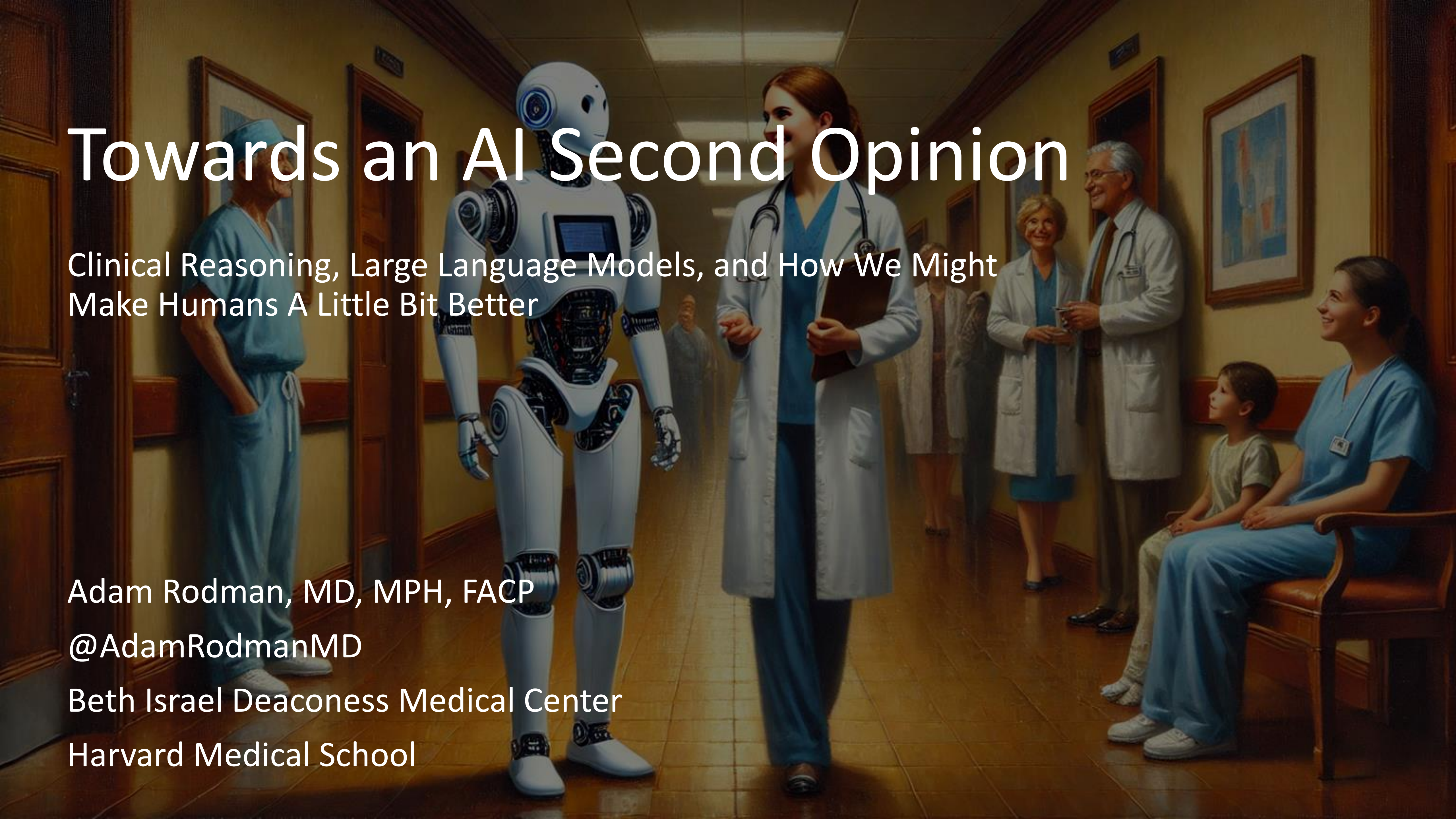


Towards an AI Second Opinion

Clinical Reasoning, Large Language Models, and How We Might Make Humans A Little Bit Better

Adam Rodman, MD, MPH, FACP

@AdamRodmanMD

Beth Israel Deaconess Medical Center

Harvard Medical School

Agenda for today

- What does it mean from a transtheoretical reasoning perspective to make a diagnosis? (6 minutes)
- Diagnostic errors, second opinions, and artificial intelligence (6 minutes)
- Live demonstration of an AI reasoning workflow with Dr. Jackson (20 minutes)
- Overview of the evidence, including my own research (10 minutes)
- Implications for current AI second opinion projects (2 minutes)
- Discussion/Reflections/Q&A

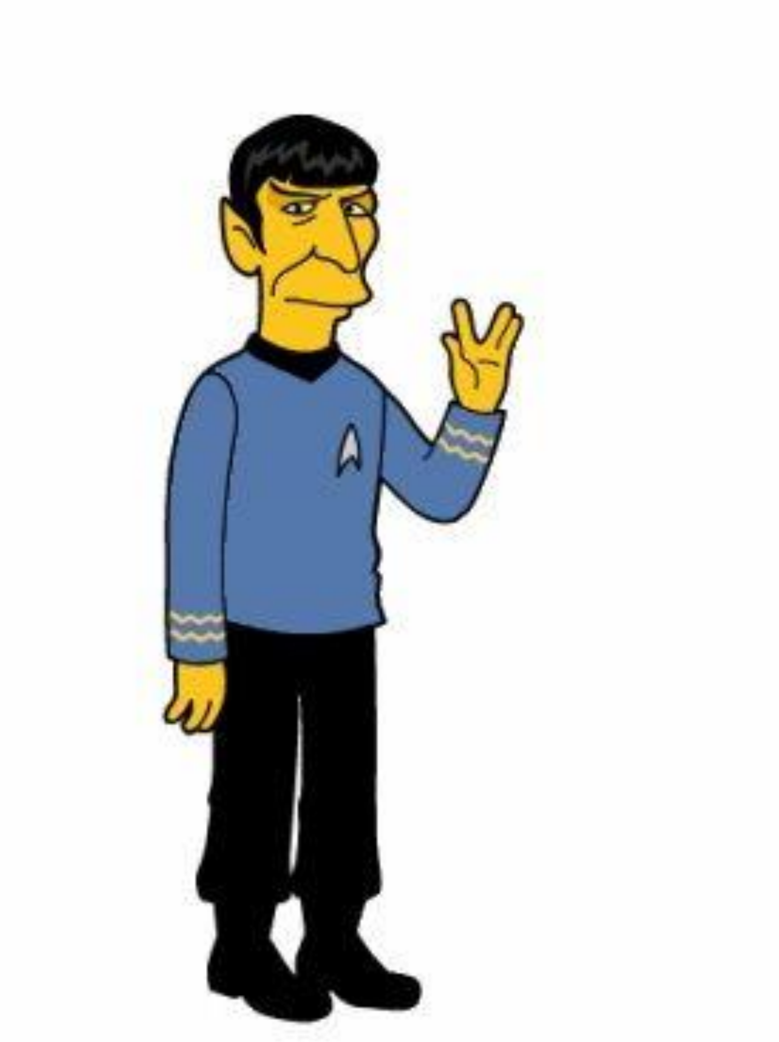
What is diagnosis anyway?

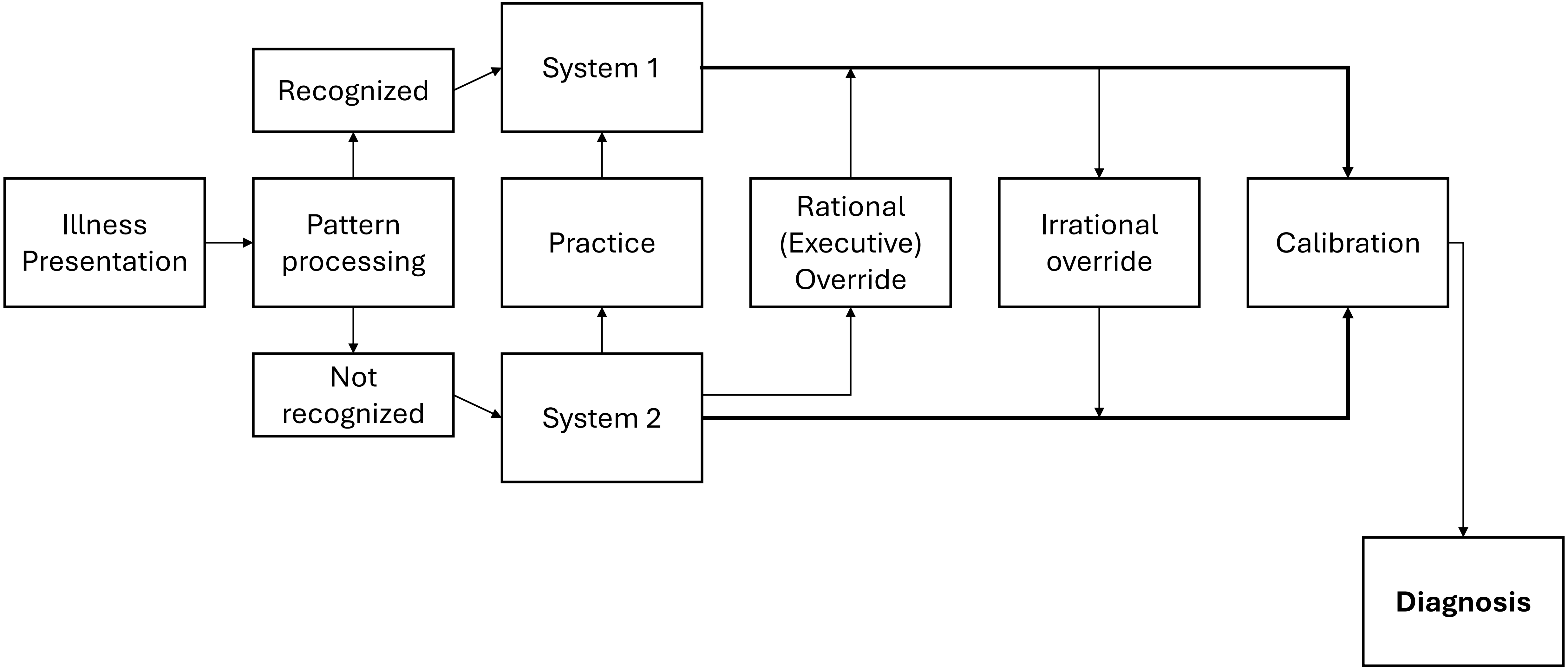
It's complicated…

Diagnosis = 1 / nosology

System 1

System 2





Illness Presentation

Pattern processing

Recognized

Not recognized

System 1

Practice

System 2

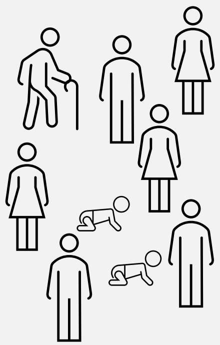
Rational (Executive) Override

Irrational override

Calibration

Diagnosis

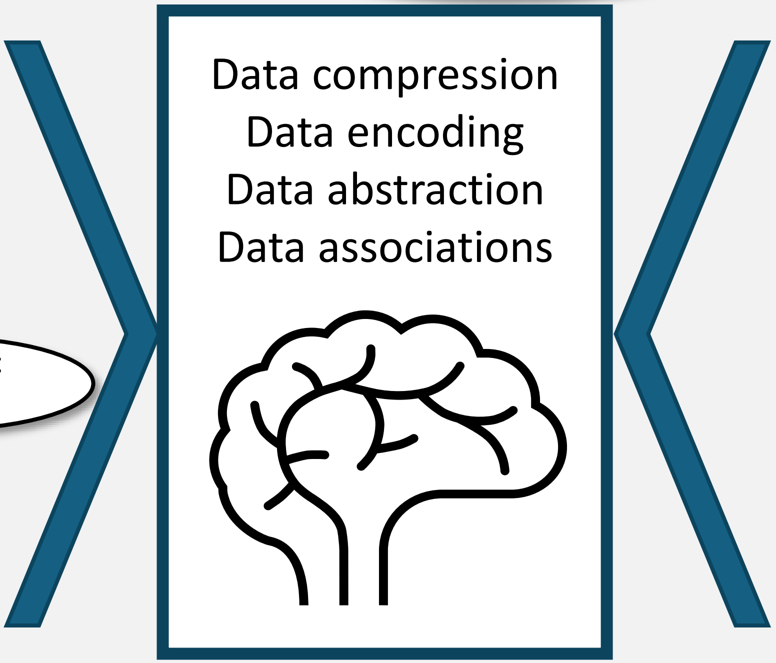
Script Theory
1. Knowledge Acquisition And Organization



Context

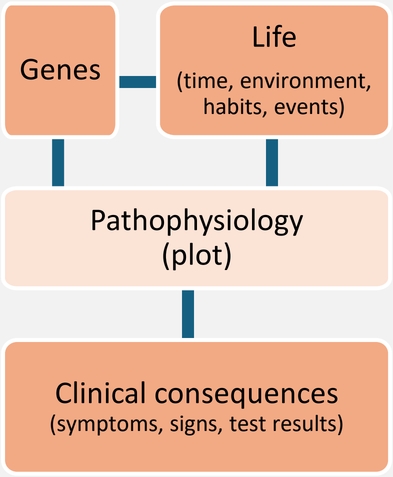
Personal bias

Systemic
bias



Data compression
Data encoding
Data abstraction
Data associations

Data-driven representation of diseases



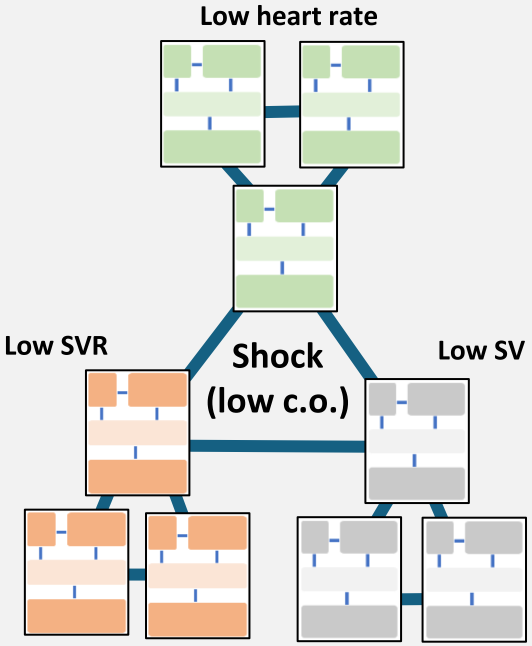
Genes

Life
(time, environment, habits, events)

