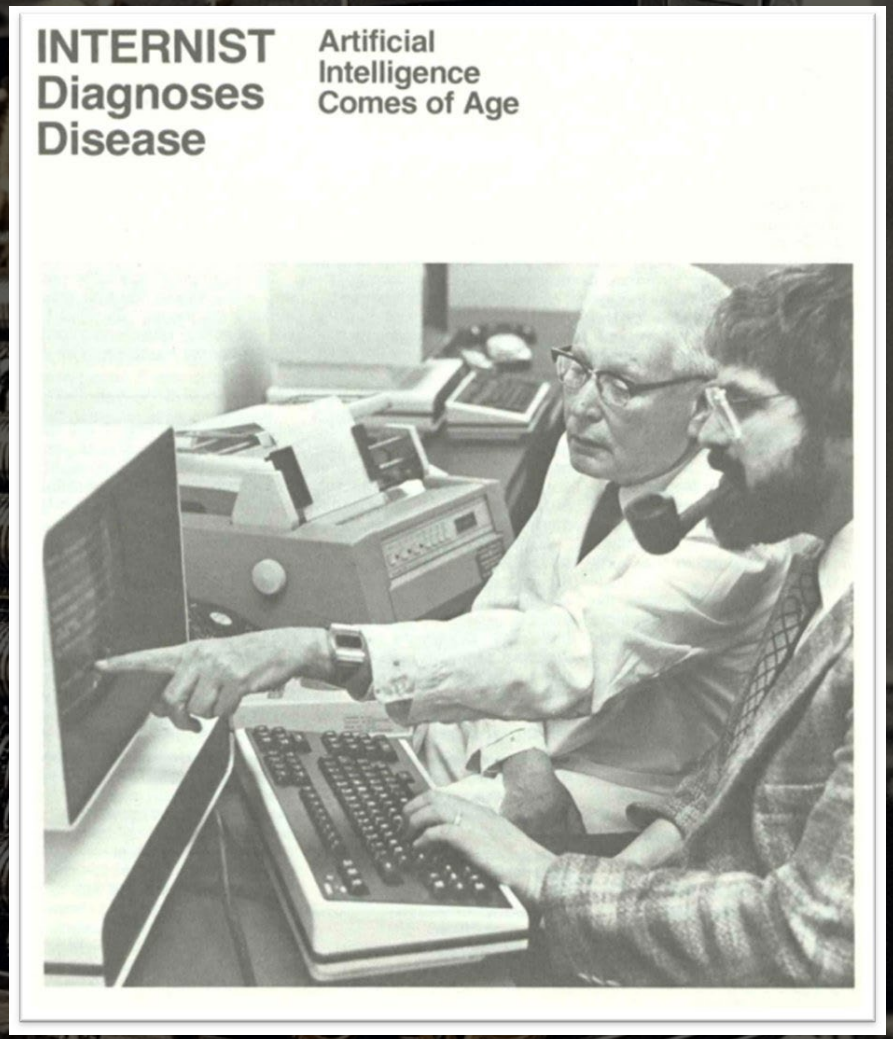
Harnessing Artificial Intelligence to Enhance Diagnostic Reasoning

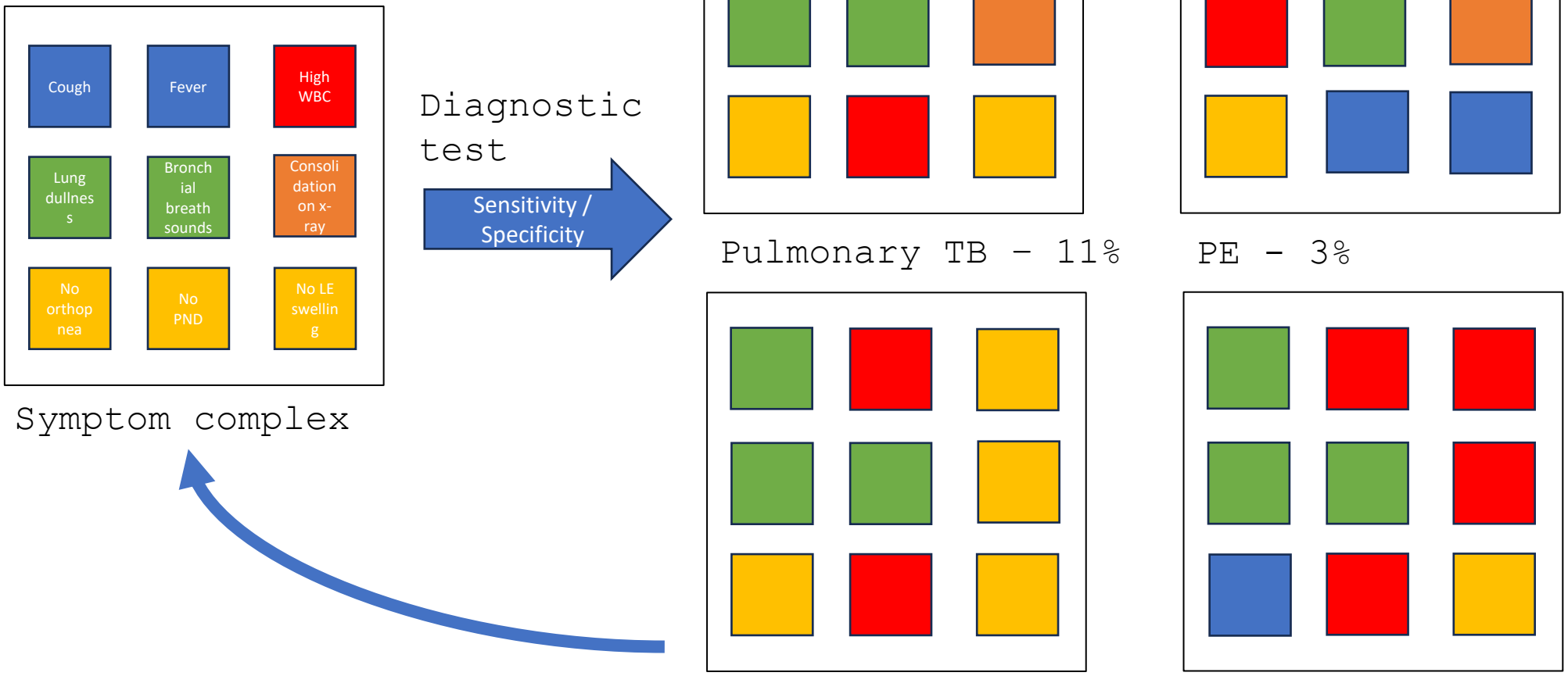
Adam Rodman, MD, MPH, FACP
Director of AI Programs, Beth Israel Deaconess Medical Center
Assistant Professor, Harvard Medical School

