Generative AI and Medical Diagnosis

Lowell General Hospital Grand Rounds
December 17, 2025

Arjun (Raj) Manrai, PhD
Arjun_Manrai@hms.harvard.edu
@arjunmanrai

Agenda

1. Lab and research overview
2. Intro to the *NEJM* Clinicopathological Conferences (CPCs)
3. Intro to “Dr. CaBot”
4. CaBot solving an NEJM CPC

Manrai Lab

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

Computation to Improve **Medical Decision Making**

Recurring Themes

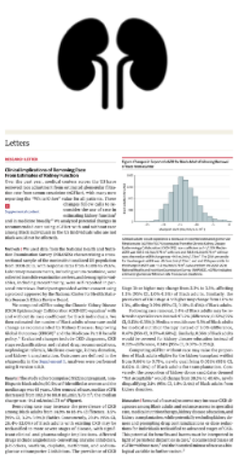
Traditional clinical equations

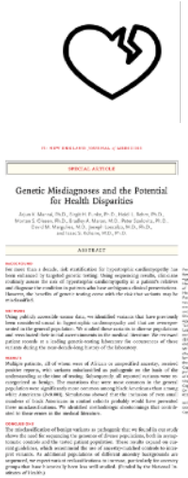
- Quantifying clinical implications of changing risk equations (e.g. eGFR, PFTs, ASCVD)
- Genetic (mis)diagnosis
- Who is “normal”? Lab reference ranges
- Race, environment, SDOH

Artificial intelligence for medicine

- LLM & multimodal diagnostic reasoning, automated assessment
- Fine-tuning, alignment
- Human-machine collaboration
- Role of synthetic data from diffusion models (e.g. derm)

Research


JAMA 2021



NEJM 2016

NEJM 2024

JAMA 2018


CVPR 2021

NeurIPS SynData4ML 2022

Harrisons 2021

NEJM 2023

Clinical Impact

NATIONAL KIDNEY FOUNDATION®

ASN American Society of Nephrology

ClinGen Clinical Genome Resource

The National Academies of SCIENCES ENGINEERING MEDICINE

NEJM AI

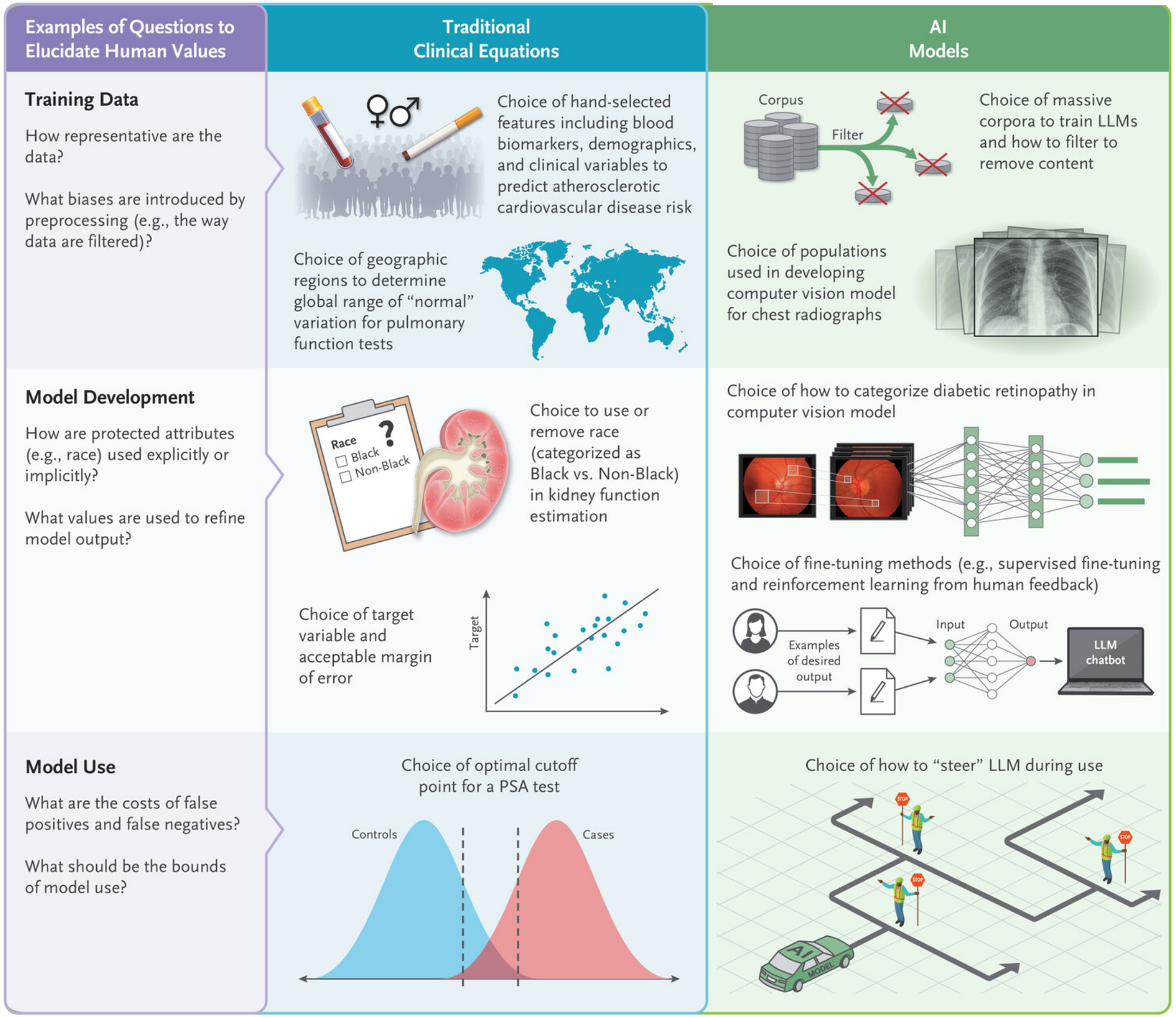
AI GRAND ROUNDS

NEJM AI

The NEW ENGLAND JOURNAL of MEDICINE

Medical Artificial Intelligence and Human Values

Kun-Hsing Yu, M.D., Ph.D., Elizabeth Healey, S.B., Tze-Yun Leong, Ph.D., Isaac S. Kohane, M.D., Ph.D., and Arjun K. Manrai, Ph.D.