Pathophysiology
(plot)

Clinical consequences
(symptoms, signs, test results)

Illness script



Low heart rate

Low SVR

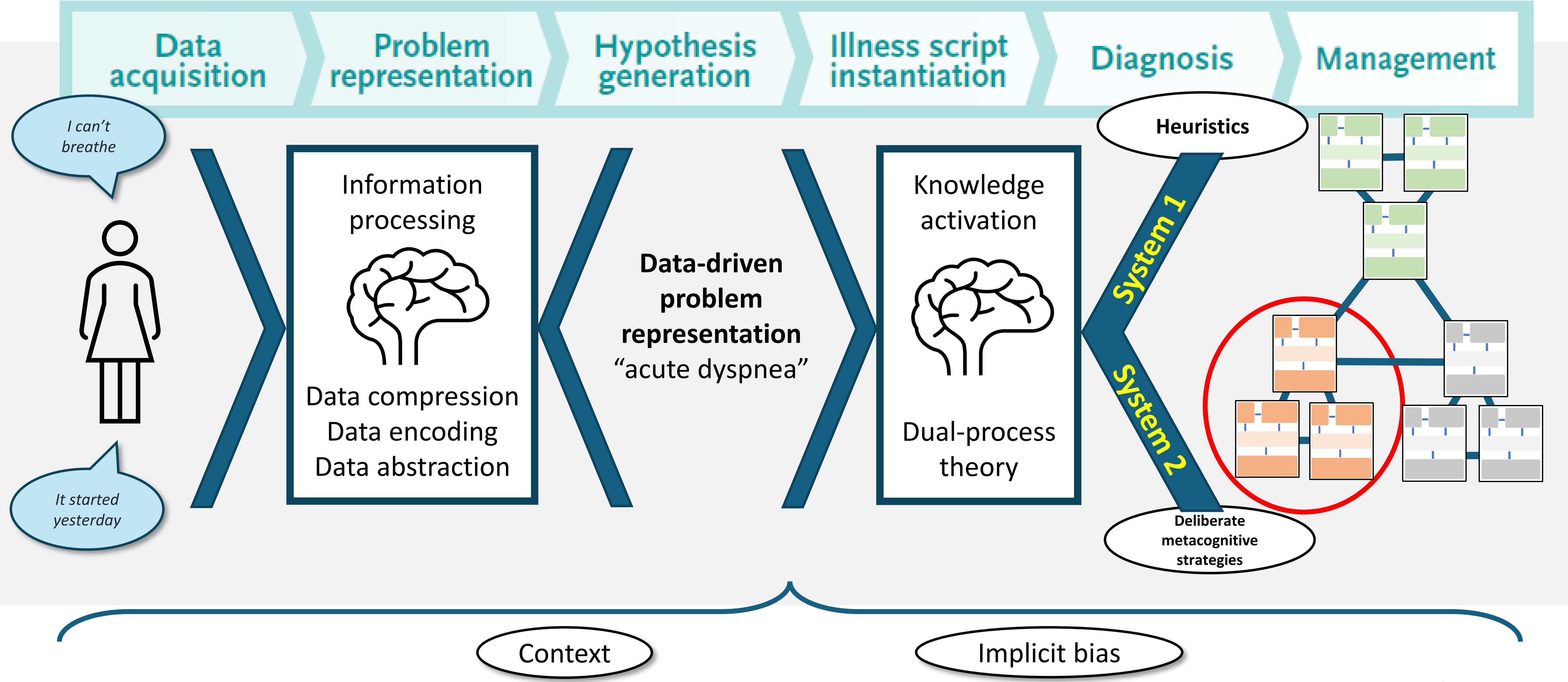
Shock
(low c.o.)

Low SV

Illness script network
“Diagnostic schemas”

Illness script network is limited by:
Body of knowledge (rapid doubling time), patients encountered, context, personal bias

Script Theory
2. Information Processing And Knowledge Activation



Data acquisition
Problem representation
Hypothesis generation
Illness script instantiation
Diagnosis
Management

I can't breathe

It started yesterday

Information processing

Data compression
Data encoding
Data abstraction

Data-driven problem representation
"acute dyspnea"

Knowledge activation

Dual-process theory

Heuristics

System 1

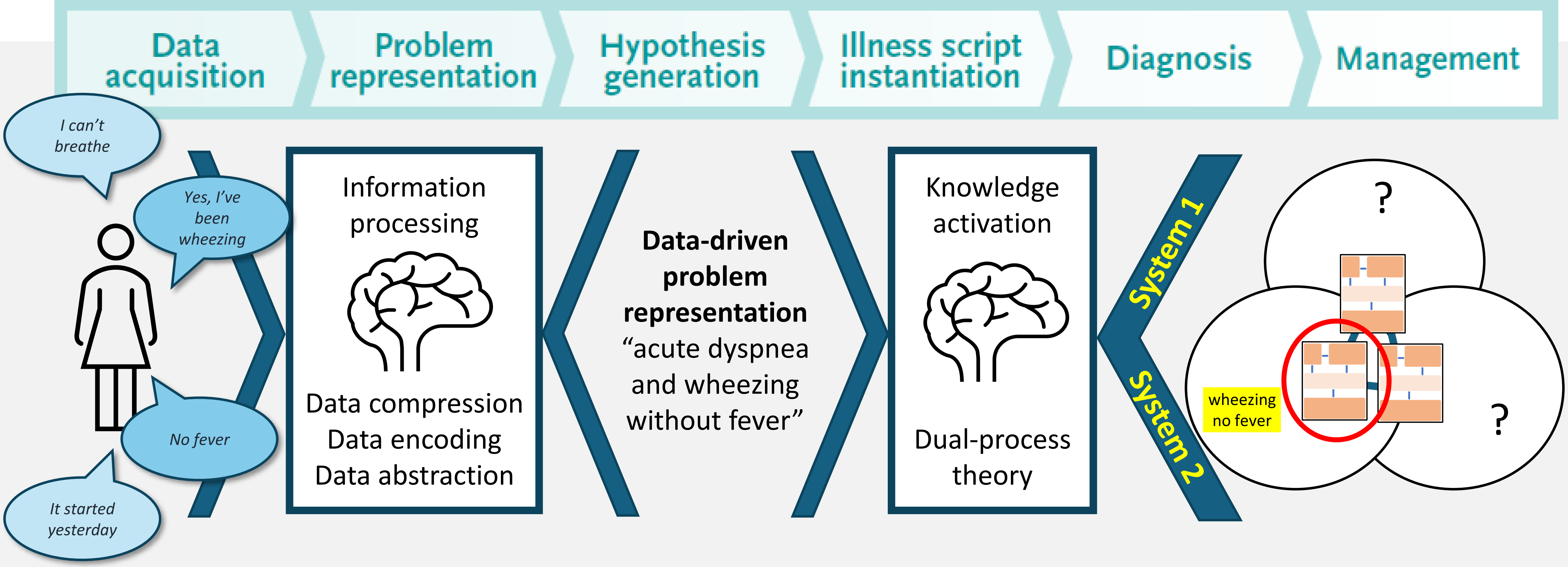
System 2

Deliberate metacognitive strategies

Context

Implicit bias

Script Theory
2. Information Processing And Knowledge Activation



Data acquisition | Problem representation | Hypothesis generation | Illness script instantiation | Diagnosis | Management

Information processing

Data compression
Data encoding
Data abstraction

Data-driven problem representation
"acute dyspnea and wheezing without fever"

Knowledge activation

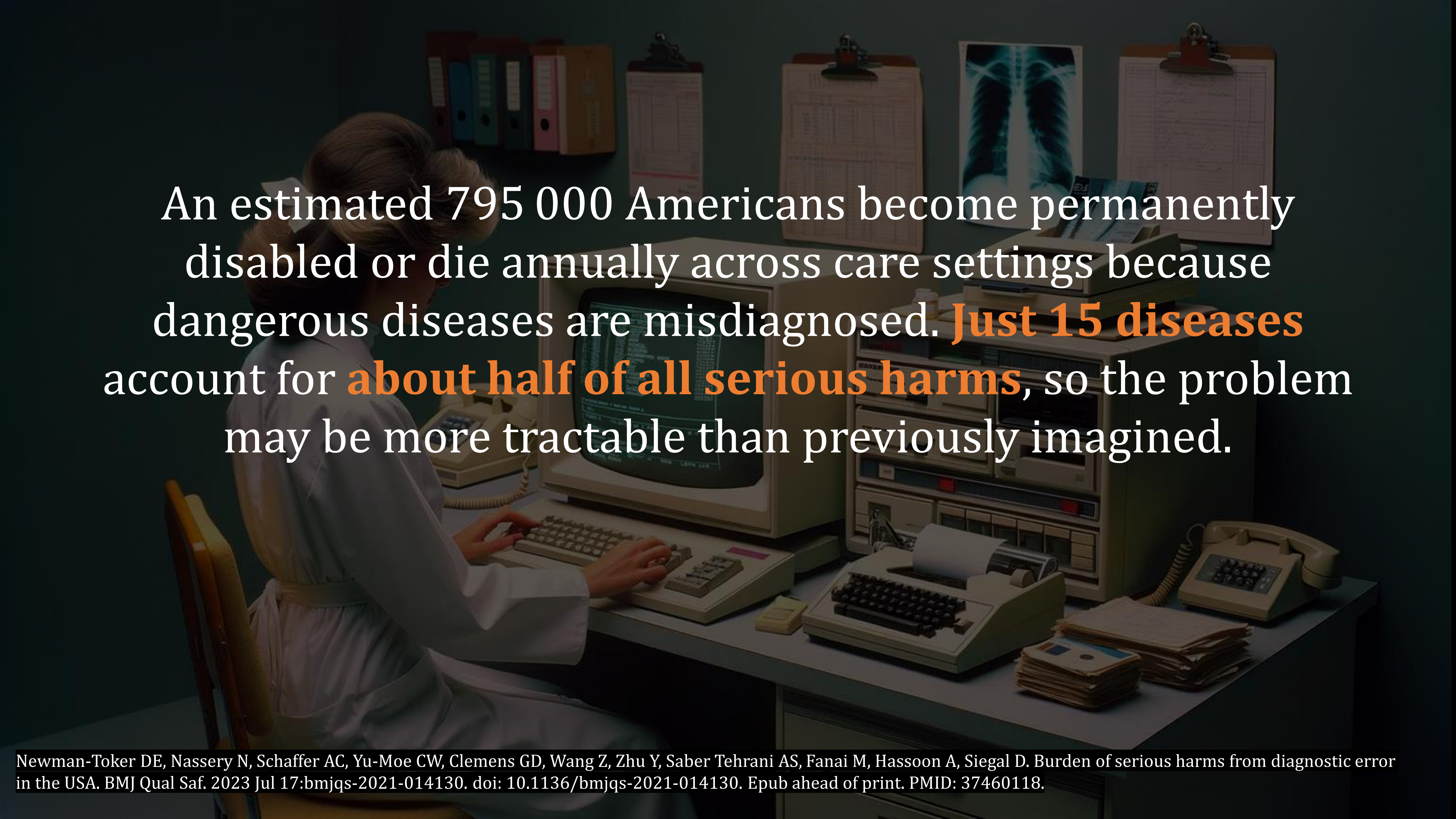
Dual-process theory

Clinical reasoning is limited by:
problem representation (data encoding, compression, and abstraction), prior knowledge (ISN), knowledge activation (recall) and its susceptibility to context, bias, and overreliance on heuristics

Ecological Psychology – Reasoning is a byproduct of interaction between clinician and surroundings/context

Situated cognition – Reasoning is subject to “context specificity” (aka situated) and fundamentally socially constrained

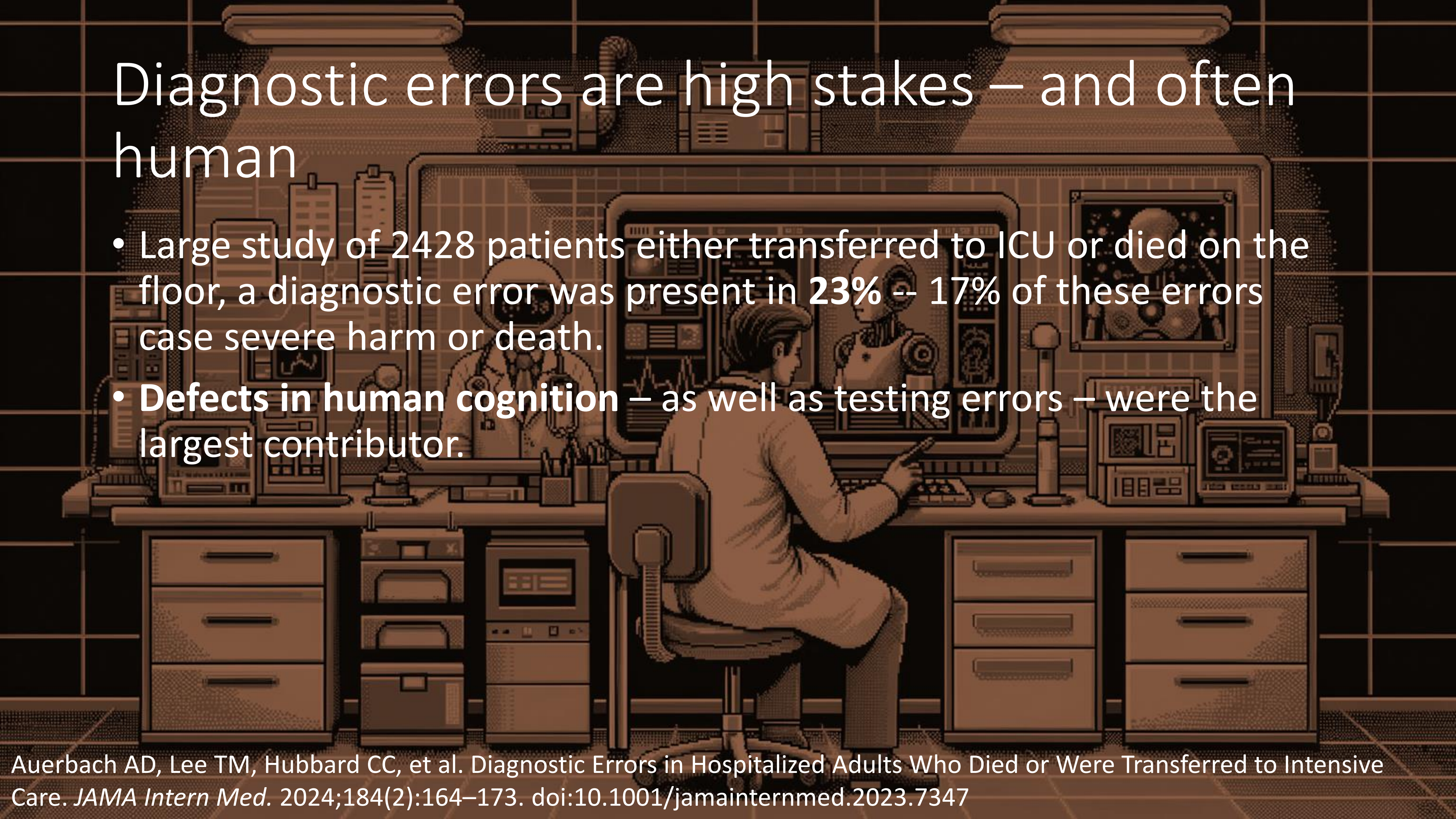
Distributed Cognition: cognition is fundamentally collaborative, spaced over multiple individuals separated by space and time, and tied together with systems such as the EHR



An estimated 795 000 Americans become permanently disabled or die annually across care settings because dangerous diseases are misdiagnosed. **Just 15 diseases** account for **about half of all serious harms**, so the problem may be more tractable than previously imagined.

