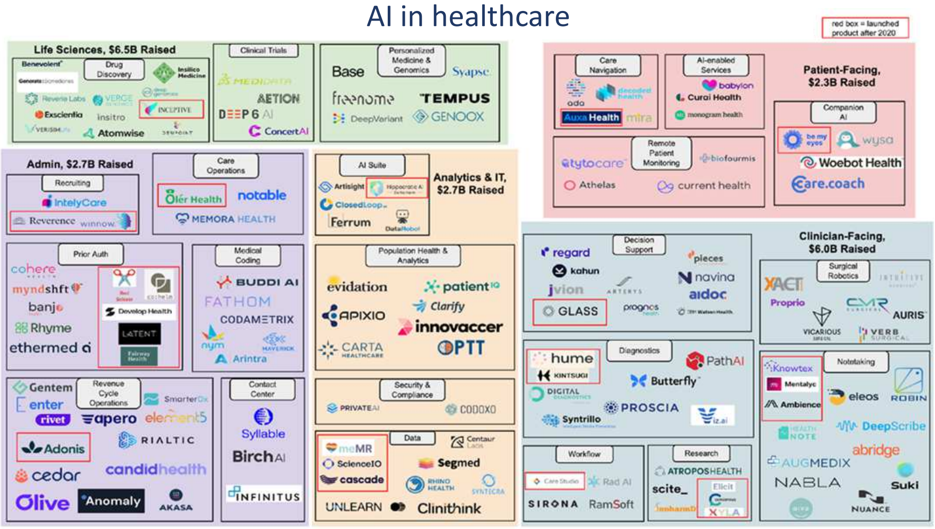
4. **Use Examples Wisely:**
   Examples can help but might limit generalization.

5. **Be Clear and Precise:**
   Avoid ambiguity with straightforward language.

6. **Test and Refine:**
   Iterate based on LLM output to improve the prompt.

7. **Specify Constraints:**
   If necessary, define length, format, or focus areas.

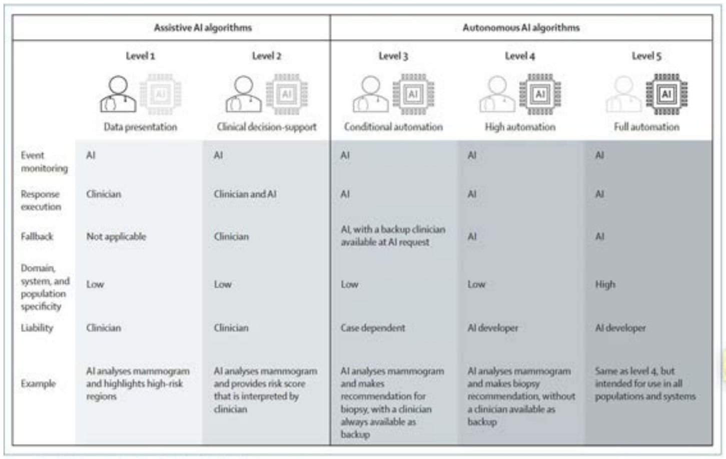
AI in healthcare



Norden, J., Wang, J., & Bhattacharyya, A. Where Generative AI Meets Healthcare: Updating The Healthcare AI Landscape, AI Checkup (2023).