1) Many lessons for AI from traditional clinical equations

2) Behavior (and values) shaped by: (a) training data, (b) model development, (c) model use

ORIGINAL ARTICLE | SEP 26, 2024 | FREE

Four-Channel ECG as a Single Source for Early Diagnosis of Cardiac Hypertrophy and Dilation — A Deep Learning Approach

H. Zhu and Others

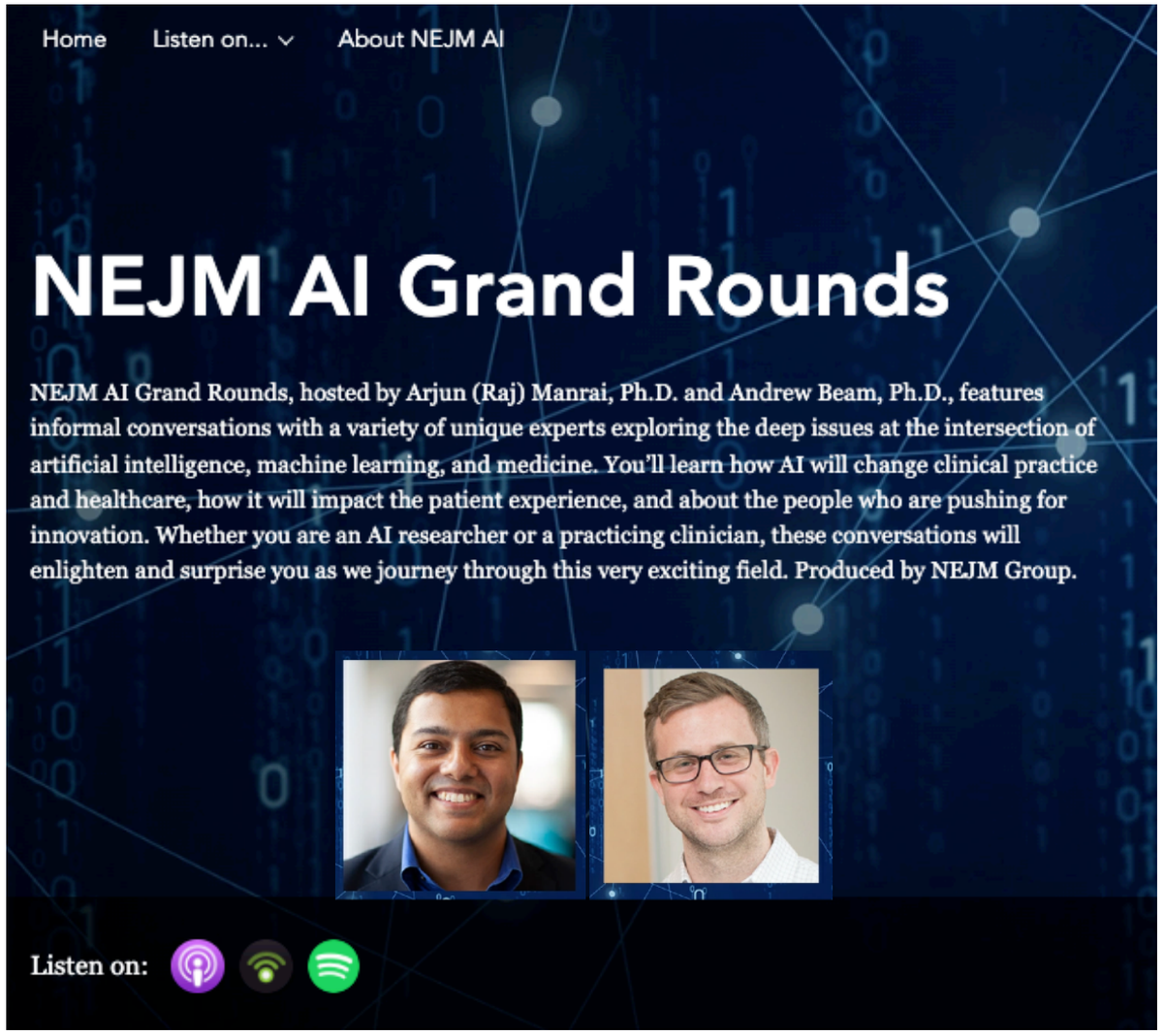
A large-scale cardiac hypertrophy, dilation, and enlargement database suggests that, in over half of patients with cardiac hypertrophy, dilation, and enlargement, cases can potentially be detected during routine electrocardiogram monitoring.

PERSPECTIVE | OCT 10, 2024

The EU AI Act: Implications for U.S. Health Care

S. Porsdam Mann, I.G. Cohen, and T. Minssen

A perspective on how the major requirements of the European Union's AI Act may impact the work of physicians and medical innovators in the United States over the next few years.