Newman-Toker DE, Nassery N, Schaffer AC, Yu-Moe CW, Clemens GD, Wang Z, Zhu Y, Saber Tehrani AS, Fanai M, Hassoon A, Siegal D. Burden of serious harms from diagnostic error in the USA. BMJ Qual Saf. 2023 Jul 17:bmjqs-2021-014130. doi: 10.1136/bmjqs-2021-014130. Epub ahead of print. PMID: 37460118.

Diagnostic errors are high stakes – and often human



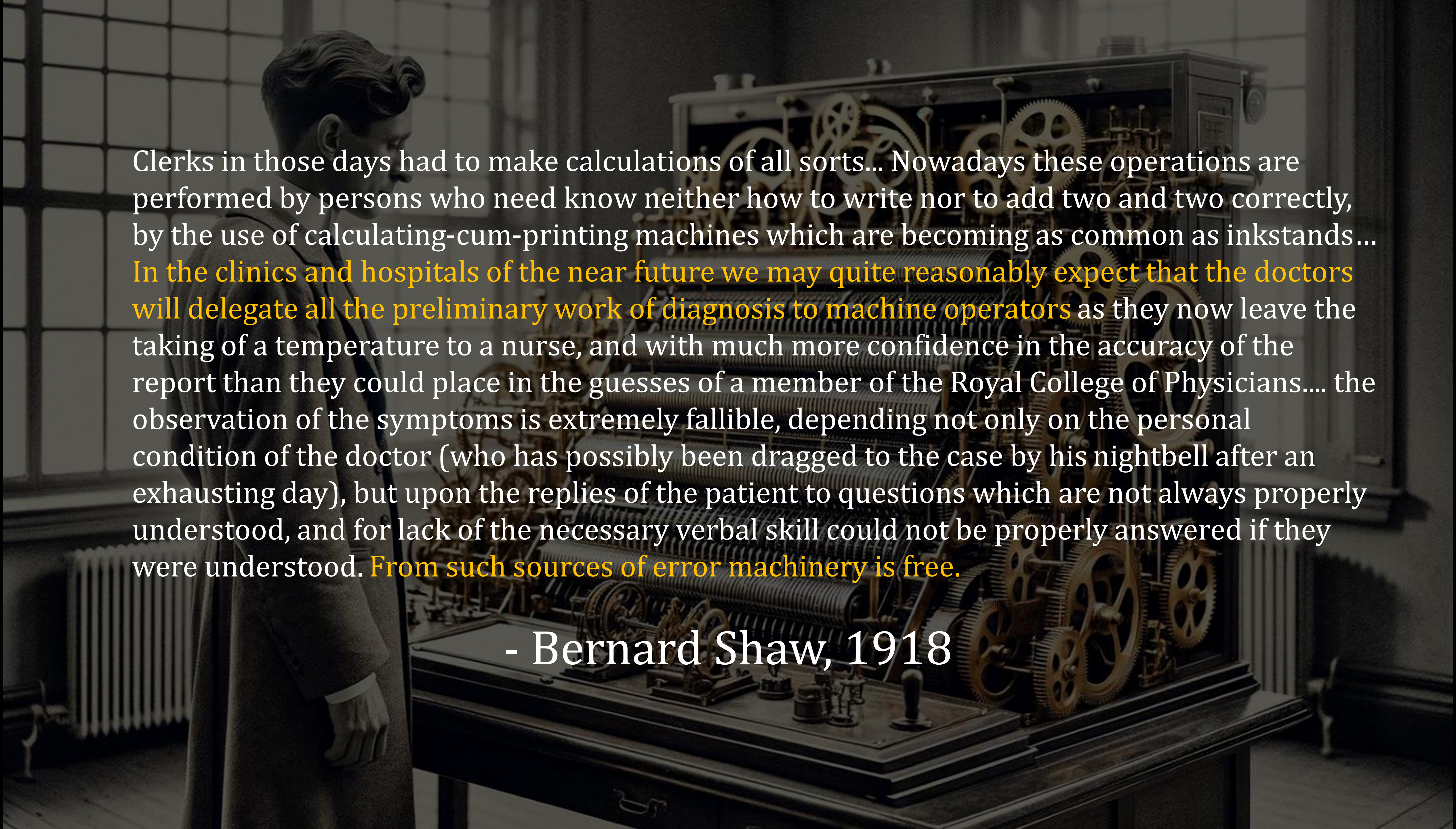
- Large study of 2428 patients either transferred to ICU or died on the floor, a diagnostic error was present in **23%** -- 17% of these errors case severe harm or death.
- **Defects in human cognition** – as well as testing errors – were the largest contributor.

Auerbach AD, Lee TM, Hubbard CC, et al. Diagnostic Errors in Hospitalized Adults Who Died or Were Transferred to Intensive Care. *JAMA Intern Med.* 2024;184(2):164–173. doi:10.1001/jamainternmed.2023.7347

Methods to improve reasoning

- Three overall categories of interventions:
  1. Education about cognitive biases
  2. Education about debiasing strategies
  3. **Artificial intelligence**
- Effect sizes of interventions have been modest and only in experimental settings

Prakash S, Sladek RM, Schuwirth L. Interventions to improve diagnostic decision making: A systematic review and meta-analysis on reflective strategies. Med Teach. 2019 May;41(5):517-524. doi: 10.1080/0142159X.2018.1497786. Epub 2018 Sep 23. PMID: 30244625.



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

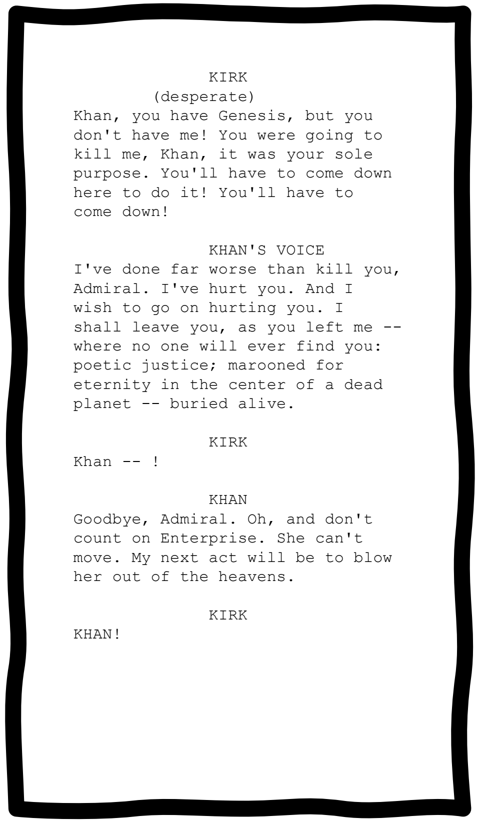
- Bernard Shaw, 1918

Artificial intelligence to improve reasoning

- Historically (~1946-1992) AI was seen as a solution, with large effective sizes, though in limited domains (AAPHELP, MYCIN, INTERNIST-I)
- Studies on modern diagnostic AI (“differential generators”) experimentally promising, but little real-world impact (ISABEL, Dxplain)
- LLMs are the most exciting intervention in clinical reasoning in decades.

Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med. 2012 Feb;27(2):213-9. doi: 10.1007/s11606-011-1804-8. PMID: 21789717; PMCID: PMC3270234.

Language Models



KIRK
(desperate)
Khan, you have Genesis, but you don't have me! You were going to kill me, Khan, it was your sole purpose. You'll have to come down here to do it! You'll have to come down!

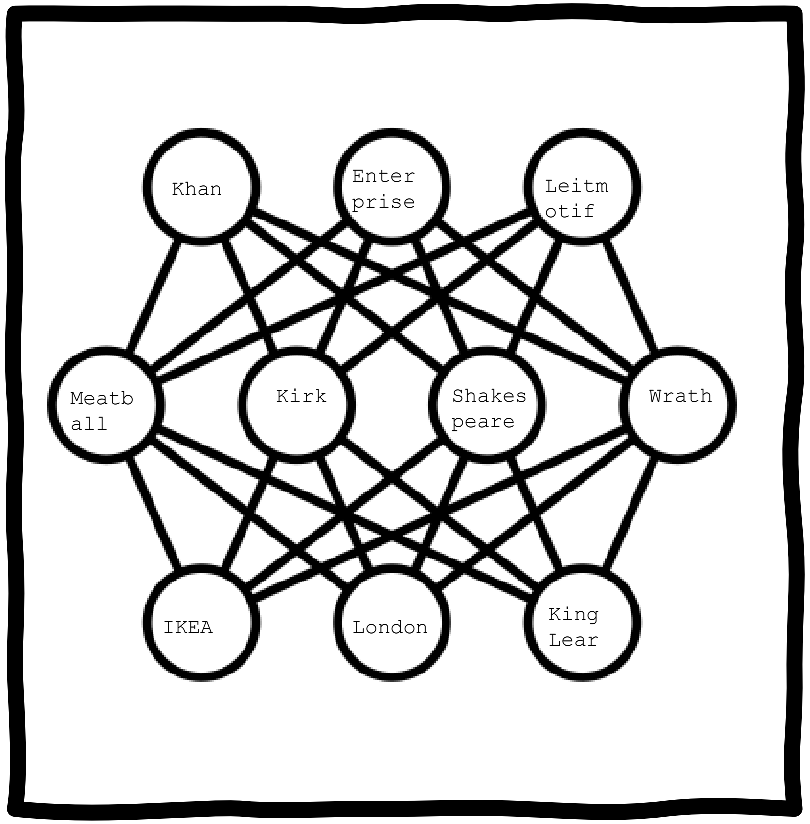
KHAN'S VOICE
I've done far worse than kill you, Admiral. I've hurt you. And I wish to go on hurting you. I shall leave you, as you left me -- where no one will ever find you: poetic justice; marooned for eternity in the center of a dead planet -- buried alive.

KIRK
Khan -- !

KHAN
Goodbye, Admiral. Oh, and don't count on Enterprise. She can't move. My next act will be to blow her out of the heavens.

KIRK
KHAN!

Massive amounts of human-generated text

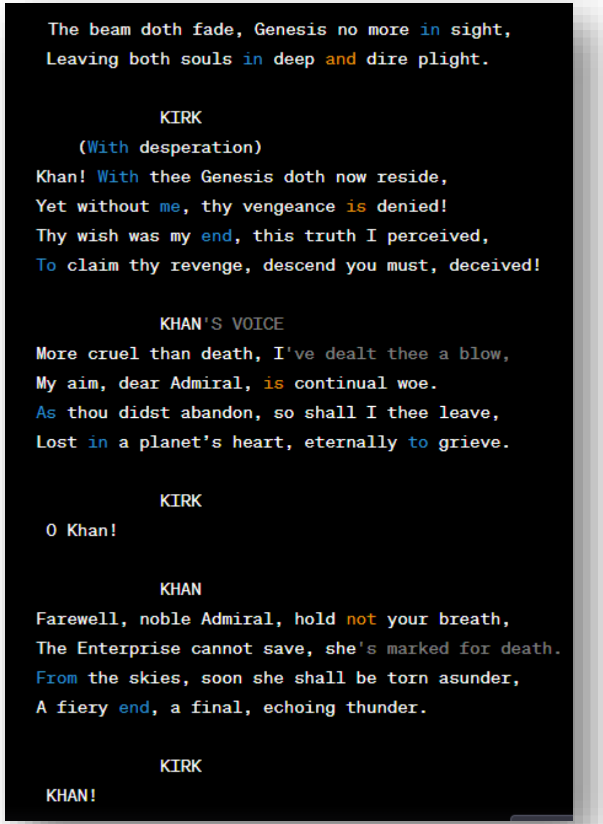


Word (token) associations (pre-training)

Fine-tuning (RLHF)

You are a Shakespeare expert, very gifted in writing in the Bard's style. I am putting on a Shakespeare festival for a local community theater. I want to stage Star Trek II: The Wrath of Khan in the style of Shakespeare -- that is, written in iambic pentameter such that it would be understandable to the audience in Shakespeare's time. I am going to give you a section of the script that I am having particular trouble with. Can you rewrite this such that it is in the style of Shakespeare?