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

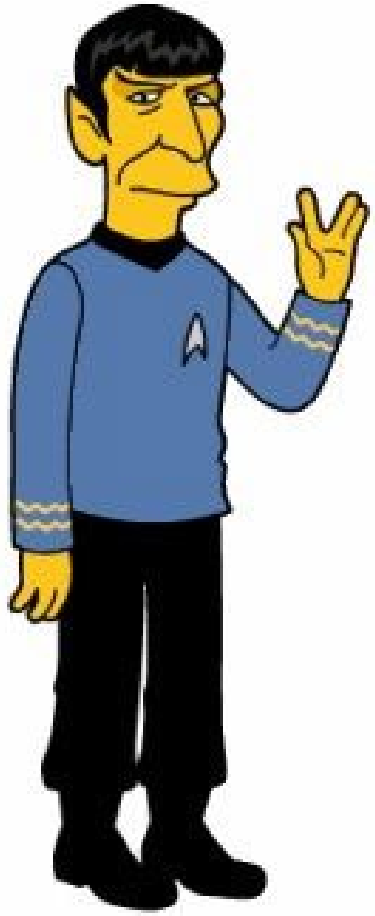


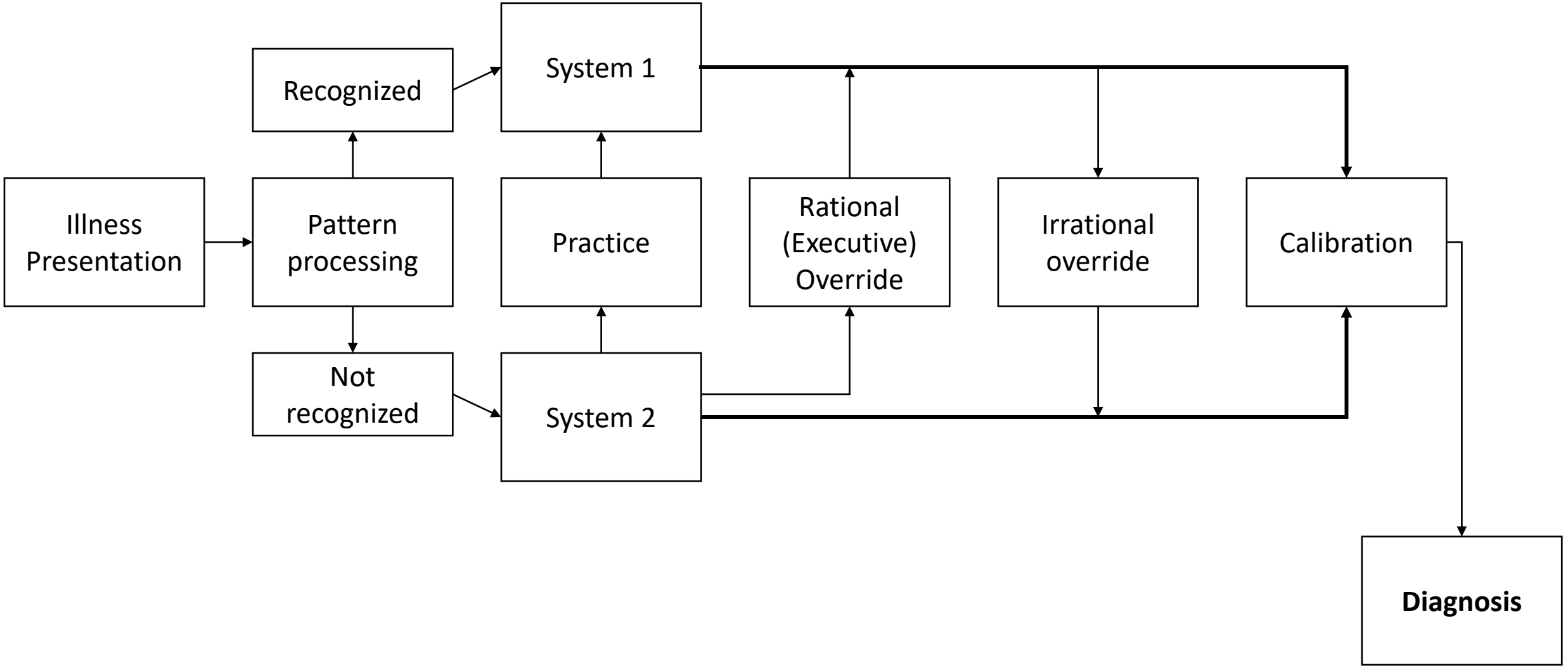
- University of Pittsburgh Alumni Magazine, 1978