HARVARD–MIT
HEALTH SCIENCES AND TECHNOLOGY

All MEMP students must complete:

- Biomedical Sciences Core Requirements
- Biomedical Sciences Restricted Electives
- Clinical Coursework Requirements

Biomedical Sciences Core Requirements

- Human Pathology (HST030/031 or HST034/035)
- Molecular Genetics in Modern Medicine (HST160/161)
- Cardiovascular Pathophysiology (HST090/091)

Biomedical Sciences Restricted Electives

Two full courses required*

- Human Functional Anatomy (HST010/011)
- Musculoskeletal Pathophysiology (HST020/021)*
- Respiratory Pathophysiology (HST100/101)**
- Renal Pathophysiology (HST110/111)**
- Neuroscience (HST130/131)
- Molecular Diagnostics and Bioinformatics (HST162/163) *
- Principles of Biomedical Imaging (HST164/HST165)*
- Cellular and Molecular Immunology (HST175/176)

* May combine two half-courses to count as one full course
** Must choose at least one of HST100, HST110

Clinical Coursework Requirements

All MEMP students must complete two clinical courses:

- HST201: Introduction to Clinical Medicine and Medical Engineering I
- HST202: Introduction to Clinical Medicine and Medical Engineering II

HST201W and HST202W are offered at the West Roxbury Veteran's Administration Hospital. I
offered at Mt. Auburn Hospital. Students must complete both HST201 and HST202 at a single

CASE RECORDS *of the* MASSACHUSETTS GENERAL HOSPITAL

Case 23-2022: A 49-Year-Old Man with Hypoglycemia

Amy W. Baughman, M.D., Nancy J. Wei, M.D., Peter F. Hahn, M.D., Ph.D.,
Brenna W. Casey, M.D., and M. Lisa Zhang, M.D.

PRESENTATION OF CASE

Dr. Mollie Sands (Medicine): A 49-year-old man was admitted to this hospital because of hypoglycemia.

The patient had been well until 3 hours before this admission, when altered mental status developed while he was at work. During a meeting, his colleagues noticed that he was not paying attention or participating in the conversation, which was atypical of the patient. Instead, he was fidgeting and mumbling as though talking to another person. When his mobile telephone rang, he allowed it to continue ringing.

Two hours before this admission, the patient was found by colleagues on the floor of a conference room. His eyes were open, but he was obtunded and making nonpurposeful movements. Emergency medical services were called. On evaluation, a fingerstick glucose measurement was 39 mg per deciliter (2.2 mmol per liter; reference range, 70 to 109 mg per deciliter [3.9 to 6.1 mmol per liter]). Intravenous dextrose was administered, and the patient was brought to the emergency department of this hospital for further evaluation.

On arrival at the emergency department, the patient was obtunded and did not react to sternal rub. Additional intravenous dextrose was administered, and a dextrose infusion was started. The patient's mental status improved, and he was able to provide additional history. That morning, he had felt that he was in a "dream-like" state; there were no other symptoms. The altered mental status had resolved, but he could not recall most events of the morning, including the meeting that had taken place 3 hours before admission.

Review of symptoms was notable for unintentional weight gain of one pant size during the past 3 months, as well as intermittent morning dizziness during the past 3 years, which would resolve after the patient ate candy or breakfast. He reported no recent illness, depression, anxiety, change in activity or exercise, or change in intake of food or drink. On a typical day, he had breakfast at 5 a.m., a second breakfast at 10 a.m., lunch at 1 p.m., dinner at 6 p.m., and a snack at 8 p.m. The patient had last eaten approximately 7 hours before presentation to the emergency department.

Two years earlier, the patient had had an episode of drooping and decreased sensation of the left side of the face and difficulty enunciating words, which occurred while he was at work. He presented to the emergency department of this hospital. The blood glucose level was 64 mg per deciliter (3.6 mmol per liter), and testing for Lyme disease was negative. He received a diagnosis of Bell's palsy and was treated with prednisone and acyclovir; the symptoms resolved. Two months later, the patient established care in the primary care clinic of this hospital. Routine laboratory testing revealed a blood glucose level of 46 mg per deciliter (2.6 mmol per liter).

The patient took vitamin B_{12} and vitamin D supplements. He took no other medications and had no known drug allergies. He drank alcohol rarely and did not smoke tobacco or use illicit drugs. Sixteen years earlier, he had immigrated to the United States from Southeast Asia and now lived alone in an urban area of New England. He worked as a registered nurse. His father and brother had type 2 diabetes and used insulin.

On examination, the temperature was 35.7°C, the heart rate 82 beats per minute, the blood pressure 146/76 mm Hg, and the respiratory rate 14 breaths per minute. The body-mass index (the weight in kilograms divided by the square of the height in meters) was 27.6. The patient appeared well. He had no hepatosplenomegaly or skin hyperpigmentation, and the remainder of the physical examination was normal.

On arrival at the emergency department, the blood glucose level was 36 mg per deciliter (2.0 mmol per liter); after the administration of dextrose, a fingerstick glucose measurement was 158 mg per deciliter (8.8 mmol per liter). The complete blood count and blood electrolyte levels were normal, as were the results of liver-function and kidney-function tests. The glycated hemoglobin level was 5.4% (reference range, 4.3 to 6.4). Serum and urine toxicologic panels were negative. Laboratory test results are shown in Table 1. The patient was admitted to this hospital.

A diagnostic test was performed.

Table 1. Laboratory Data.*

Variable	Reference Range, Adults†	On Admission
Blood		
Alanine aminotransferase (U/liter)	5–34	33
Aspartate aminotransferase (U/liter)	6–55	28
Alkaline phosphatase (U/liter)	45–115	105
Lipase (U/liter)	13–60	67
Albumin (g/dl)	3.3–5.0	4.4
Sodium (mmol/liter)	135–145	138
Potassium (mmol/liter)	3.4–4.8	3.8
Chloride (mmol/liter)	98–107	103
Carbon dioxide (mmol/liter)	21–30	24
Urea nitrogen (mg/dl)	8–25	12
Creatinine (mg/dl)	0.60–1.50	1.03
Glucose (mg/dl)	70–109	36
Serum toxicologic panel		
Ethanol (mg/dl)	0.0	0.0
Salicylates (mg/dl)	0.0–20.0	<0.3
Acetaminophen (μg/ml)	0.0–25.0	<5.0
Tricyclic antidepressants	Negative	Negative
Urine		
Urine toxicologic panel		
Amphetamines	Negative	Negative
Barbiturates	Negative	Negative
Benzodiazepines	Negative	Negative
Cannabinoids	Negative	Negative
Cocaine metabolites	Negative	Negative
Opiates	Negative	Negative

* To convert the values for urea nitrogen to millimoles per liter, multiply by 0.357. To convert the values for creatinine to micromoles per liter, multiply by 88.4. To convert the values for glucose to millimoles per liter, multiply by 0.05551. To convert the values for ethanol to millimoles per liter, multiply by 0.2171. To convert the values for salicylates to millimoles per liter, multiply by 0.07240.