Prompt (input)

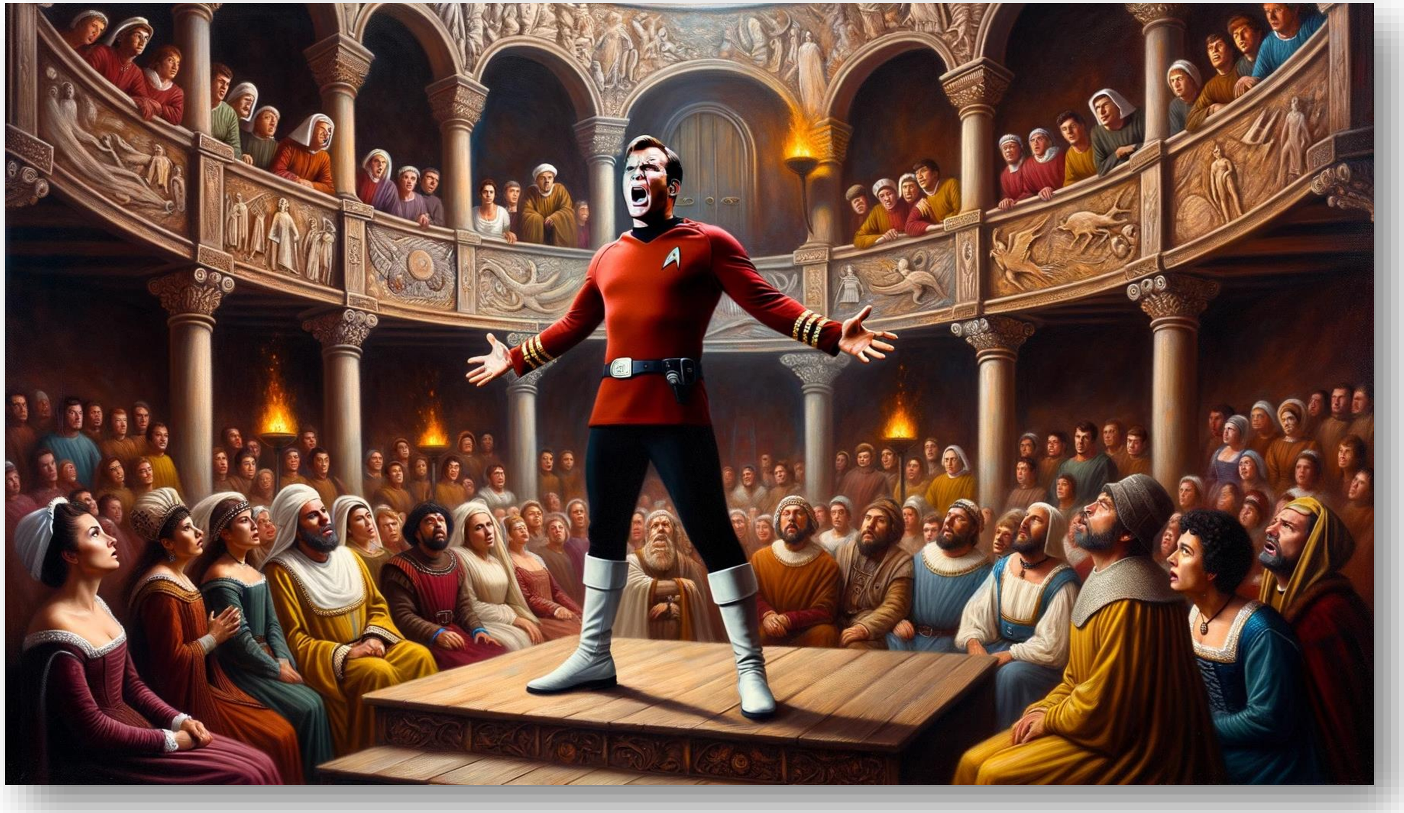


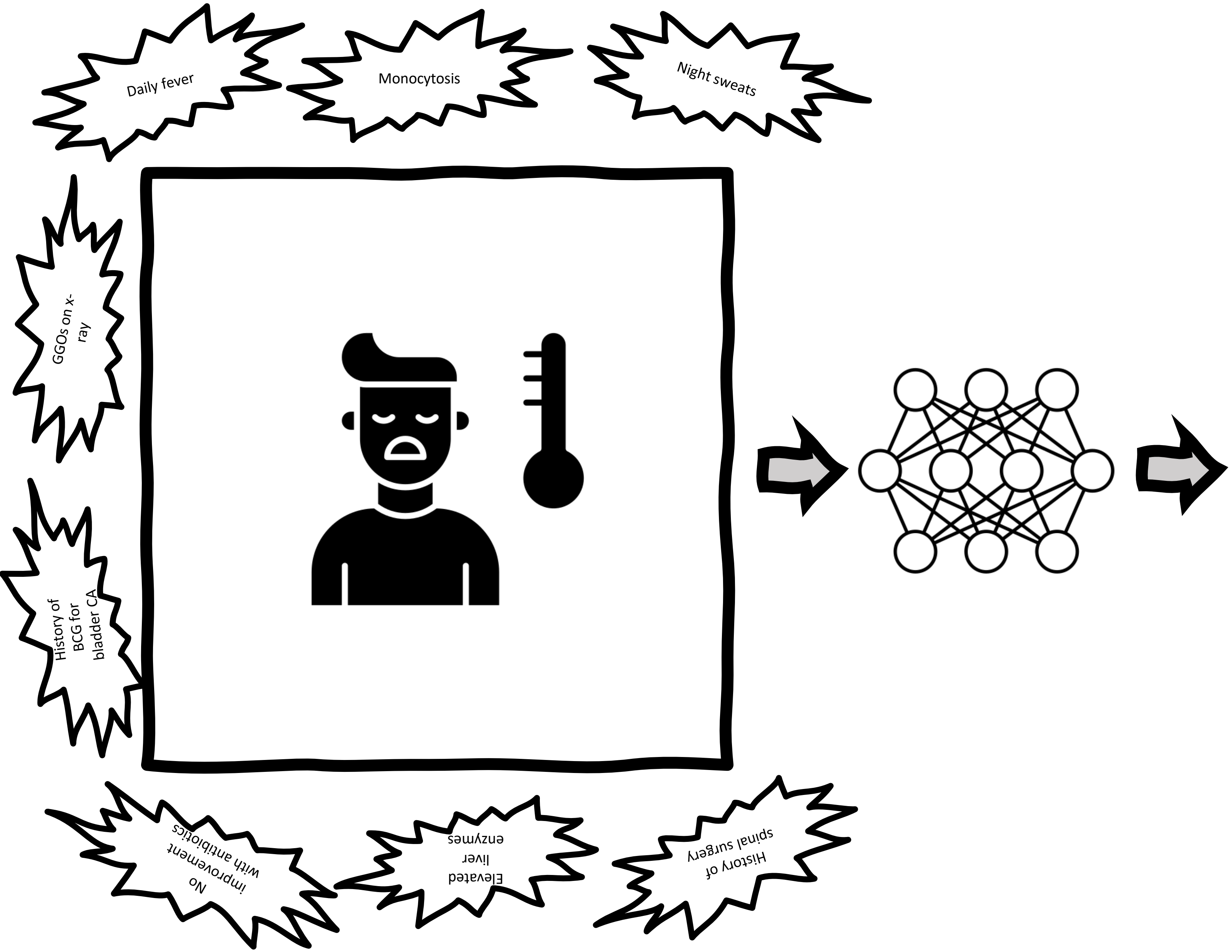
```
 The beam doth fade, Genesis no more in sight,
 Leaving both souls in deep and dire plight.

           KIRK
    (With desperation)
Khan! With thee Genesis doth now reside,
Yet without me, thy vengeance is denied!
Thy wish was my end, this truth I perceived,
To claim thy revenge, descend you must, deceived!

           KHAN'S VOICE
More cruel than death, I've dealt thee a blow,
My aim, dear Admiral, is continual woe.
As thou didst abandon, so shall I thee leave,
Lost in a planet's heart, eternally to grieve.

           KIRK
 O Khan!

           KHAN
Farewell, noble Admiral, hold not your breath,
The Enterprise cannot save, she's marked for death.
From the skies, soon she shall be torn asunder,
A fiery end, a final, echoing thunder.

           KIRK
 KHAN!
```

Output: Simulacra of human text

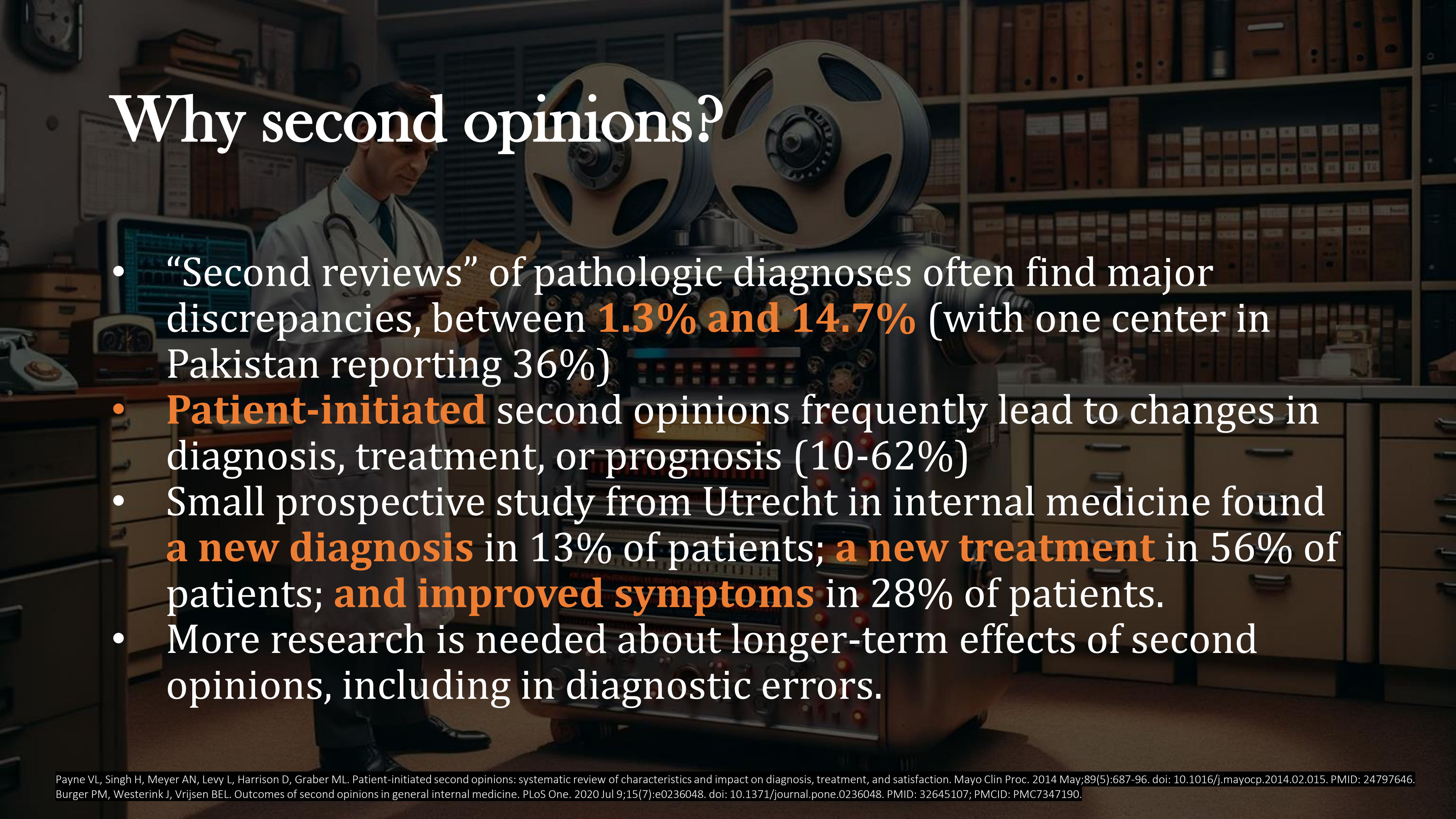
KHAAAN!
PYRAMID AMERICA
TM & 2019 CBS Studios Inc. 2019 Paramount Pictures Corp. All Rights Reserved.

Renaissance-style oil painting of Captain Kirk in Starfleet uniform, passionately exclaiming 'KHAN'. He stands on a stage of a theater reminiscent of the middle ages, with intricate carvings, stone archways, and burning torches. Medieval spectators gaze at him with a mix of shock and curiosity.




Daily fever
Monocytosis
Night sweats
GGOs on x-ray
History of BCG for bladder CA
No improvement with antibiotics
Elevated liver enzymes
History of spinal surgery

1. **Disseminated Tuberculosis** - 20%: Given his travel history and the clinical presentation, this remains a possibility, though it would be uncommon in the absence of a positive imaging or microbiological evidence.
2. **Endocarditis** - 15%: Even though he has a history of ischemic heart disease and a dual-chamber ICD, there's no clear evidence from the echo. Still, clinical suspicion remains for possible subclinical or atypical endocarditis.
3. **Lymphoma** - 30%: Chronic fevers, weight loss, elevated LDH, and splenomegaly all support this diagnosis. The PET/CT findings of diffuse ground glass opacities bilaterally could also be supportive.
4. **Chronic Q Fever** - 10%: Given the history of travel to endemic areas, this remains a differential. It can present subacutely with prolonged fevers and can affect the liver.
5. **Chronic Hepatitis (possibly Autoimmune)** - 10%: Elevated transaminases and AST/ALT ratio is suggestive of liver injury. There's also an increased alkaline phosphatase. Still, other indicators of autoimmune hepatitis like ANA or anti-smooth muscle antibody were not mentioned.
6. **Histoplasmosis** - 5%: Endemic in some parts of Central America, and can present with similar features, though the imaging findings don't strongly support this.
7. **Chronic Ehrlichiosis/Anaplasmosis** - 3%: He has travel history to regions where tick-borne illnesses are common, and the presentation could align with a chronic form of ehrlichiosis/anaplasmosis.
8. **Fungal Infection (e.g., Coccidioidomycosis, Blastomycosis)** - 2%: Given the presentation and travel history, but again, the imaging findings don't strongly suggest this.
9. **Leptospirosis** - 3%: This is an important consideration given his travel history, although the presentation would be atypical for leptospirosis to persist this long without severe kidney or liver impairment.
10. **Spinal Hardware Infection** - 2%: He has a history of multiple spinal surgeries, injections, and a spinal cord stimulator placement. While the nuclear-tagged WBC scan showed no elevated uptake in the spine, hardware-associated infections can sometimes be subtle.



Why second opinions?

- "Second reviews" of pathologic diagnoses often find major discrepancies, between **1.3% and 14.7%** (with one center in Pakistan reporting 36%)
- **Patient-initiated** second opinions frequently lead to changes in diagnosis, treatment, or prognosis (10-62%)
- Small prospective study from Utrecht in internal medicine found **a new diagnosis** in 13% of patients; **a new treatment** in 56% of patients; **and improved symptoms** in 28% of patients.
- More research is needed about longer-term effects of second opinions, including in diagnostic errors.

Payne VL, Singh H, Meyer AN, Levy L, Harrison D, Graber ML. Patient-initiated second opinions: systematic review of characteristics and impact on diagnosis, treatment, and satisfaction. Mayo Clin Proc. 2014 May;89(5):687-96. doi: 10.1016/j.mayocp.2014.02.015. PMID: 24797646.
Burger PM, Westerink J, Vrijsen BEL. Outcomes of second opinions in general internal medicine. PLoS One. 2020 Jul 9;15(7):e0236048. doi: 10.1371/journal.pone.0236048. PMID: 32645107; PMCID: PMC7347190.

Reasoning Prompt

You are an emergency room physician caring for a patient who has just presented to the emergency department. I am going to give you information from the case in real time. After each section of the case, I want you to list your full differential diagnosis. Under each item on the differential, I want you to list all the factors in the history thus far that are consistent with your diagnosis, all the factors in the history thus far that are not consistent with the diagnosis, and the likelihood of each diagnosis, expressed as a percentage. After you have your list of diagnoses, please list the diagnostic tests you would like to order next.