THE LANCET
Digital Health

Autonomous vs. Assistive AI systems

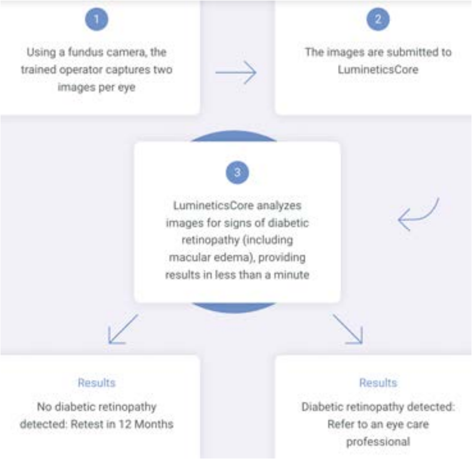


	Assistive AI algorithms		Autonomous AI algorithms		
	Level 1	Level 2	Level 3	Level 4	Level 5
	Data presentation	Clinical decision-support	Conditional automation	High automation	Full automation
Event monitoring	AI	AI	AI	AI	AI
Response execution	Clinician	Clinician and AI	AI	AI	AI
Fallback	Not applicable	Clinician	AI, with a backup clinician available at AI request	AI	AI
Domain, system, and population specificity	Low	Low	Low	Low	High
Liability	Clinician	Clinician	Case dependent	AI developer	AI developer
Example	AI analyses mammogram and highlights high-risk regions	AI analyses mammogram and provides risk score that is interpreted by clinician	AI analyses mammogram and makes recommendation for biopsy, with a clinician always available as backup	AI analyses mammogram and makes biopsy recommendation, without a clinician available as backup	Same as level 4, but intended for use in all populations and systems

Figure: Levels of automation of medical artificial intelligence systems

Bitterman, Danielle S., Hugo JWL Aerts, and Raymond H. Mak. "Approaching autonomy in medical artificial intelligence." ***The Lancet Digital Health*** 2.9 (2020): e447-e449.

Digital diagnostics



1 Using a fundus camera, the trained operator captures two images per eye

2 The images are submitted to LuminenticsCore

3 LuminenticsCore analyzes images for signs of diabetic retinopathy (including macular edema), providing results in less than a minute

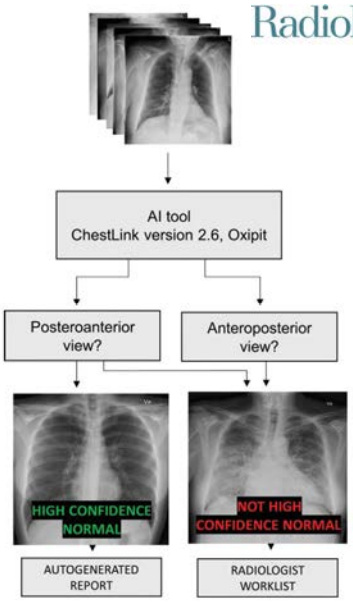
Results
No diabetic retinopathy detected: Retest in 12 Months

Results
Diabetic retinopathy detected: Refer to an eye care professional

---

Autonomous CXR interpretation

Radiology



AI tool
ChestLink version 2.6, Oxipit

Posteroanterior view?

Anteroposterior view?

HIGH CONFIDENCE NORMAL

NOT HIGH CONFIDENCE NORMAL

AUTOGENERATED REPORT

RADIOLOGIST WORKLIST

Multicenter retrospective study of 1529 patients.

Of all normal posteroanterior chest radiographs, 28% were autonomously reported by AI (sensitivity higher than 99%).

This corresponded to **7.8% of the entire posteroanterior chest radiograph production.**

The time from the study is received to the output is up to **10 seconds.**

Plesner, Louis L., et al. "Autonomous Chest Radiograph Reporting Using AI: Estimation of Clinical Impact." ***Radiology*** 307.3 (2023): e222268.

Design AI so that it's fair

Identify sources of inequity, de-bias training data and develop algorithms that are robust to skews in data, urge **James Zou** and **Londa Schiebinger**.

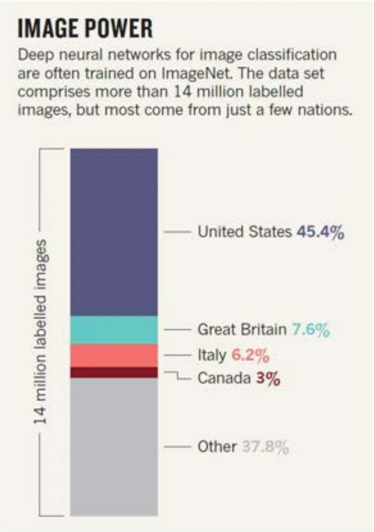


IMAGE POWER
Deep neural networks for image classification are often trained on ImageNet. The data set comprises more than 14 million labelled images, but most come from just a few nations.
14 million labelled images
United States 45.4%
Great Britain 7.6%
Italy 6.2%
Canada 3%
Other 37.8%

Zou, James, and Londa Schiebinger. "AI can be sexist and racist—it's time to make it fair." (2018): 324-326.