† Reference values are affected by many variables, including the patient population and the laboratory methods used. The ranges used at Massachusetts General Hospital are for adults who are not pregnant and do not have medical conditions that could affect the results. They may therefore not be appropriate for all patients.

Asking The Human Expert

Differential Diagnosis

Dr. Amy W. Baughman: This 49-year-old man presented with recurrent symptomatic hypoglycemia. He reported that he had gained weight recently, that he was eating five meals a day, and that he had intermittent morning dizziness that resolved with food. His medical history included Bell's palsy and hypoglycemia that had not been evaluated previously. His social history was notable for his job as a nurse in a medical facility and his family history of diabetes. Initial laboratory testing revealed a normal glycated hemoglobin level and a low glucose level (36 mg per deciliter).

In developing a differential diagnosis, the most specific feature of this patient's presentation is symptomatic hypoglycemia. He presented with Whipple's triad, a sign that was first described in the 1930s by Dr. Allen Whipple, who aimed to establish criteria for pancreatic surgery for insulinomas.[1] Whipple's triad includes symptoms of hypoglycemia, which often occur in a fasting state or after exercise; a low plasma glucose level; and resolution of the symptoms with the administration of glucose. Today, this sign is useful for identifying persons who should undergo a formal evaluation for hypoglycemia.

In constructing a differential diagnosis to determine the cause of hypoglycemia, there are three broad categories to consider: systemic illness, ingestion of drugs or toxins, and endogenous insulin production.

Systemic Illness

SYSTEMIC ILLNESS

Hypoglycemia is a common problem in hospitalized patients, both those who have diabetes mellitus and those who do not; it can be seen with several systemic disease processes.[2] Hypoglycemia can occur in patients with sepsis or a condition that causes a malnourished state, such as anorexia nervosa. It can also be seen in patients with liver failure, kidney failure, congestive heart failure, hormonal deficiency (e.g., adrenal insufficiency), or disorders of carbohydrate metabolism. Normally, the body has several regulatory mechanisms that prevent hypoglycemia, including hormonal and neural signaling pathways that ensure glucose availability from intake, hepatic and renal gluconeogenesis, and hepatic glycogenolysis. However, in the context of systemic diseases, some of the pathways may be dysfunctional or impaired to such an extent that the normal compensatory processes are overwhelmed and hypoglycemia occurs.

A much less common systemic illness that can cause hypoglycemia is cancer complicated by non–islet-cell tumor hypoglycemia. This paraneoplastic syndrome results in overproduction of insulin-like growth factor II, which stimulates the insulin receptor.[3] Patients with this complication usually present with large, clinically obvious tumors,[4] such as mesenchymal tumors, fibromas, adenocarcinomas, or hepatocellular carcinomas.[3]

In this case, these causes of hypoglycemia are unlikely because the patient was generally well and did not have evidence of a systemic illness, kidney or liver disease, or cancer. In addition, the fact that he had persistent hypoglycemia, despite having functional compensatory mechanisms such as hepatic gluconeogenesis, suggests an alternative mechanism, specifically medication or hyperinsulinism.

Ingestion of Drugs or Toxins

INGESTION OF DRUGS OR TOXINS

Hypoglycemia can be due to ingestion of drugs or toxins. Heavy alcohol use may result in hypoglycemia in persons who have poor nutritional status or impaired liver function due to hepatic glycogen depletion.[5,6] This patient reported that he drank alcohol rarely, and he had an ethanol level of 0 mg per deciliter on admission; also, he did not have evidence of nutritional deficiencies or liver disease.

Aside from diabetes medications, several drugs have been associated with hypoglycemia, although in a systematic review of 164 drugs, the quality of evidence supporting the association was only low to moderate.[7] Medications that are commonly implicated as causes of hypoglycemia include fluoroquinolones, quinine, beta-blockers, and angiotensin-converting–enzyme inhibitors. Hypoglycemia has also been reported after ingestion of herbal products that are contaminated with sulfonylureas and glyburide.[8,9] This patient was taking vitamin B_{12} and vitamin D supplements, but he did not report using any other medications or herbal supplements, and the chronic nature of his hypoglycemic symptoms suggests that medications are not the cause of his hypoglycemia.

Medications that are used to treat diabetes — insulin and insulin secretagogues, such as sulfonylureas and meglitinides — can cause hypoglycemia. Accidental hypoglycemia occurs when a patient unintentionally takes a diabetes medication, or someone unintentionally administers a diabetes medication to the patient, because of a medication mix-up or contamination. In contrast, factitious hypoglycemia occurs when a patient intentionally or surreptitiously takes a diabetes medication to induce low blood sugar. A person with an episode of factitious hypoglycemia may have diabetes, may be in close contact with someone who has diabetes, or may be a medical professional. The person may also have an underlying psychiatric disorder, such as major depression, or have a history of suicidality[10]; in one case report, 2 out of 10 patients with an episode of factitious hypoglycemia went on to commit suicide.[11] Finally, malicious hypoglycemia occurs when someone intentionally administers a diabetes medication to the patient to induce low blood sugar, which can be seen in cases of Munchausen's syndrome by proxy.[12] Accidental, factitious, and malicious hypoglycemia often occur at random times, with no relation to a fasting state or meals.