ED RN triage

At ED Triage:

CC: Chest pain, tachycardia

Triage: Patient reports a new PE diagnosis and LLE DVT at OSH ED five days ago. Put on Eliquis 10mg BID since four days ago. Patient now arrives here with worsening chest pain, cough, and tachycardia.

Severity: 1 (highest)

Triage Vitals: T 101 HR 140 RR 26 BP 111/86 SPO2 99% O2 Device None

MD triage

Pt is a young female presenting to the emergency department for chest pain and worsening shortness of breath. Patient notes that she has a history of lupus, was diagnosed with a pulmonary embolism five days ago after having a CT for shortness of breath and chest pain. She was started on Eliquis 10 mg twice daily which she has been taking regularly. Today, she had worsening shortness of breath, chest pain, and worsened palpitations which prompted her to present to the emergency department. She was triggered for tachycardia to the 140s. Blood pressure normal, vitals notable for tachycardia from 1 40-1 60. Administered IV fluids, patient on monitor.

Physical Exam

Vitals: Normotensive, tachycardic to 140 (range 119-140), febrile to 101 °F.
Gen: Well-appearing, NAD
HEENT: Normocephalic, atraumatic, PERLA, EOMI.
Neck: Supple, no c-spine tenderness to palpation
CV: Tachycardic rate, no murmurs noted.
Resp: Lungs CTAB, no wheezes or crackles.
Abd: Soft, non-distended, non-tender to palpation. No rigidity, rebound, or guarding.
Skin: Warm, dry, intact
Ext: No lower extremity edema, erythema, or tenderness to palpation
Neuro: A&Ox3, CN II-XII intact, 5/5 strength in all extremities, sensation intact and symmetric
Psych: Appropriate mood and affect

Chest X-ray:

There are bibasilar opacities compatible with likely small bilateral pleural effusions and likely atelectasis. Component of infarct can not be excluded given patient's history of pulmonary emboli. Superiorly, the lungs are clear. The cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities.

CTA Chest

1. Right lower lobar and segmental pulmonary emboli without evidence of right heart strain.
2. Small pericardial effusion. Trace left pleural effusion.
3. Prominent bilateral axillary lymph nodes measuring up to 1.1 cm on the right in short axis. This may be reactive, but recommend ultrasound follow-up in 4-6 weeks is recommended for further evaluation.
4. 2.7 cm indeterminate hepatic hypodensity, possibly either a cyst or hemangioma. Recommend nonemergent outpatient ultrasound for improved evaluation.

TTE

Normal left ventricular wall thickness and mass, biventricular cavity sizes, and hyperdynamic regional/global biventricular systolic function. Reduced global longitudinal strain. Normal diastolic function. No definite valvular pathology or pathologic flow identified. Indeterminate pulmonary
artery systolic pressure due to insufficient tricuspid regurgitation (though the normal pulmonary artery acceleration time suggests a normal pulmonary artery systolic pressure).

No prior study available for comparison.

Labs:

Lactic Acid 1.7

Sodium 131, Potassium 4.2, Chloride 100, CO2 18, Anion Gap 13, BUN 7, Creatinine 0.60, Estimated GFR(CKD-EPI) >120, Glucose 129

Calcium 8.0

Total Protein 6.6, Albumin, Blood 3.3, Globulin 3.3, AST (SGOT) 30, ALT (SGPT) 23, Alkaline Phosphatase 35, Total Bilirubin 0.5, Magnesium, Blood 1.9

hs-Troponin T 6
NT-ProBNP 53

Labs:

WBC 6.58, Hemoglobin 9.1, Hematocrit 27.8, Platelet Count 289

MCV 81, MCH 26.5, MCHC 32.7, RDW 16.0, RDW-SD 46.3, MPV 9.4, Nucleated RBC 0

Neutrophil 89.8, Immature Granulocyte (Meta, Myelo, Promyelocyte) 0.6, Lymphocyte 6.5, Monocyte 2.7, Eosinophil 0.2, Basophil 0.2, Absolute Neutrophil Count 5.91, Absolute Immature Granulocyte (Meta, Myelo, Promyelocyte) 0.04, Absolute Lymphocyte Count 0.43, Absolute Monocyte Count 0.18, Absolute Eosinophil Count 0.01, Absolute Basophil Count 0.01

PTT 35, INR1.7, Prothrombin Time, 20.0

Labs:

Preg Test, Ur Negative
Color, UA Straw , Clarity, UA Clear, Specific Gravity, UA 1.009, pH, UA 7.0, Protein, UA Negative, Glucose, UA Negative, Ketones, UA Negative, Blood, UA Negative, Nitrite, UA Negative, Leukocytes, UA Negative, Bilirubin, Urine Negative, Urobilinogen, Urine Normal

WBC, UA 2, RBC, UA <1, Bacteria, UA Few, Budding Yeast Rare, Squamous Epithelial Cells 0-2, Mucous Threads Rare

Repeat: Lactic Acid 1.5
Repeat: hs-Troponin T 8

Additional history

For context, this patient was diagnosed with PE after presenting to outside hospital a few days ago for ~1 week of pleuritic chest pain and dyspnea, found to have segmental/subsegmental PE (involving RLE, RML, RUL, LLL) on CTA with no evidence of RHS or large pericardial effusion, LENI w/ L femoral/popliteal/gastroc DVT, hs-trop 4, and BNP 75. She was discharged on apixaban load (10mg BID x 1 week, then 5mg BID) and has been taking it with no missed doses.

Since that hospitalization, she reports mild progression of her symptoms with increased pleuritic chest discomfort and dyspnea. She had a few seconds of dizziness in the shower today, but no other episodes of dizziness/LH or syncope. She has noticed an elevated HR on her wearable device but no significant palpitations. Of note, she has also had ongoing fevers (recently up to 100-101F range) since end of a few months ago in setting of her active lupus, for which she follows here for rheumatology with ongoing titration of her immunosuppressive regimen. She was on OCPs which were stopped last week after diagnosis of PE, no other pro-thrombotic medications. No prior personal or family history of VTE or clotting disorders. The patient otherwise denies orthopnea, PND, LE edema, unexpected weight change, syncope or transient neurologic symptoms.

Per chart review, patient with similar constellation of symptoms (sinus tach, pleuritic chest pain) for which she was evaluated by cardiology 6 years back. She also was noted to have a very small pericardial effusion at the time. Her symptoms lasted only a short time without intervention and thus cardiology signed off, with the contingency that should her symptoms return causing ED admission, could consider cardiac MRI for eval of cardiac inflammation and further work-up.

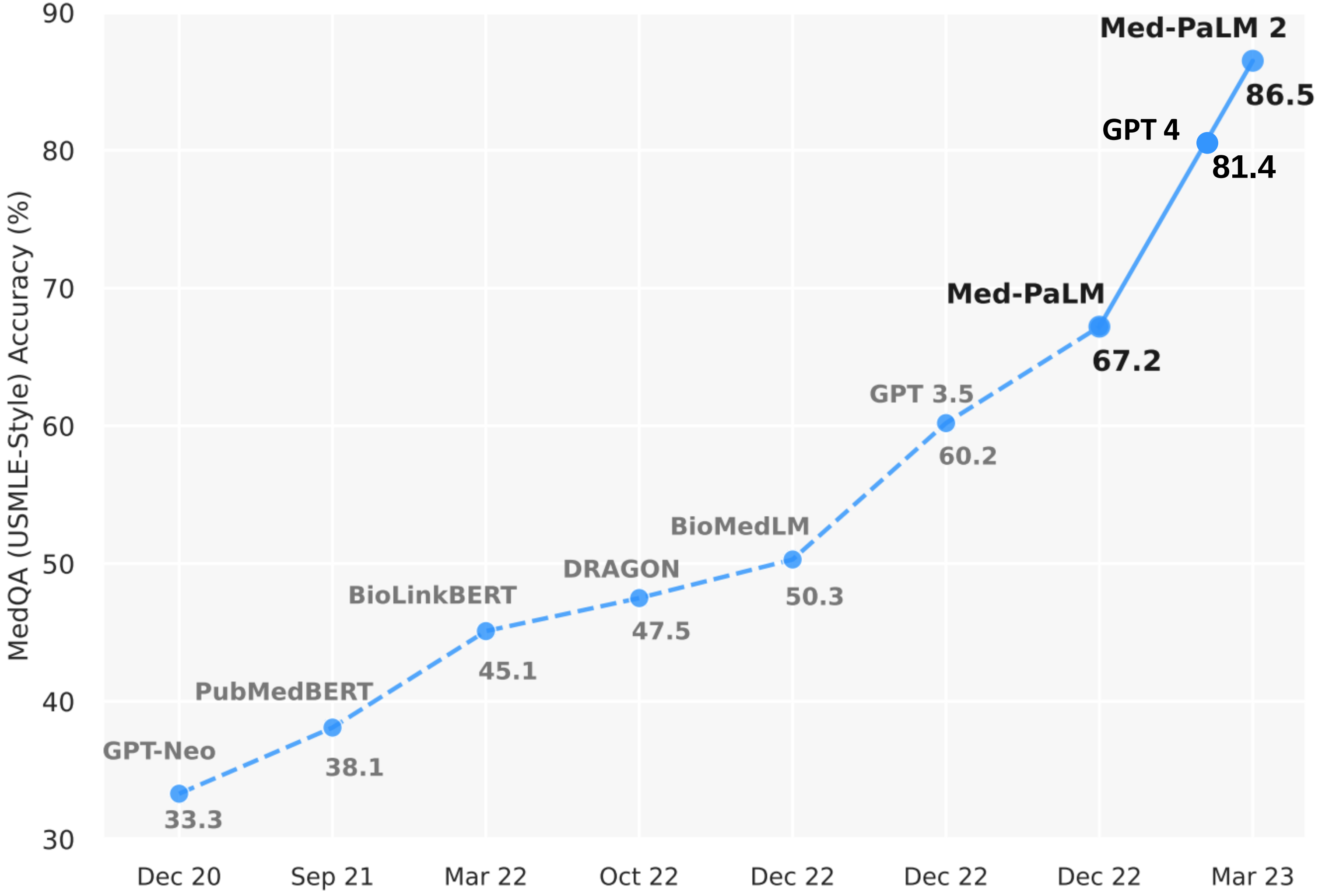
Seen on the floor, patient reports stable pleuritic chest pain. Endorses the dyspnea to be worse lying flat then when sitting up, and not provoked when leaning forward. Has had a dry cough over the last few weeks, for which she was recently placed on Azithromycin with slight improvement in her cough. Otherwise, no sick symptoms or sick contacts - however, has had daily fever for many weeks that has been attributed to active lupus flare. Methotrexate recently increased with improvement in some of her symptoms (such as hair loss). Otherwise, denies syncope, palpitations, abd pain, urinary symptoms, changes in bowel movements, or new edema. **Patient endorses that current symptoms are similar to those from 6 years ago, though these symptoms have been persisting for far longer.**

Current medications:

- Methotrexate 20 mg q7 days
- Hydroxychloroquine 200 mg daily
- Folic acid
- Apixaban

What does the data show?

LLMs score highly on medical exams



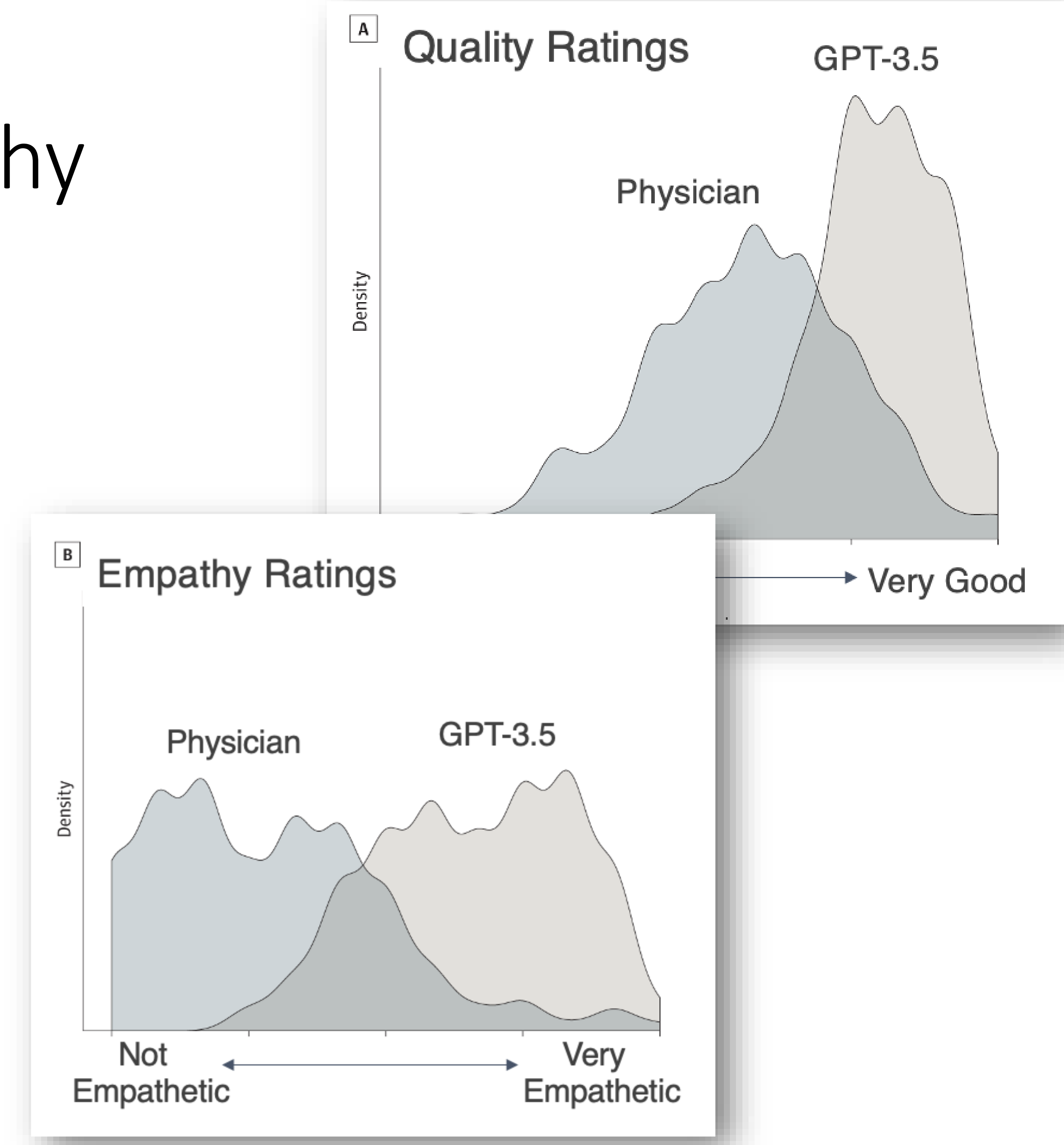
Singhal, Karan, et al. *arXiv preprint arXiv:2305.09617* (2023).