Cough
Fever
High WBC
Lung dullness
Bronchial breath sounds
Consolidation on x-ray
No orthopnea
No PND
No LE swelling
Symptom complex
Diagnostic test
Sensitivity / Specificity
Pneumonia – 61%
CHF – 6%
Pulmonary TB – 11%
PE – 3%
Disease complexes

System 1

System 2




Illness Presentation
Pattern processing
Recognized
Not recognized
System 1
System 2
Practice
Rational (Executive) Override
Irrational override
Calibration
Diagnosis

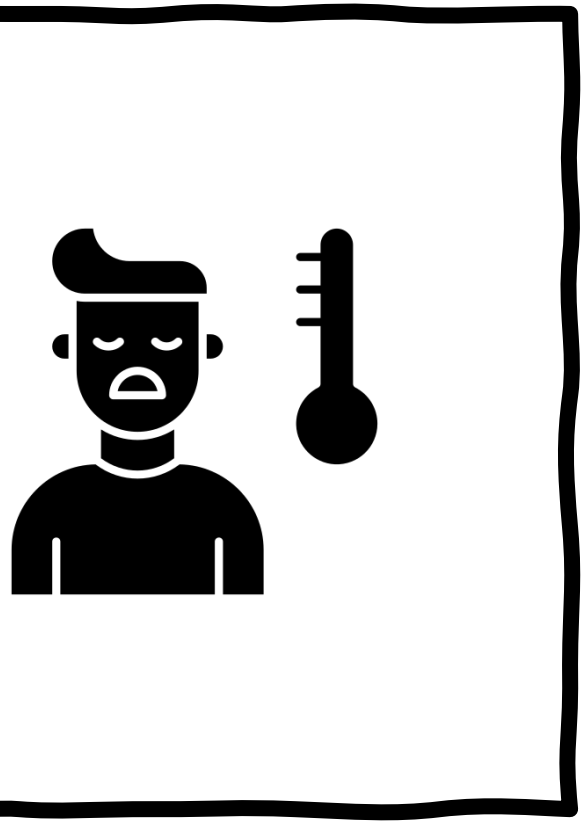
Daily fever

Monocytosis

Night sweats

GGOs on x-ray

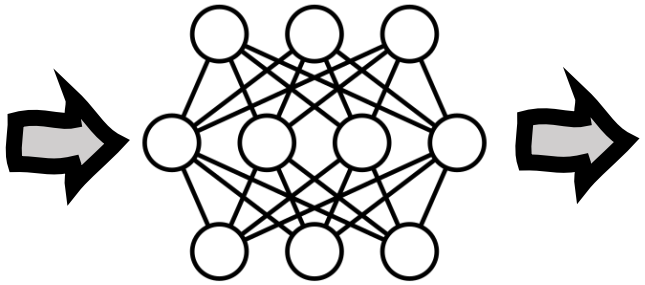
History of BCG for bladder CA



No improvement with antibiotics

Elevated liver enzymes

History of spinal surgery

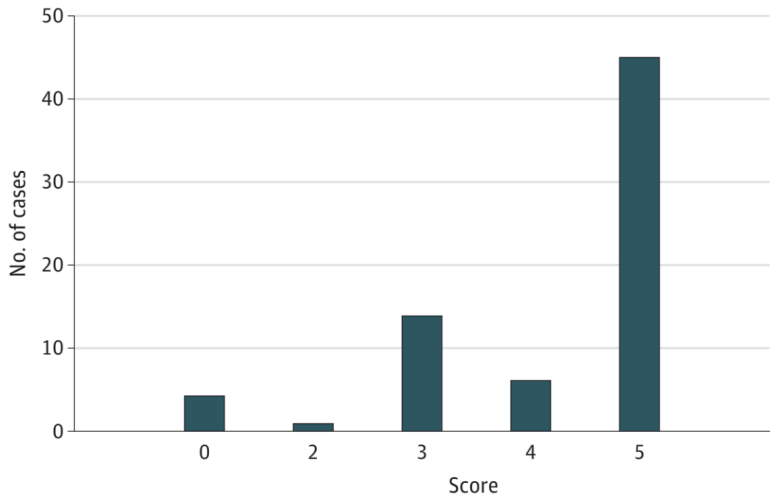


1. **Disseminated Tuberculosis** - 20%: Given his travel history and the clinical presentation, this remains a possibility, though it would be uncommon in the absence of a positive imaging or microbiological evidence.
2. **Endocarditis** - 15%: Even though he has a history of ischemic heart disease and 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.