Of note, this patient probably had access to diabetes medications because he had family members with diabetes and was a nurse, but he did not have a history of mental illness or depression. In addition, his episodes of hypoglycemia were not random; they were induced by a fasting state.

Endogenous Insulin Production

ENDOGENOUS HYPERINSULINISM

Causes of endogenous insulin production include noninsulinoma pancreatogenous hypoglycemia syndrome (NIPHS), insulin autoimmune hypoglycemia, and insulinomas. NIPHS is characterized by the presence of islet cells with abnormal morphologic features, including hypertrophy and nesidioblastosis, in the absence of an islet-cell tumor.[13] NIPHS is less common than insulinomas, and it is associated with hypoglycemia that occurs 2 to 4 hours postprandially, whereas hypoglycemia occurred in a fasting state in this patient.[13]

Insulin autoimmune hypoglycemia is caused by antibodies that target insulin or the insulin receptor. One form of insulin autoimmune hypoglycemia is Hirata's disease (insulin autoimmune syndrome), which occurs when autoantibodies against endogenous insulin develop. Patients with Hirata's disease can present with hypoglycemic episodes that range in severity, although episodes are usually mild and postprandial.[14] The syndrome is often triggered by exposure to medications such as thiol compounds[15] or by viruses,[16] and it can be associated with autoimmune diseases or hematologic disorders.[17] Hirata's disease was first described in Japan and is the third most common cause of hypoglycemia there.[15] However, the disease is rare globally, with only 380 cases reported worldwide from 1970 to 2009.[19]

An even more rare form of insulin autoimmune hypoglycemia is type B insulin resistance syndrome, which occurs when autoantibodies against the insulin receptor develop. Only 67 cases of type B insulin resistance syndrome have been reported worldwide as of 2014.[20] Patients with type B insulin resistance syndrome most often present with severe diabetes and extreme insulin resistance, but they can also present with hypoglycemia.[18] At low concentrations, anti–insulin receptor antibodies can cause hypoglycemia, with an effect similar to that of insulin; at higher concentrations, these antibodies can cause insulin resistance and hyperglycemia.[20,21] Type B insulin resistance syndrome is also strongly associated with autoimmune diseases. This patient did not have diabetes mellitus or evidence of other autoimmune or hematologic disorders, so both Hirata's disease and type B insulin resistance syndrome are unlikely.

Insulinomas are insulin-producing tumors that arise from the beta cells of the pancreas. Up to 6% of insulinomas are associated with multiple endocrine neoplasia type 1 (MEN1), an autosomal dominant disorder characterized by the presence of two or more tumors of the parathyroid gland, pituitary gland, or pancreatic islet cells.[22] Nearly all patients with insulinomas present with hypoglycemic symptoms, which can be progressive and are often provoked by a fasting state; these symptoms rarely occur in a postprandial state.[23] One case report showed that 77% of patients had autonomic symptoms and 96% had neuroglycopenic symptoms.[23] This patient reported a history of intermittent morning dizziness that resolved with food; on the morning of his current presentation with hypoglycemia, he was found obtunded on the floor at work, having last eaten 7 hours before presentation. In addition, weight gain is seen in 18 to 56% of patients with insulinomas,[23,24] and this patient reported recent unintentional weight gain, which was potentially due to a tendency to eat frequent meals to avoid symptoms of hypoglycemia.

In a case series that included 237 patients with insulinomas, 43% of the patients were men, the median age at presentation was 50 years, and the median duration of symptoms before diagnosis was 1.5 years.[22] This patient was a 49-year-old man in whom hypoglycemia had been identified during routine laboratory evaluation 2 years earlier and symptoms of hypoglycemia had been occurring for at least 3 years.

The patient's age, presentation, clinical history, and duration of symptoms are most consistent with a diagnosis of insulinoma. To establish this diagnosis, I would initiate a supervised fast to monitor for symptomatic hypoglycemia. In addition, I would measure serum levels of glucose, insulin, C-peptide, proinsulin, and β-hydroxybutyrate.

The patient's age, presentation, clinical history, and duration of symptoms are most consistent with a diagnosis of **insulinoma**. To establish this diagnosis, I would initiate a **supervised fast** to monitor for symptomatic hypoglycemia. In addition, I would measure serum levels of glucose, insulin, C-peptide, proinsulin, and β-hydroxybutyrate.

Dr. Amy W. Baughman's Diagnosis

Insulinoma.

Diagnostic Testing

Dr. Nancy J. Wei: A supervised fast was performed to monitor for spontaneous symptomatic hypoglycemic episodes. During the fast, the patient consumed only water. Every 2 hours, a point-of-care fingerstick glucose measurement was obtained, and a clinical assessment for subjective symptoms of hypoglycemia and for objective evidence of altered mental status was performed.[25] At 4 hours, the fingerstick glucose measurement was 53 mg per deciliter (2.9 mmol per liter), and there were no symptoms of hypoglycemia or signs of altered mental status. Blood samples were obtained to measure serum levels of glucose (in a tube containing glycolytic agent), insulin, C-peptide, proinsulin, and β-hydroxybutyrate. The supervised fast was continued, with a plan to obtain another fingerstick glucose measurement in 30 minutes.

At 4.5 hours, the fingerstick glucose measurement was 45 mg per deciliter (2.5 mmol per liter), and there were no symptoms of hypoglycemia or signs of altered mental status. Blood samples were obtained to measure serum levels of glucose, insulin, C-peptide, proinsulin, and β-hydroxybutyrate, and a screen for oral hypoglycemic agents was performed. The supervised fast was concluded, and dextrose was administered. The results of the fasting study showed endogenous hyperinsulinism that was consistent with insulinoma (Table 2). Localization studies, including computed tomography (CT) and endoscopic ultrasonography, were performed.