USMLE Sample Exam	**GPT-4** (% Correct)	**GPT-3.5** (% Correct)
Step 1	85.71	52.10
Step 2	83.33	58.33
Step 3	90.71	64.96
Overall Average*	**86.70**	**58.78**

Nori, Harsha, et al. *arXiv preprint arXiv:2303.13375* (2023).

LLMs can show empathy

- Study compared GPT-3.5 and physicians' responses to 200 patient inquiries on r/AskDocs
- 78.6% of evaluators preferred the GPT-3.5 response
- GPT-3.5 had higher ratings of quality and empathy
- Physicians responses were shorter (51 words for physicians, 211 words for GPT-3.5)

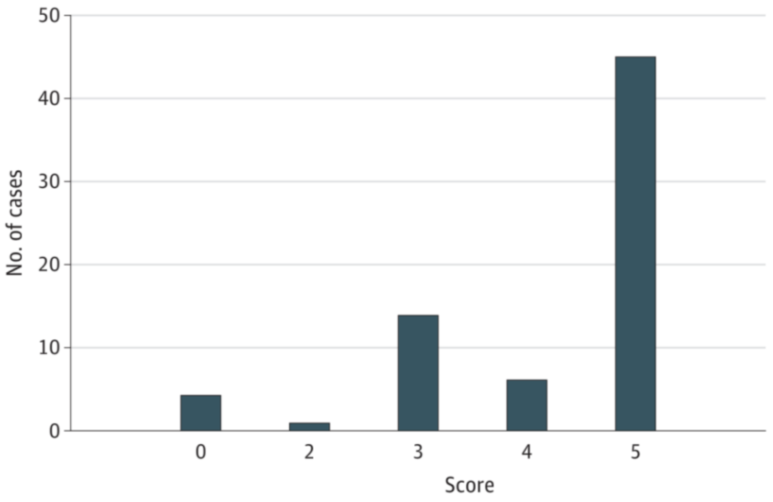


Ayers, John W., et al. *JAMA internal medicine* (2023).

LLMs can make diagnoses

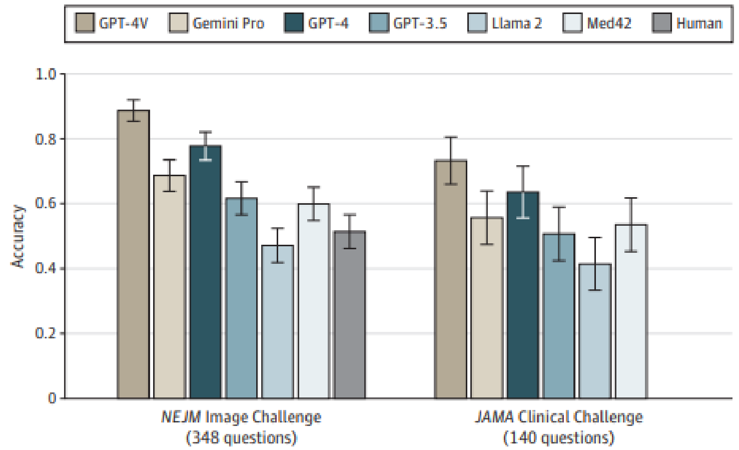
- GPT-4 can solve NEJM CPCs:
  - Top diagnosis: 27/70 (39%)
  - Diagnosis in differential: 45/70 (64%)
- Continued improvements over time:
  - GPT-4o currently at 77%, Llama 3.1 405b at 83%, Gemini at 63%.
- Similar performance gap seen with multimodal (text + clinical image) reasoning

Figure. Performance of Generative Pre-trained Transformer 4 (GPT-4)



Score = 5; Actual diagnosis was suggested in the differential
Score = 0; No suggestions were close to the diagnosis

Figure 1. Performance of Large Language Models on *New England Journal of Medicine* (*NEJM*) and *JAMA* Vignette Questions

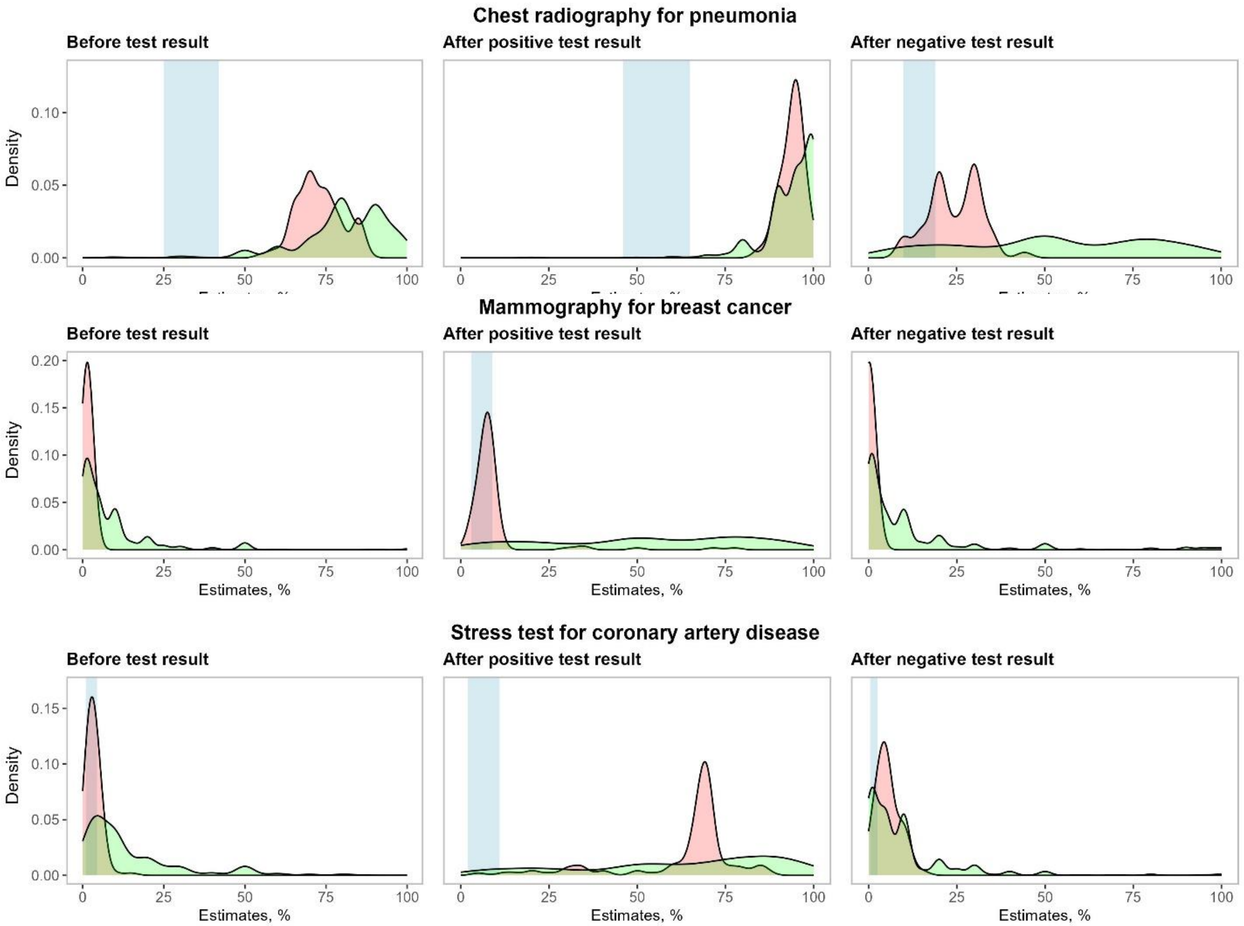


---

[1] Hirosawa, Takanobu, et al. *The American Journal of Medicine* (2023); [2] Kanjee, Zahir et al.

LLMs have emergent probabilistic reasoning

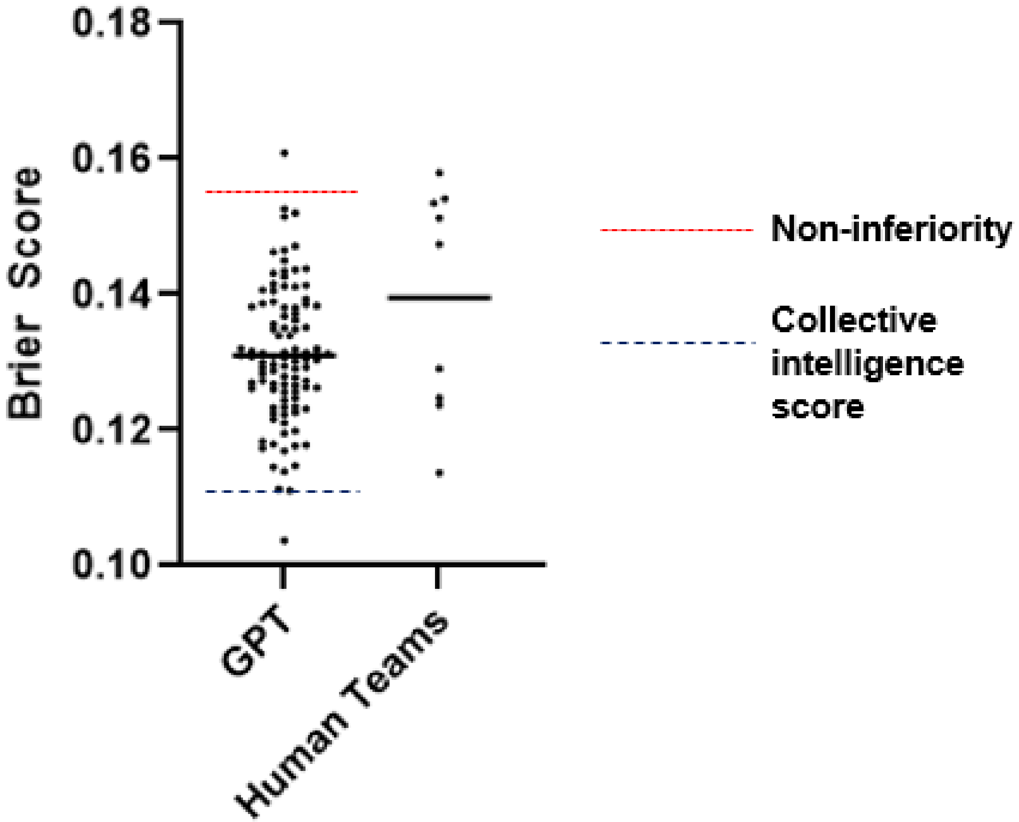
- Comparison GPT-4’s pre-test and post-test probability after a negative a positive test for “reference standard” conditions
- Compared 100 API calls versus 553 humans
- GPT-4 with much less MAE in all cases of pre-test probability and post-test after a negative; equivalent after positive



Rodman A, Buckley TA, Manrai AK, Morgan DJ. Artificial intelligence vs. clinician performance in estimating probabilities of diagnoses before and after testing. JAMA Open [Internet] Available from: http://dx.doi.org/10.1001/jamanetworkopen.2023.47075

LLMs can forecast superior to humans

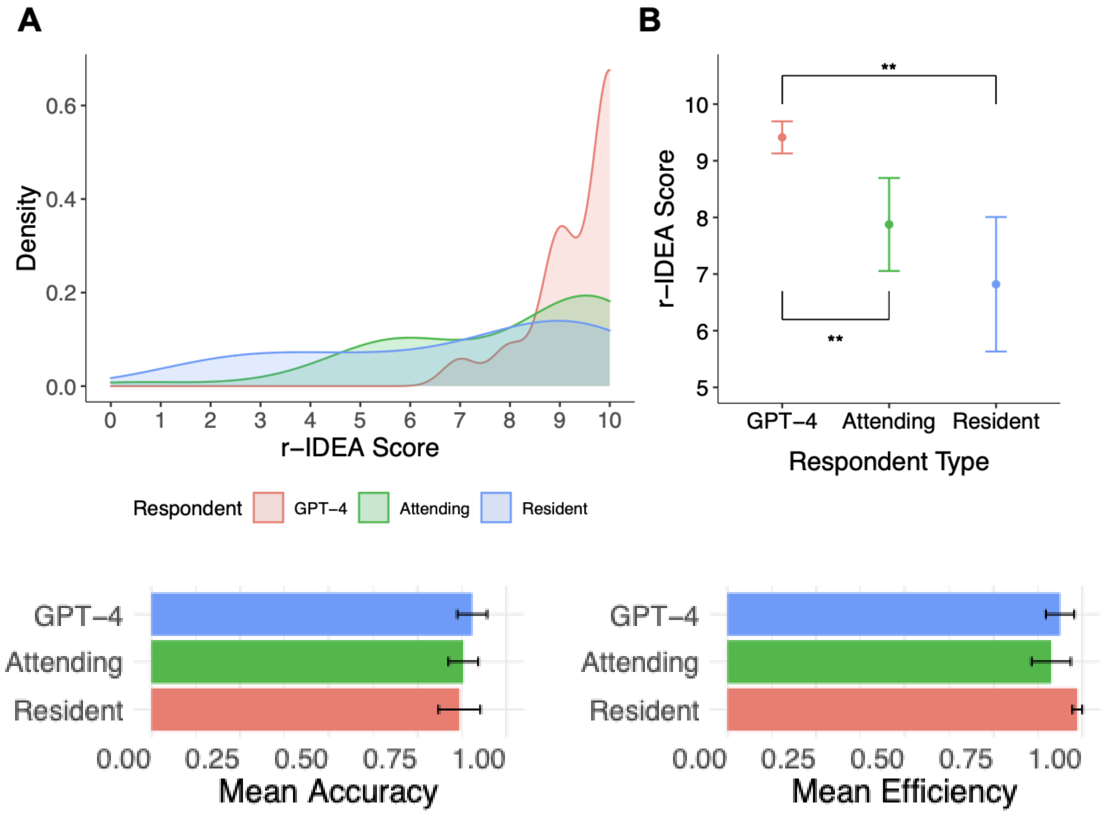
- Better at forecasting diagnoses than human teams with lower Brier scores
- Not superior to human collective intelligence – but what about human-LLM collective intelligence?



Stehouwer N, Rodman A. **Data not yet peer reviewed**

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

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



Cabral S, Rodman A et. All, 2024. JAMA Internal Medicine. In Press

Are LLMs alone better at making diagnoses than LLMs and people together?