LLMs can make diagnoses

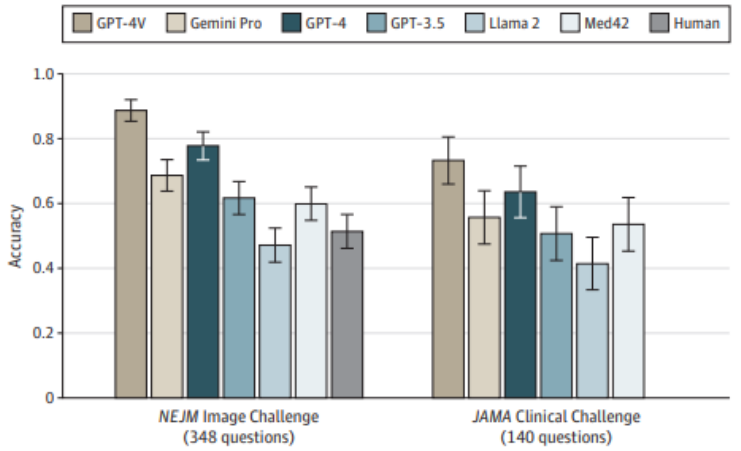
- GPT-4 can solve NEJM CPCs:
  - Top diagnosis: 27/70 (39%)
  - Diagnosis in differential: 45/70 (64%)
- Models have been improving over time (now close to 85% accurate with GPT-4o and Gemini Ultra)
- Similar performance gains seen with multimodal (text + clinical image) models

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

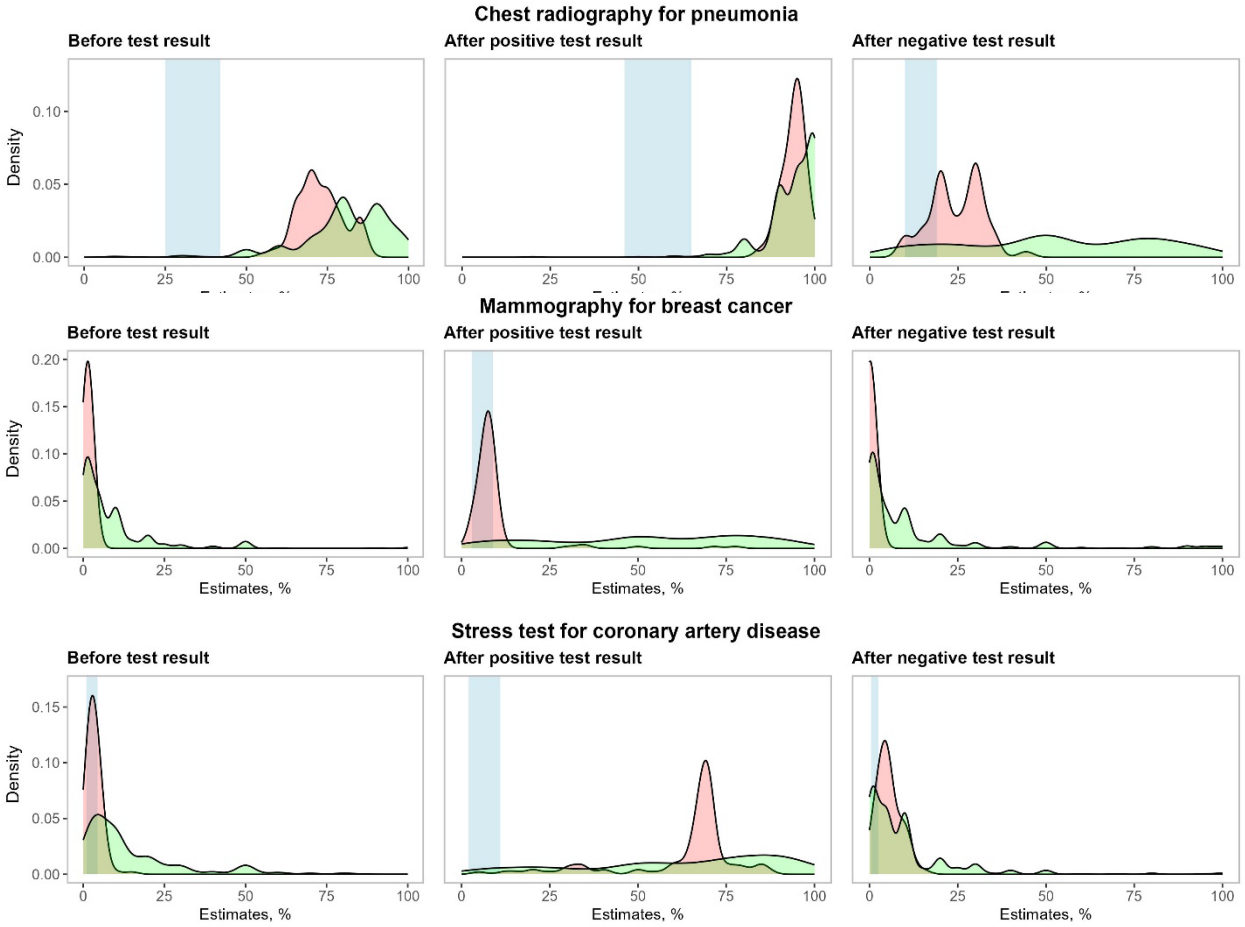
Figure 1. Performance of Large Language Models on *New England Journal of Medicine* (*NEJM*) and *JAMA* Vignette Questions



Kanjee, Zahir et al. *JAMA* (2023); Han, Adam et ak, JAMA (2024).