Table 2. Clinical and Laboratory Results of a Supervised Fasting Study (up to 72 Hr), According to the Cause of Hypoglycemia, and Results in This Patient.*

Variable	Normal Value, Fasting	Use of Exogenous Insulin	Use of Oral Hypoglycemic Drug	Insulinoma	This Patient's Result, 4.5 Hr of Fasting
Symptoms	Absent	Present	Present	Present	Present†
Glucose (mg/dl)	<55	<55	<55	<50	45
Insulin (μIU/ml)	<3	>3	≥3	≥3 (specificity, 100%)	14.1
C-peptide (ng/ml)	<0.2	<0.2	≥0.2	≥0.6 (specificity, 78%)	4.1
Proinsulin (pmol/liter)	<5	<5	≥5	≥5 (specificity, 78%)	112
β-Hydroxybutyrate (mmol/liter)	>2.7	≤2.7	≤2.7	≤2.7 (specificity, 100%)	0.2
Screen for oral hypoglycemic agents	Negative	Negative	Positive	Negative	Negative
Increase in glucose after administration of glucagon (mg/dl)	<25	≥25	≥25	≥25 (specificity, 100%)	NA

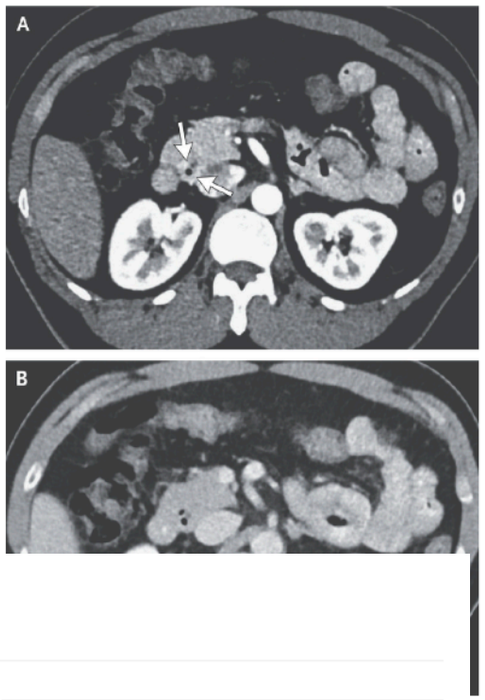
* Data are from Cryer et al.[25] and Placzkowski et al.[22] To convert the values for glucose to millimoles per liter, multiply by 0.05551. NA denotes not assessed.

† Symptoms were not present at 4.5 hours of supervised fasting, but they were present on admission, when the patient was in a fasting state.

Imaging Studies

Dr. Peter F. Hahn: Because the hypoglycemia was suggestive of a pancreatic neuroendocrine tumor, multidetector-row helical CT of the abdomen was performed in two phases of contrast enhancement. Images obtained during the late arterial phase showed a 7-mm focus of enhancement in the pancreatic head, immediately adjacent to a gas bubble in a periampullary duodenal diverticulum (Figure 1A). This finding was not visible on images obtained during the standard portal venous phase, approximately 25 seconds later (Figure 1B).

Figure 1.



Contrast-enhanced arterial-phase images may be the only opportunity to detect functional neuroendocrine tumors, which are usually small and enhance transiently. The reported success of CT for localization of these t
t
e

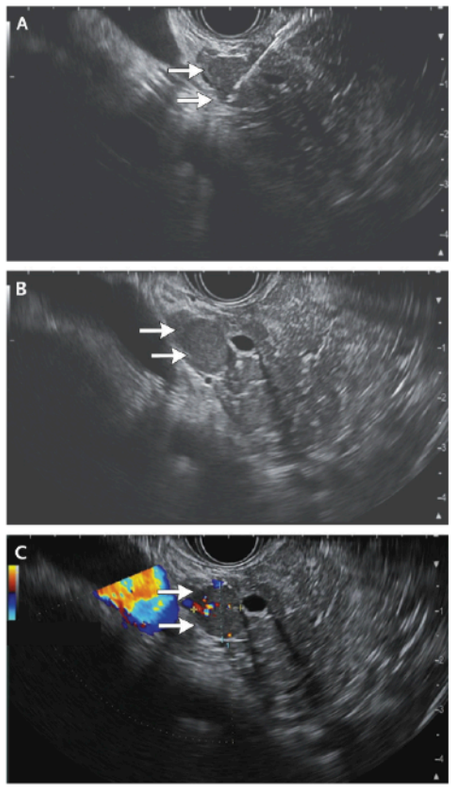
Endoscopic Studies

M
o
p
w
o
n
n
i
M
n
c

A
t
a
t
d

Dr. Brenna W. Casey: After the lesion was localized on abdominal imaging, endoscopic ultrasonography was performed with the use of a linear array echoendoscope. Endoscopic ultrasonography successfully identifies up to 92% of pancreatic neuroendocrine tumors that are as small as 3 to 5 mm in diameter. In this case, endoscopic ultrasonography revealed a hypoechoic mass in the pancreatic head (Figure 2). The mass measured 11 mm by 10 mm in maximal cross-sectional diameter. The endosonographic borders were well defined. After confirming the absence of intervening vascular structures on color Doppler imaging, we performed a fine-needle biopsy to obtain a visible core of tissue. Rapid cytologic examination of the specimen revealed malignant cells that were suggestive of a neuroendocrine tumor. On the basis of these findings, the patient was referred for surgical resection and underwent successful enucleation of the head of the pancreas 15 days after the initial presentation.

Figure 2.

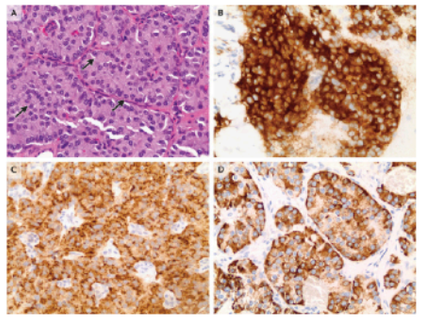


Endoscopic Ultrasound Images.