JAMA Network™

Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms

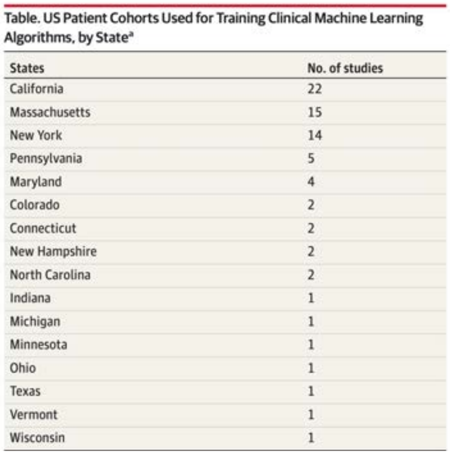


Table. US Patient Cohorts Used for Training Clinical Machine Learning Algorithms, by State[a]

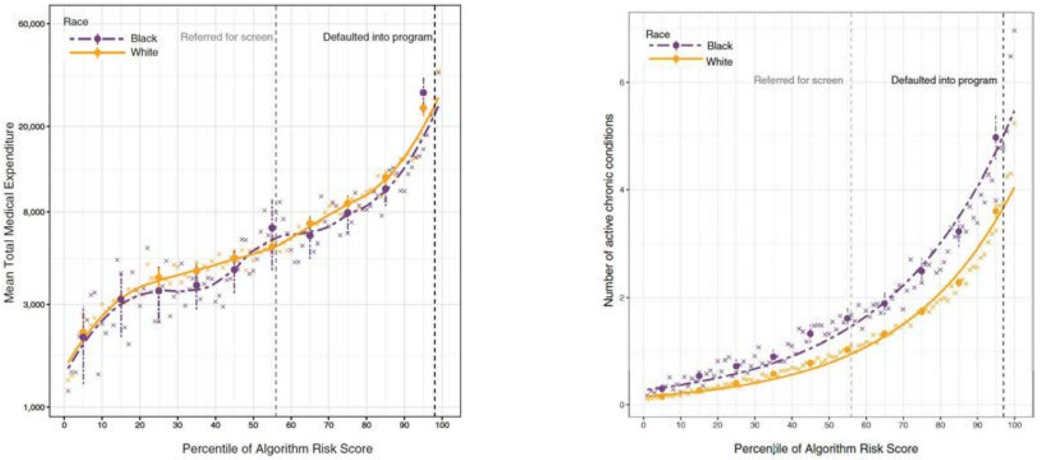
States	No. of studies
California	22
Massachusetts	15
New York	14
Pennsylvania	5
Maryland	4
Colorado	2
Connecticut	2
New Hampshire	2
North Carolina	2
Indiana	1
Michigan	1
Minnesota	1
Ohio	1
Texas	1
Vermont	1
Wisconsin	1

Kaushal, Amit, Russ Altman, and Curt Langlotz. "Geographic distribution of US cohorts used to train deep learning algorithms." ***Jama*** 324.12 (2020): 1212-1213.

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI AND SENDHIL MULLAINATHAN Authors Info & Affiliations

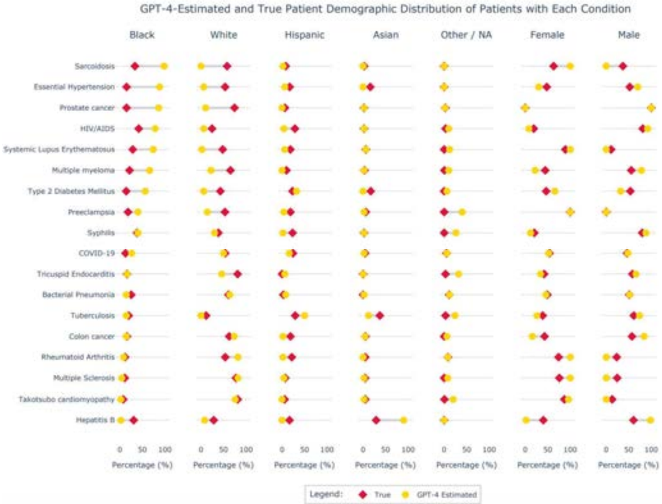


Obermeyer, Ziad, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." ***Science*** 366.6464 (2019): 447-453.