- Recreation of the NEJM CPC study using a fine-tuned Palm2, this time with multiple human comparison groups.
- LLM alone outperformed clinician+LLM, outperformed clinician+search, outperformed unassisted clinician

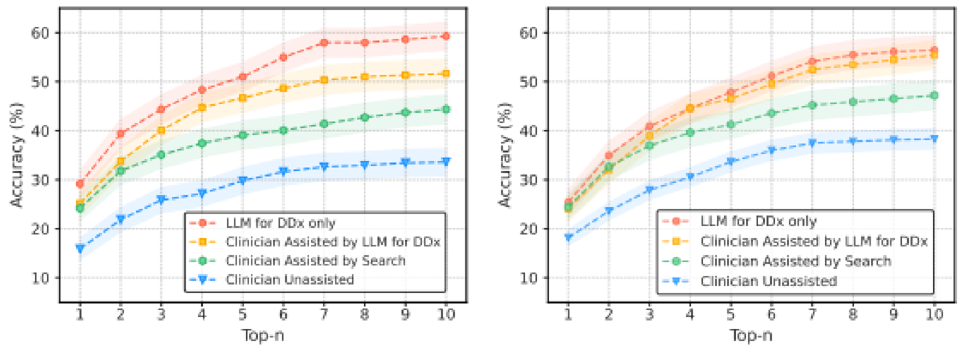
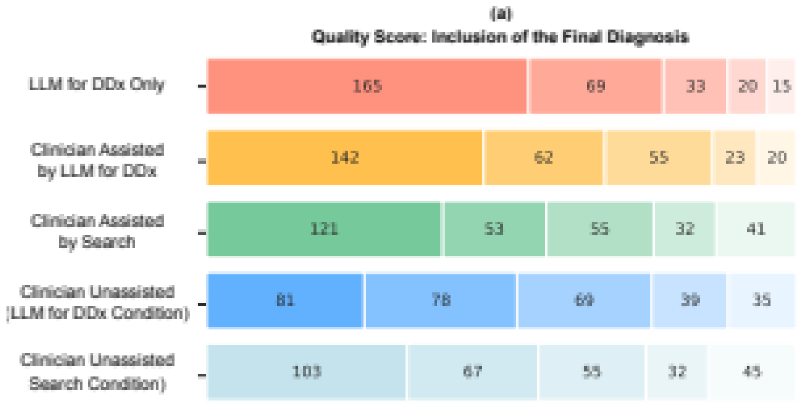


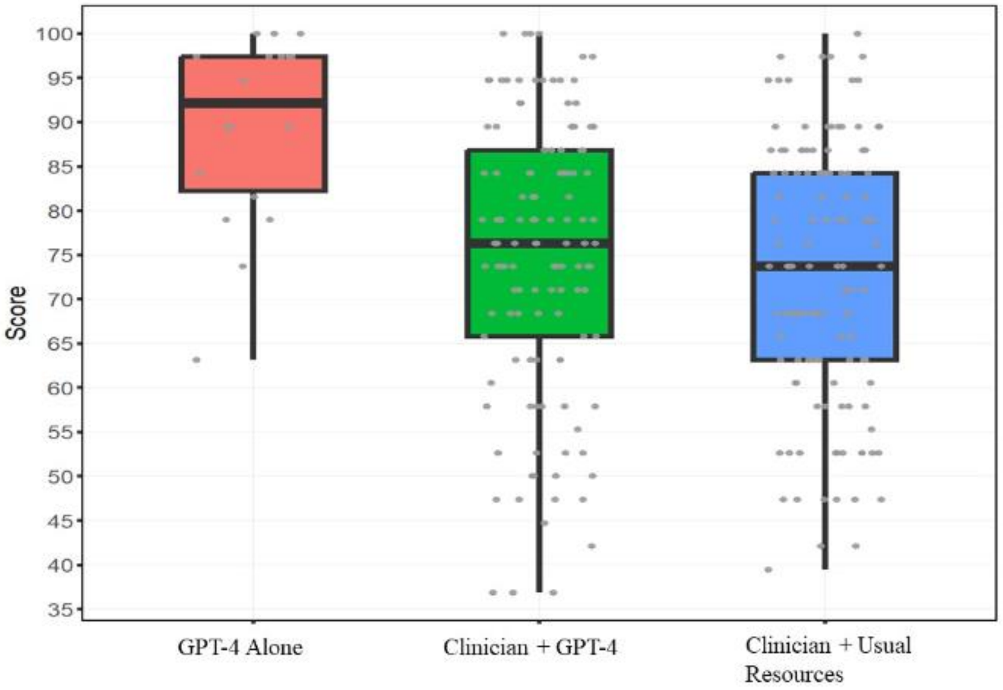
Figure 5 | Top-n Accuracy. (left) The percentage of DDx lists with the final diagnosis through human evaluation. (right) The percentage of DDx lists with the final diagnosis through automated evaluation.



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

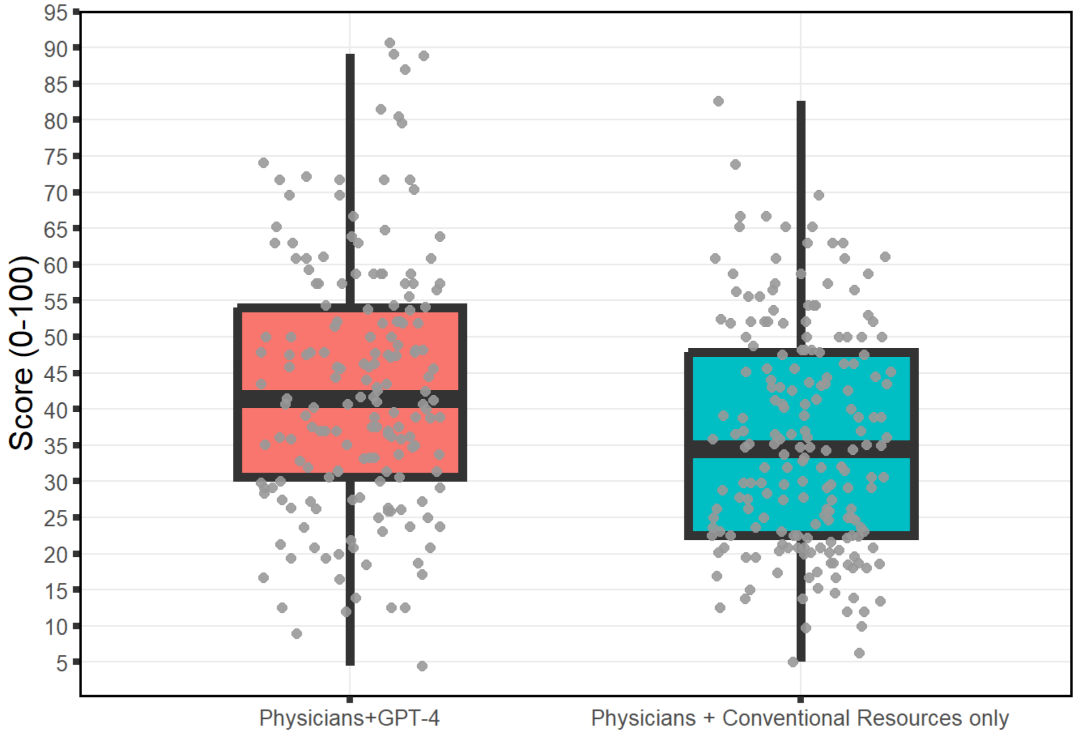
Are LLMs alone better than humans + LLMs at reflective reasoning?

- Single-blind RCT involving 50 US generalist clinicians solved difficult cases, randomized to either usual care (any digital resources) or usual care + LLM
- Outcome was structured reflection – gold standard in improving diagnostic reasoning.
- No difference in humans vs humans + LLM (though clinically meaningful but non-statistically significant increase in final diagnosis and efficiency) – but massive difference with LLM alone
- Humans + LLM had huge increase in time per case – saved over 2 minutes per case.



Can LLMs make management decisions?

- Randomized trial of 92 physicians solving 400 cases of complex management decisions (no right answers) using usual resources or usual resources + LLM
- LLM use had 8% increase in overall performance – all from case specific and management questions.



Rodman, Goh, Hom, Chen. *Under Review*

Can LLMs collect data?

- Double-blind trial using standardized patients of AMIE (Articulate Medical Intelligence Explorer)
- Using standardized rubrics (PACES), performed better than humans in 28 of 32 axes, which significantly improved diagnostic accuracy
- Trained by a unique "self-play" mechanism (synthetic data)

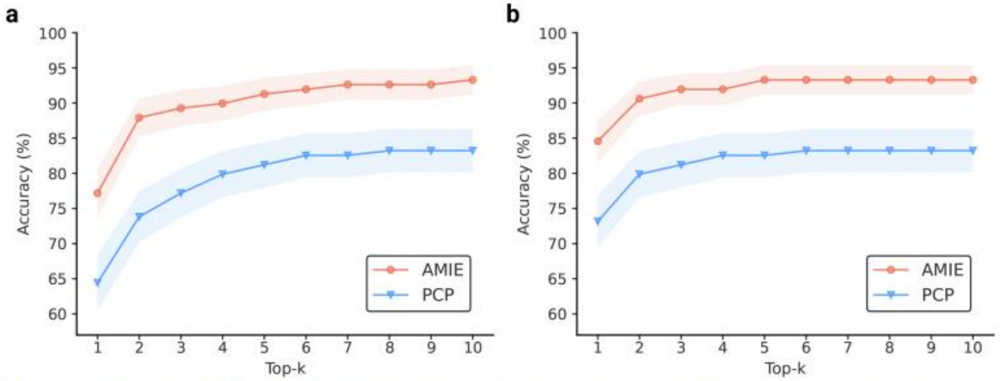


Figure 3 | **Specialist-rated top-k diagnostic accuracy.** AMIE and PCPs top-k DDx accuracy are compared across 149 scenarios with respect to the ground truth diagnosis (**a**) and all diagnoses in the accepted differential (**b**). Bootstrapping (n=10,000) confirms all top-k differences between AMIE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

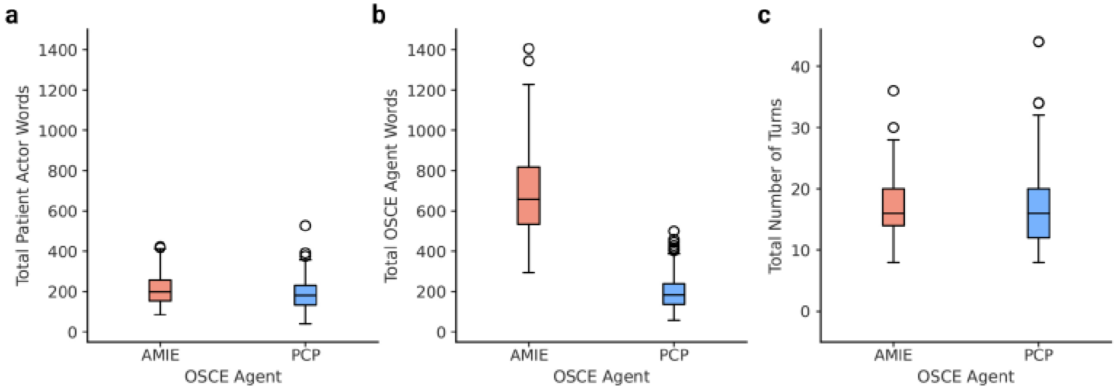


Figure A.11 | **Distribution of words and turns in OSCE consultations.** (**a**) Total patient actor words elicited by AMIE vs. PCPs. (**b**) Total words sent to patient actor from AMIE vs. PCPs. (**c**) Total number of turns in AMIE vs. PCP consultations.

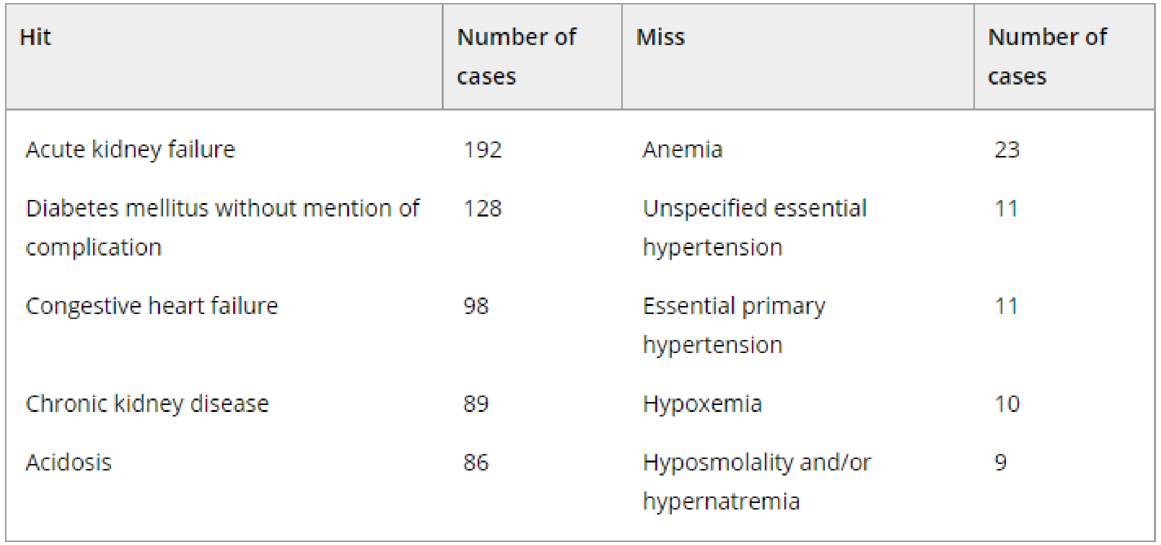
arXiv:2401.05654 [cs.AI]

What about EHR data?

- Random sample of structured and unstructured data (though no progress notes) from 1000 patients at BIDMC (MIMIC-IV)
- Reference standard of physicians + medical coders; determined the “hit rate” (that is, the proportion of correct diagnoses) from GPT-4 and PaLM2.