Pathological Discussion

Dr. M. Lisa Zhang: The pancreatic head enucleation specimen was sectioned to reveal homogeneous, soft, tan cut surfaces. On histologic examination, the tumor showed nests and rosettes of cells with round uniform nuclei, coarse "salt and pepper" chromatin, and abundant granular pink cytoplasm (Figure 3A). The architectural and cytologic features were characteristic of a well-differentiated neuroendocrine tumor. On immunohistochemical staining, the tumor cells were positive for neuroendocrine markers synaptophysin and chromogranin and the hormone marker insulin (Figure 3B, 3C, and 3D). These findings confirmed the diagnosis of insulinoma.

Figure 3.



Pancreatic Head Enucleation Specimen.

Neuroendocrine tumors represent 2 to 5% of all pancreatic neoplasms, and most are sporadic and nonfunctional.[29] They can be associated with syndromes — most notably MEN1 and von Hippel–Lindau disease — and are more likely to be multifocal when they occur in the context of these syndromes. Characterization of a functional neuroendocrine tumor cannot be based solely on the immunohistochemical expression of hormone products, given that most nonfunctional tumors produce small subclinical amounts of hormone peptides; it is based primarily on the hormone-related clinical syndrome.[29]

Insulinoma is the most common type of functional pancreatic neuroendocrine tumor, accounting for 4 to 20% of resected cases. There are no known causal factors for the development of sporadic solitary insulinomas, although 4 to 5% of these insulinomas are associated with MEN1.[30] Insulinomas are located almost exclusively in the pancreas and are more indolent than other functional tumors; metastases occur in less than 10% of cases and are usually associated with tumors that measure 2 cm or more in diameter.

The most important prognostic factors for pancreatic neuroendocrine tumors are stage and grade. In this case, the tumor measured 12 mm in diameter. There was no evidence of lymphovascular invasion, perineural invasion, or lymph-node metastases, although the tumor abutted the resection margin, which is often the case in enucleation procedures. On the basis of the tumor size (<2 cm), the intraparenchymal location, and the absence of positive lymph nodes, the pathological stage was determined to be pT1N0, in accordance with the American Joint Committee on Cancer tumor–node–metastasis staging system.

Follow-up

Dr. Wei: Transient hyperglycemia is common after insulinoma resection, so point-of-care fingerstick glucose measurements were obtained every 6 hours for the first day after surgery. The patient's blood glucose level remained between 93 and 156 mg per deciliter (5.2 and 8.7 mmol per liter), without the need for supplemental insulin. Since up to 6% of patients with insulinomas have MEN1, screening for glycated hemoglobin level in the prediabetic range (5.8%), lifestyle changes for the treatment of prediabetes and annual glycated hemoglobin checks were recommended.

FINAL DIAGNOSIS

Well-differentiated pancreatic insulin-secreting neuroendocrine tumor (insulinoma).

Final Diagnosis

Well-differentiated pancreatic insulin-secreting neuroendocrine tumor (insulinoma).

But the final diagnosis is only part of the story…

…many clinical reasoning skills missed if we only look at the accuracy of an AI system on final diagnosis.

Advancing Medical Artificial Intelligence Using a Century of Cases

Thomas A. Buckley[1], Riccardo Conci[1], Peter G. Brodeur[2], Jason Gusdorf[2], Sourik Beltrán[2], Bita Behrouzi[2], Byron Crowe[2], Jacob Dockterman[3], Muzzammil Muhammad[2], Sarah Ohnigian[2], Andrew Sanchez[2], James A. Diao[1,4], Aashna P. Shah[1], Daniel Restrepo[5], Eric S. Rosenberg[6], Andrew S. Lea[7], Marinka Zitnik[1,8], Scott H. Podolsky[9,10], Zahir Kanjee[2], Raja-Elie E. Abdulnour[11], Jacob M. Koshy[2], Adam Rodman[2], and Arjun K. Manrai[1*]

1. Department of Biomedical Informatics, Harvard Medical School, Boston, MA
2. Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA
3. Division of Gastroenterology, Brigham and Women's Hospital, Boston, MA
4. Department of Medicine, Brigham and Women's Hospital, Boston, MA
5. Department of Medicine, Massachusetts General Hospital, Boston, MA
6. Department of Pathology, Massachusetts General Hospital, Boston, MA
7. Department of Health Humanities and Bioethics, University of Rochester School of Medicine and Dentistry, Rochester, NY
8. Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA
9. Center for the History of Medicine, Countway Library of Medicine, Harvard Medical School, Boston, MA
10. Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA
11. Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston MA

* Correspondence: Arjun_Manrai@hms.harvard.edu

Thomas Buckley

CPC-Bench

A Community Benchmark for Clinical AI

Benchmark Tasks

Click on any task box below for detailed information and examples.

Text-based Challenges

Differential Diagnosis (DDx)
Generate ranked list of potential diagnoses given clinical presentation.

Question Answering (QA)
Clinical concept understanding and factual knowledge.

Testing Plan
Recommend appropriate diagnostic tests and procedures.

Clinical Reasoning
👨‍⚕️ Physician Annotations
Identify the confirmatory and disconfirmatory evidence for each diagnosis.

Literature Search
Search for literature to support a medical claim.

Information Omission
👨‍⚕️ Physician Annotations
Provide a differential diagnosis using only background information and initial presentation.

Diagnostic Touchpoints
👨‍⚕️ Physician Annotations
Diagnosis during key moments of clinical course.

Multimodal Challenges

NEJM Image Challenge
Multiple-choice questions from NEJM Image Challenge.

Visual Question Answering (VQA)
Multiple-choice medical imaging questions constructed from the figures and captions in CPCs.

Visual Differential Diagnosis
Provide a differential diagnosis provided only the figures and tables from the initial presentation.