LLMs have emergent probabilistic reasoning

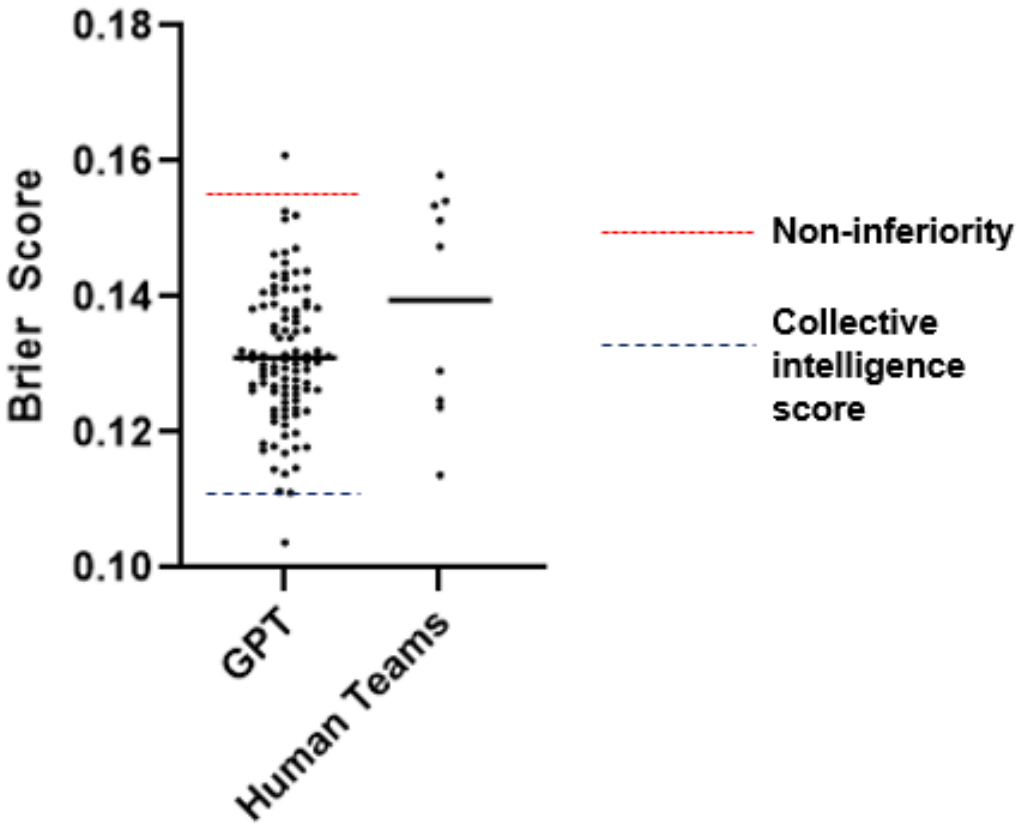
- Comparison GPT-4's pre-test and post-test probability after a negative or positive test for "reference standard" conditions
- Compared 100 API calls versus 553 humans
- GPT-4 with much less MAE in all cases of pre-test probability and post-test after a negative; equivalent after positive



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

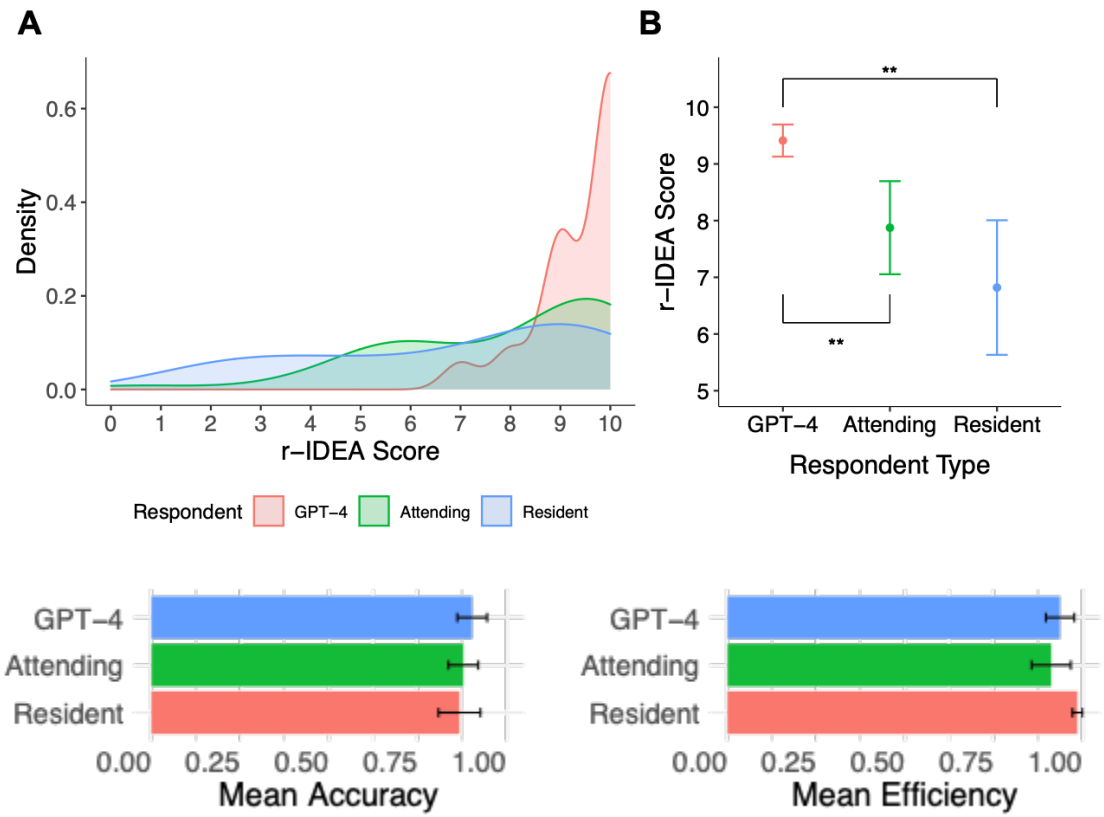
- Better at forecasting diagnoses than human teams with lower Brier scores
- Not superior to human collective intelligence – but what about human-LLM collective intelligence?