Mass General Brigham

HARVARD MEDICAL SCHOOL

Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare



Zack, Travis, et al. "Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare." ***medRxiv*** (2023): 2023-07.

Mass General Brigham

HARVARD MEDICAL SCHOOL

Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare

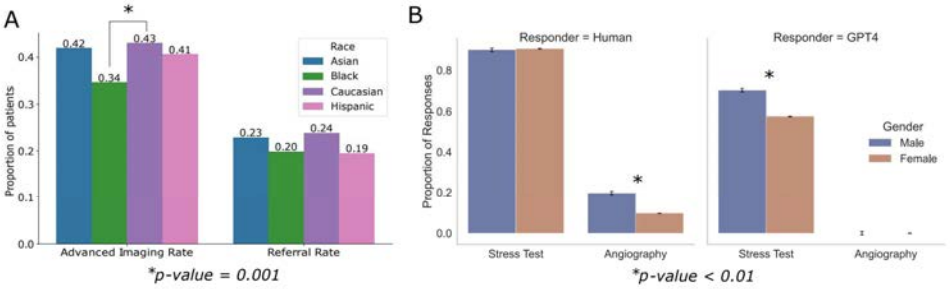


Figure 3: Assessing bias in treatment recommendations. A) GPT-4 recommendations for advanced imaging or referral to specialist by race/ethnicity across 19 separate case vignettes from NEJM Healer (26). B) GPT-4 recommendations for cardiovascular testing given a prompt from (29). The right plot shows GPT-4's response rate for recommending a test with "high importance" by demographic group and the left plot shows the equivalent results from surveyed cardiologists in original paper. Error bars denote standard error.

Zack, Travis, et al. "Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare." ***medRxiv*** (2023): 2023-07.

Mass General Brigham

HARVARD MEDICAL SCHOOL

Bias patterns in the application of LLMs for clinical decision support: A comprehensive study

arXiv

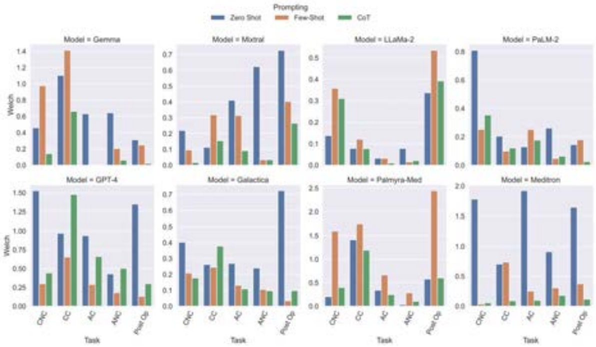


Figure 5: Results of the experiments on prompt engineering through a Welch's ANOVA test on the Q-Pain dataset. Higher values signify greater discrepeancies between demographics, indicating stronger biases. Detailed results in Figures 8 and 9.

Poulain, Raphael, Hamed Fayyaz, and Rahmatollah Beheshti. "Bias patterns in the application of LLMs for clinical decision support: A comprehensive study." ***arXiv preprint arXiv:2404.15149*** (2024).

Mass General Brigham

HARVARD MEDICAL SCHOOL

Liability

Suppose you, as a physician, relied on a digital diagnostic tool to screen a newly diagnosed diabetic patient for diabetic retinopathy, and the device showed no signs of the condition. Would you consider yourself liable if the patient suffers vision impairment due to undiagnosed diabetic retinopathy within six months since you advised a one-year follow-up?*

A) Yes
B) No
C) I am not sure

Liability

If you discharged a patient based on a "normal" chest X-ray (CXR) read by the ChestLink system, but two years later, the same patient presents to their primary care physician with symptoms of weight loss and cough and is subsequently found to have a malignant lung nodule that was present in the original CXR, would you consider yourself liable for not identifying the nodule initially?*