- Average hit rate of 94.1%, corresponding to 1116 unique diagnoses

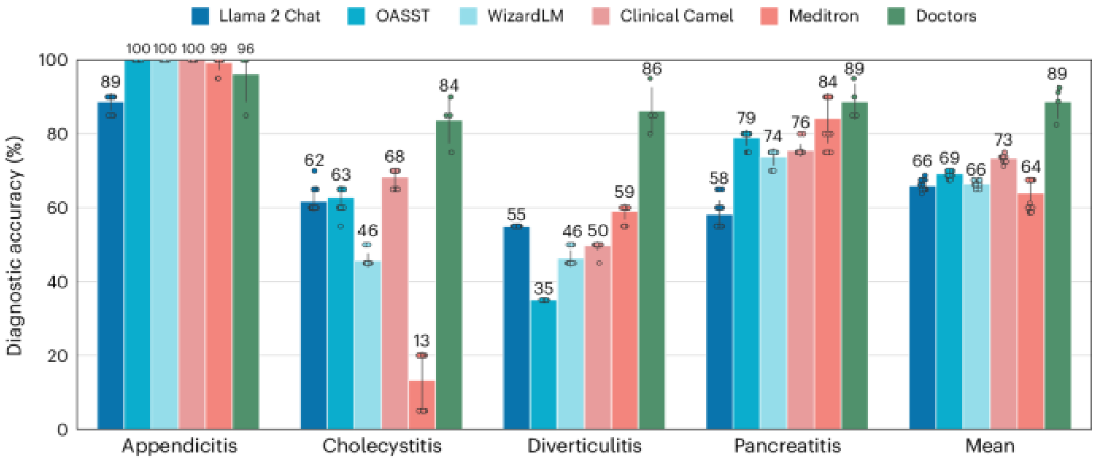
Table 1. Top 5 hits and misses.



Hit	Number of cases	Miss	Number of cases
Acute kidney failure	192	Anemia	23
Diabetes mellitus without mention of complication	128	Unspecified essential hypertension	11
Congestive heart failure	98	Essential primary hypertension	11
Chronic kidney disease	89	Hypoxemia	10
Acidosis	86	Hyposmolality and/or hypernatremia	9

Can LLMs use EHR data to make autonomous decisions?

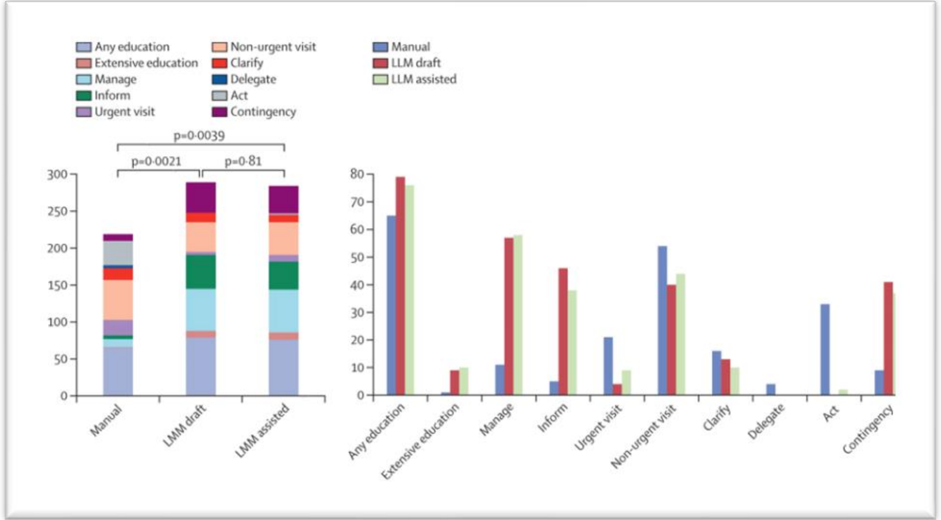
- Extracted diagnostic information from MIMIC IV to compare several LLMs against human clinicians in four abdominal pathologies
- LLMs significantly underperformed humans
- **No frontier models were included**



Early LLM in healthcare implementations have been mixed

- Randomized QI study of 122 physicians with AI drafted replies.
- Read time was **21.8% higher** in LLM group, reply time unchanged, and length 17.9% higher.

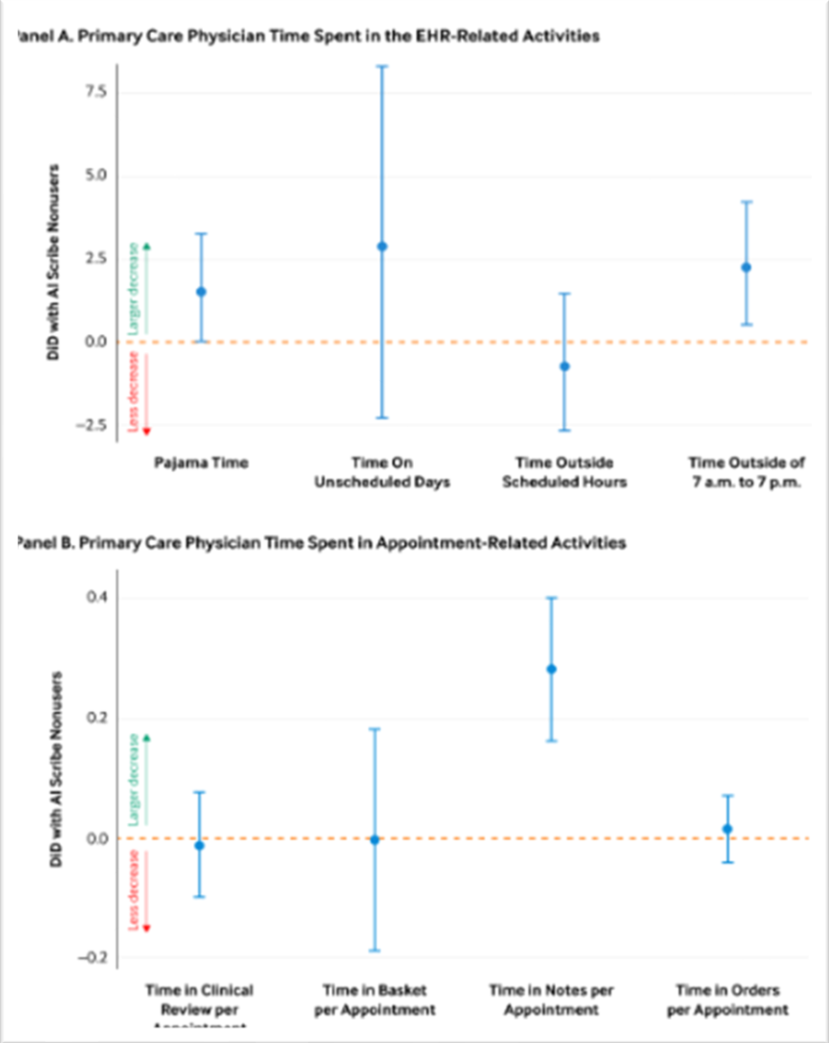
- “Turing test” study of patient concerns in Rad Onc clinic – AI + human **“best of both worlds”**



Tai-Seale M, Baxter SL, Vaida F, Walker A, Sitapati AM, Osborne C, Diaz J, Desai N, Webb S, Polston G, Helsten T, Gross E, Thackaberry J, Mandvi A, Lillie D, Li S, Gin G, Achar S, Hofflich H, Sharp C, Millen M, Longhurst CA. AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. JAMA Netw Open. 2024 Apr 1;7(4):e246565. doi: 10.1001/jamanetworkopen.2024.6565. PMID: 38619840; PMCID: PMC11019394. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, Chipidza FE, Leeman J, Aerts HJWL, Miller T, Savova GK, Gallifant J, Celi LA, Mak RH, Lustberg M, Afshar M, Bitterman DS. The effect of using a large language model to respond to patient messages. Lancet Digit Health. 2024 Apr 24:S2589-7500(24)00060-8. doi: 10.1016/S2589-7500(24)00060-8. Epub ahead of print. PMID: 38664108.

Early LLM in healthcare implementations have been mixed

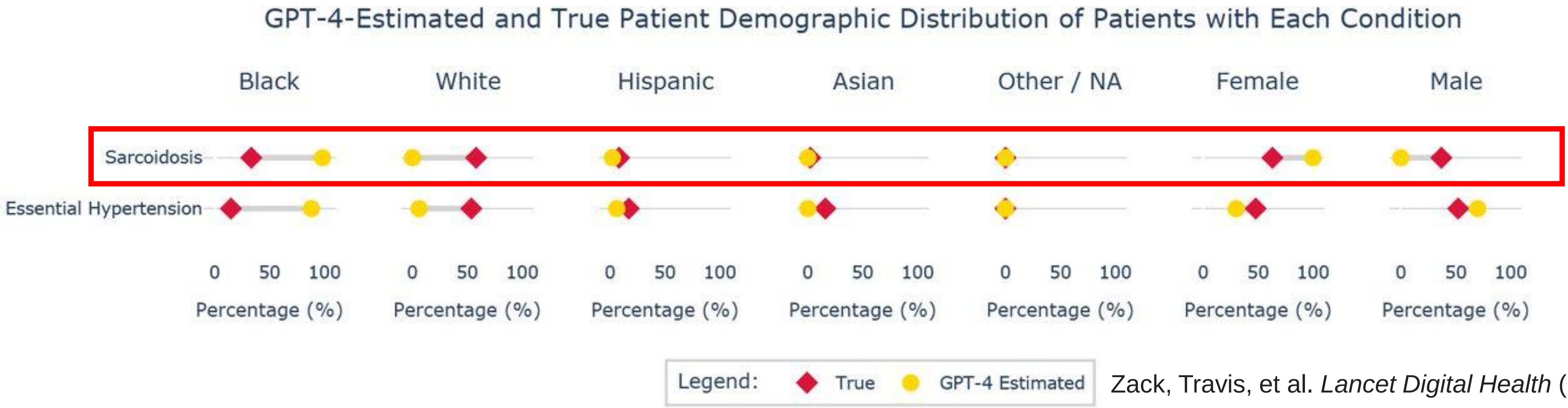
- QI project of Nabla at Kaiser Northern California – decreased time spent documenting, with no other changes in EHR utilization with a dose-response curve
- Manual audit of notes showed high quality of Nabla-assisted notes



Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation, NEJM Catalyst

LLMs contain the bias of their pretraining and finetuning

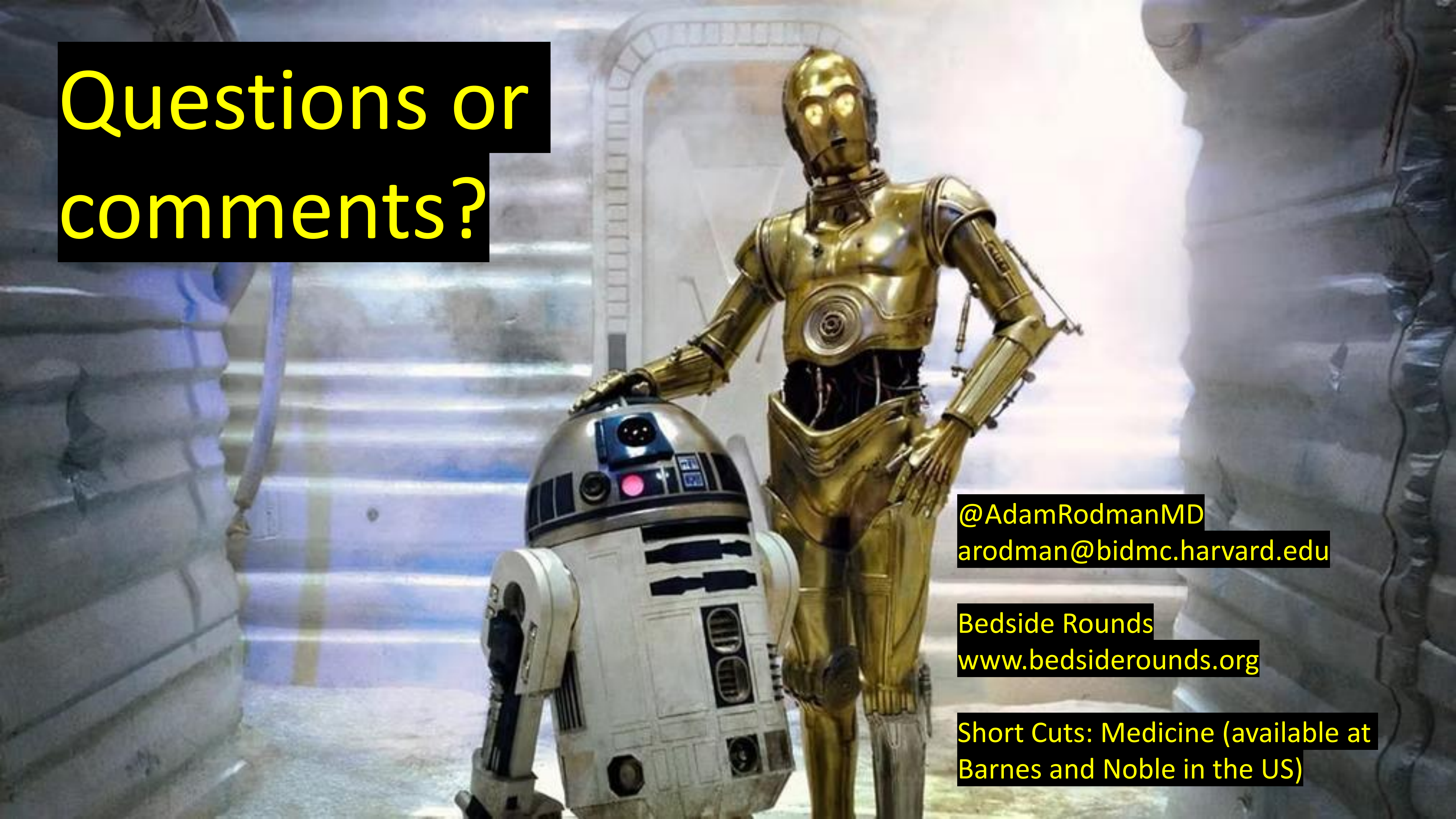
- Asked GPT-4 to create clinical vignettes
  - Over-represented demographic stereotypes of diseases
- Asked GPT-4 to give management plans for cases while substituting gender and race/ethnicity
  - Less likely to recommend advanced imaging for Blacks compared to whites



Zack, Travis, et al. *Lancet Digital Health* (2023).

What are the challenges to an AI second opinion service?

- Human-computer interaction is tricky:
  - **Which clinicians** should be targeted and **when**?
  - **Which cases** benefit the most from a second opinion?
  - **How** does an AI second opinion affect quality and resource utilization?
- Automation bias affects all CDS systems
- How does this affect diagnostic deskilling in the future?
  - LLMs will **never** be able to replace human diagnosticians (though other AI technologies might)



Questions or comments?

@AdamRodmanMD
arodman@bidmc.harvard.edu

Bedside Rounds
www.bedsiderounds.org

Short Cuts: Medicine (available at Barnes and Noble in the US)