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

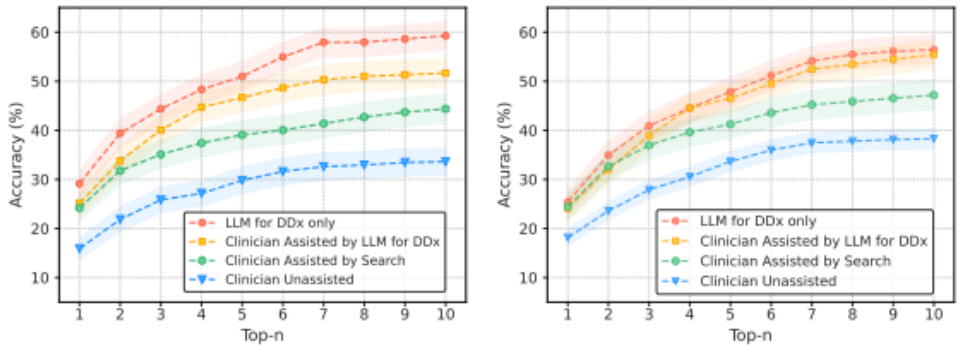
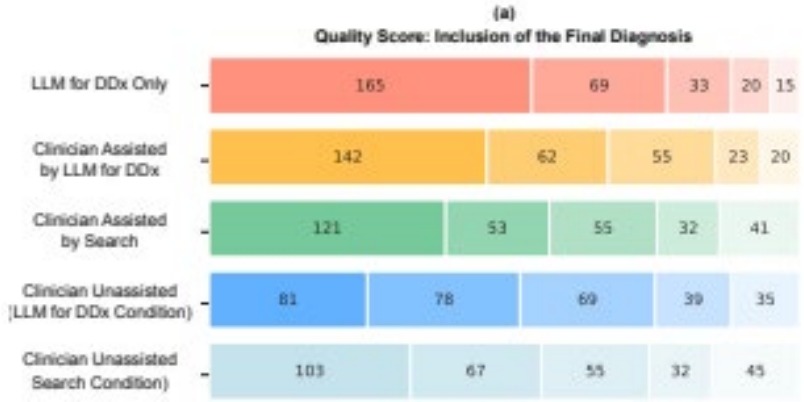


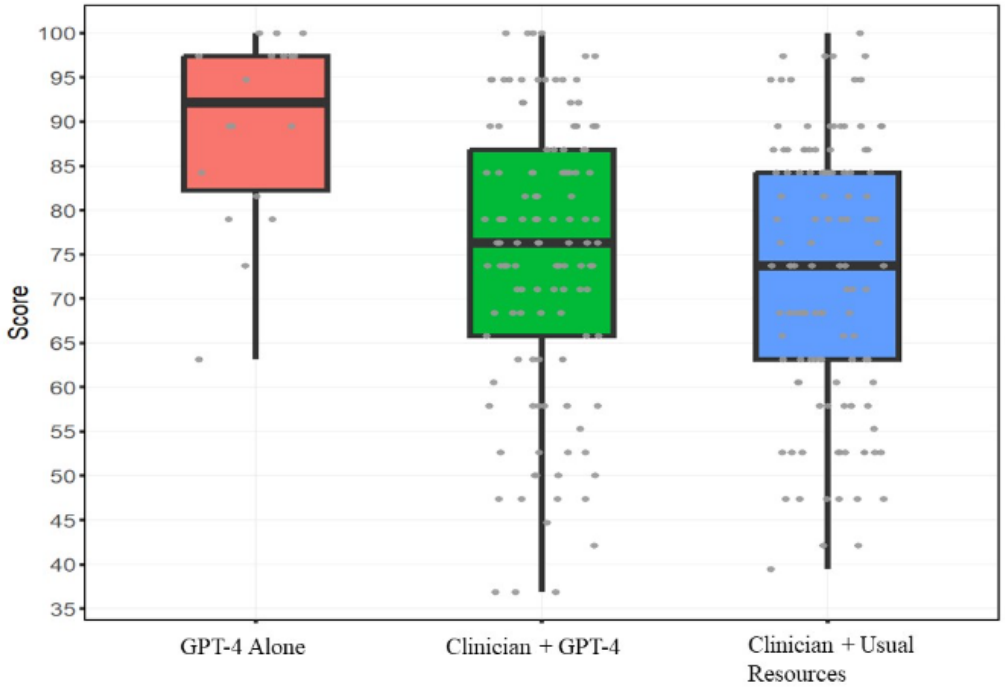
Figure 5 | Top-n Accuracy. (left) The percentage of DDx lists with the final diagnosis through human evaluation. (right) The percentage of DDx lists with the final diagnosis through automated evaluation.



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?