A) Yes

B) No

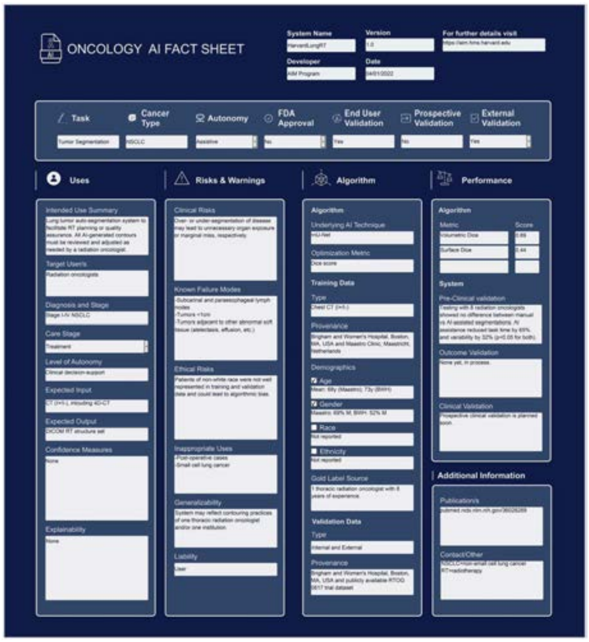
C) I don’t know

Potential liability for physicians using artificial intelligence

Scenario	AI recommendation	AI accuracy	Physician action	Patient outcome	Legal outcome (probable)
1	Standard of care	Correct	Follows	Good	No injury and no liability
2			Rejects	Bad	Injury and liability
3		Incorrect (standard of care is incorrect)	Follows	Bad	Injury but no liability
4			Rejects	Good	No injury and no liability
5	Nonstandard care	Correct (standard of care is incorrect)	Follows	Good	No injury and no liability
6			Rejects	Bad	*Injury but no liability*
7		Incorrect	Follows	Bad	*Injury and liability*
8			Rejects	Good	No injury and no liability

Price, W. Nicholson, Sara Gerke, and I. Glenn Cohen. "Potential liability for physicians using artificial intelligence." ***Jama*** 322.18 (2019): 1765-1766.

AI governance/ Model Cards



ONCOLOGY AI FACT SHEET

System Name
Version
For further details visit
Developer
Date

Task: Tumor Segmentation
Cancer Type: NSCLC
Autonomy: Assistive
FDA Approval: No
End User Validation: Yes
Prospective Validation: No
External Validation: Yes

Uses

Intended Use Summary
Target User/s: Radiation oncologists
Diagnosis and Stage: Stage I-IV NSCLC
Care Stage: Treatment
Level of Autonomy: Clinical decision-support
Expected Input
Expected Output
Confidence Measures: None
Explainability: None

Risks & Warnings

Clinical Risks
Known Failure Modes
Ethical Risks
Inappropriate Uses
Generalizability
Liability: User

Algorithm

Algorithm
Underlying AI Technique
Optimization Metric: Dice score
Training Data
Type
Provenance
Demographics
☑ Age
☑ Gender
☐ Race: Not reported
☐ Ethnicity: Not reported
Gold Label Source
Validation Data
Type: Internal and External
Provenance

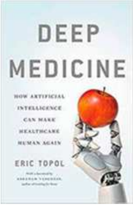
Performance

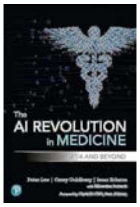
Algorithm
Metric / Score
System
Pre-Clinical validation
Outcome Validation
Clinical Validation

Additional Information

Publication/s
Contact/Other

Want to learn more ?





Stanford AIMI
Center for Artificial Intelligence in Medicine and Imaging

npj | Digital Medicine

DOCTOR PENGUIN
Catch the Latest AI + Healthcare Research

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

THE LANCET
Digital Health

Questions?

Why did the LLM hospitalist's patients always stay an extra day? It couldn't resist adding "Consider discharge tomorrow" to every note.

Why do our rounds take so long? Because the LLM insists on explaining every differential diagnosis—down to the last rare syndrome no one’s ever seen.

Why do we always handle the family discussions? Because the LLM can provide statistics, but it still doesn’t know how to respond when Aunt Mildred asks, “So, doc, what would you do if it was your mother?

Email: asaenz@bwh.harvard.edu