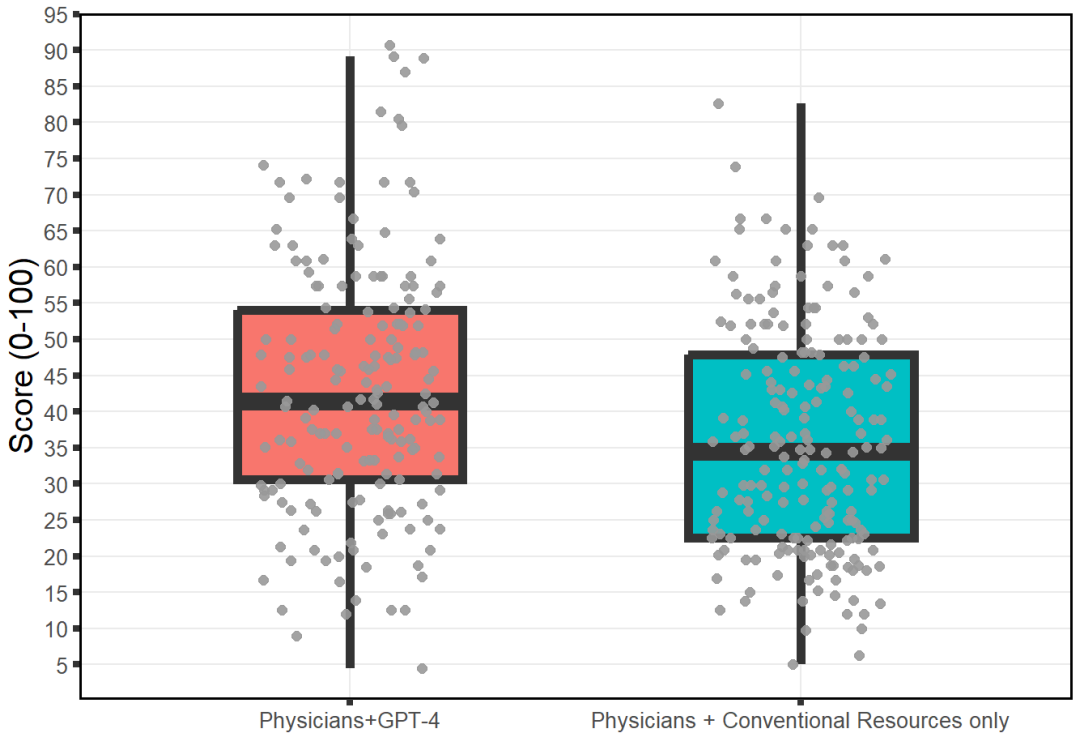
- Randomized involving 50 US generalist clinicians solved difficult cases, randomized to either usual care (any digital resources) or usual care + LLM
- Outcome was structured reflection – gold standard in improving diagnostic reasoning.
- No difference in humans vs humans + LLM (though clinically meaningful 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.



Goh, Rodman, Chen. *JNO* (In Press)

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.
- Considerably (2 minutes) slower in AI group



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)

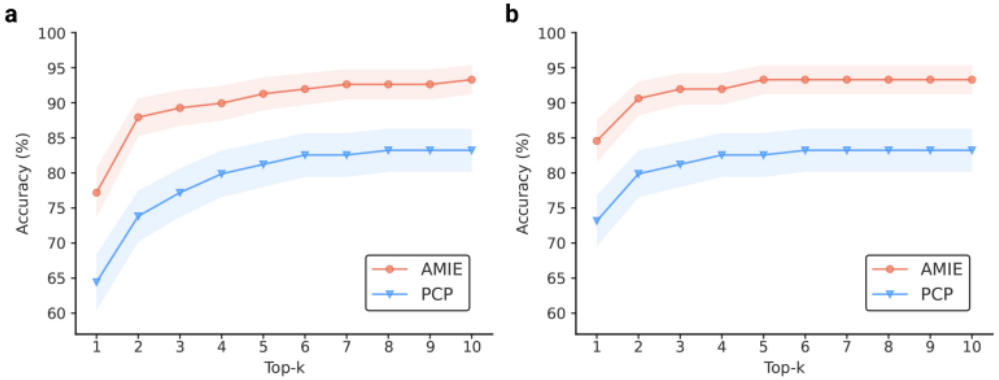


Figure 3 | Specialist-rated top-k diagnostic accuracy. AMIE and PCPs top-k DDx accuracy are compared across 149 scenarios with respect to the ground truth diagnosis (**a**) and all diagnoses in the accepted differential (**b**). Bootstrapping (n=10,000) confirms all top-k differences between AMIE and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

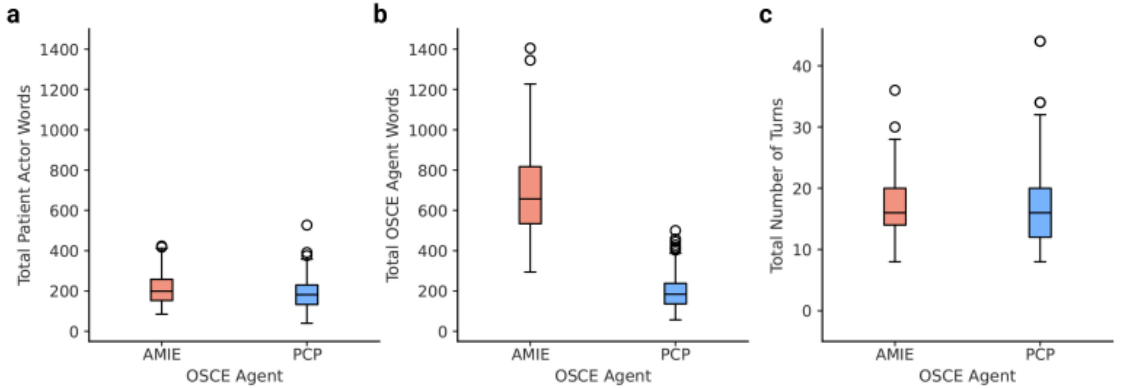


Figure A.11 | Distribution of words and turns in OSCE consultations. (**a**) Total patient actor words elicited by AMIE vs. PCPs. (**b**) Total words sent to patient actor from AMIE vs. PCPs. (**c**) Total number of turns in AMIE vs. PCP consultations.

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.

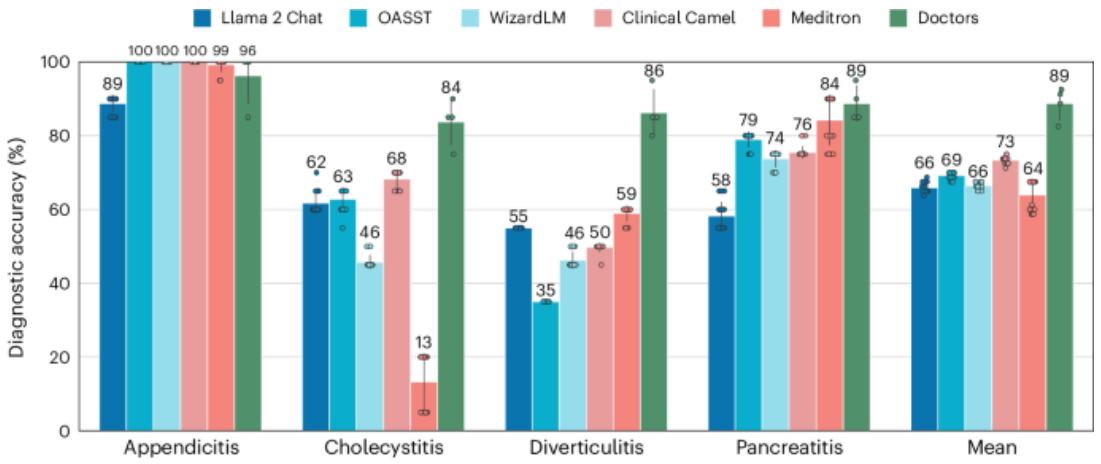
- Average hit rate of 94.1%, corresponding to 1116 unique diagnoses

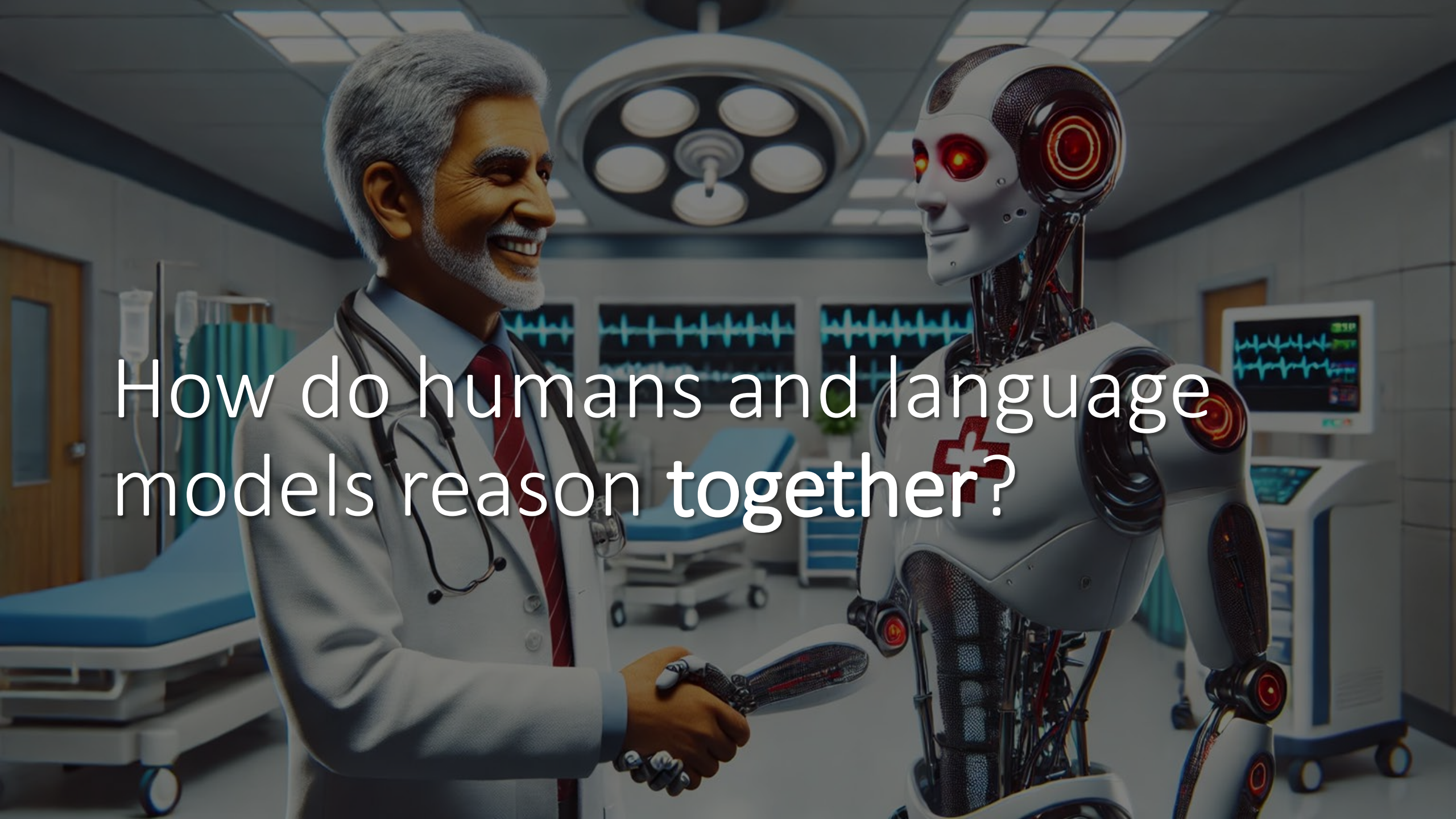
Table 1. Top 5 hits and misses.

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




How do humans and language models reason **together**?