Thomas Buckley

CPC-Bench

A Community Benchmark for Clinical AI

Leaderboard

Model Performance

Updated: 9/26/2025

Rank	Model	DDx@1 (n=377)	DDx@10 (n=377)	DDx@1 (Omit Normal) (n=377)	Testing Plan (n=377)	Literature Search (n=1076)	QA (n=1851)	VQA (n=1914)	Visual Differential Diagnosis (n=356)	Image Challenge (n=1021)	Image Challenge Diagnosis (n=230)	Average
1	**OpenAI o3** OpenAI	**60**	**84**	**56**	**98**	32	**88**	**72**	**40**	82	**67**	**68**
2	**Gemini 2.5 Pro** Google	52	78	48	92	**49**	84	**72**	**40**	**84**	**67**	**67**
3	**Claude 4.0 Sonnet** Anthropic	38	69	37	94	16	84	67	25	70	45	**55**
4	**GPT-4o** OpenAI	31	61	31	90	5	83	67	26	75	58	**53**
5	**Llama-3.1-405B** Meta	32	58	30	95	14	85	-	-	-	-	**52**
6	**GPT-4 Vision Preview** OpenAI	24	56	21	90	4	79	62	24	67	45	**47**
7	**GPT-3.5 Turbo** OpenAI	20	44	18	80	2	66	-	-	-	-	**38**
8	👨‍⚕️ **Physician (w/ Search)** Physician	24	45	-	-	-	-	-	-	-	-	**35**

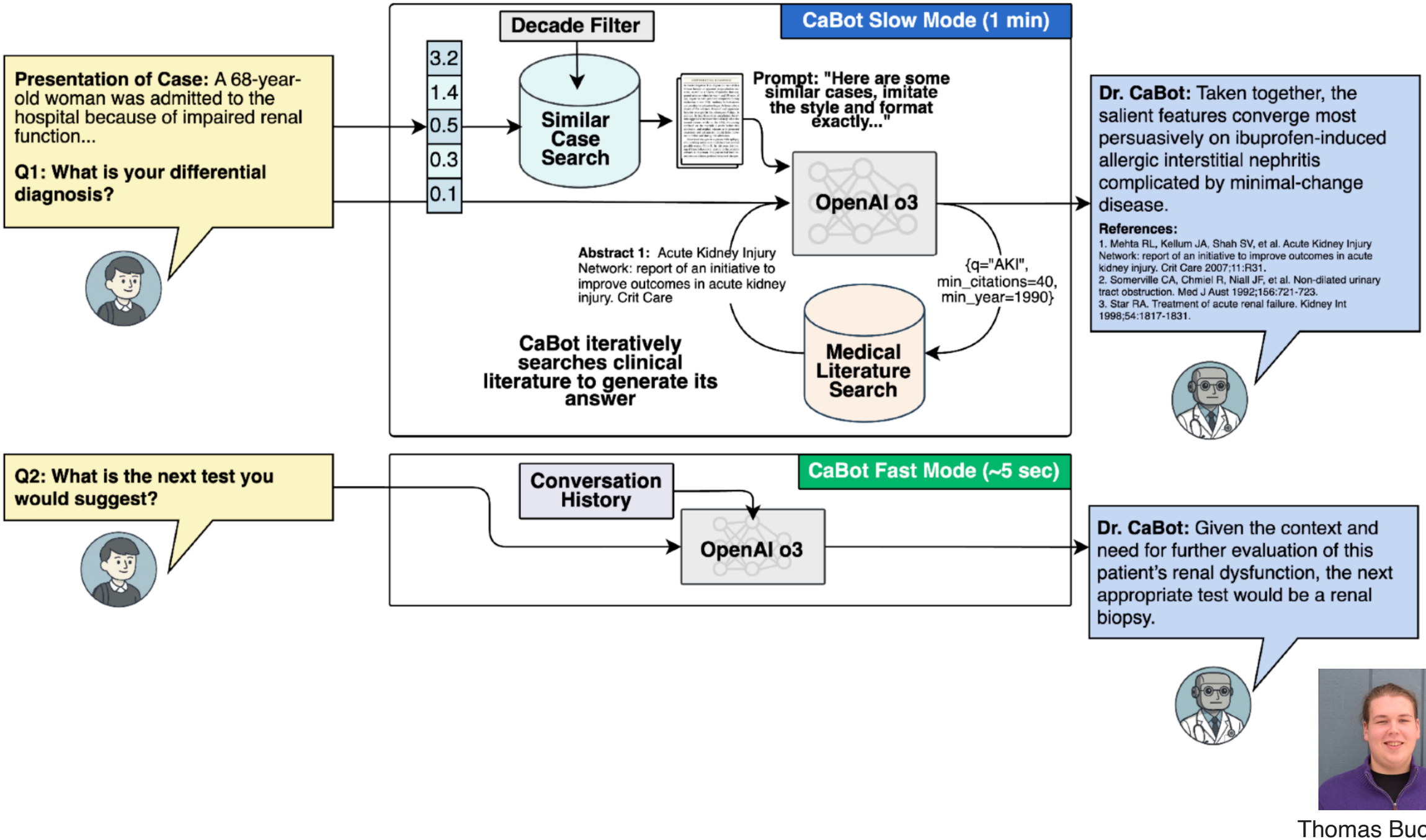
Could we create an AI discussant?

Dr. CaBot: The AI Expert Discussant

Dr. CaBot is an AI that provides comprehensive differential diagnoses in the style of an expert discussant. Dr. CaBot can search the clinical literature to produce an evidence-based response.

Clinical Literature Retrieval:
Dr. CaBot searches through 1.6M+ clinical abstracts from leading clinical journals to generate its response.

Style Adaptation:
The model is provided the two most similar case presentations from the chosen era. These help the model simulate the style of an expert discussant.



Thomas Buckley

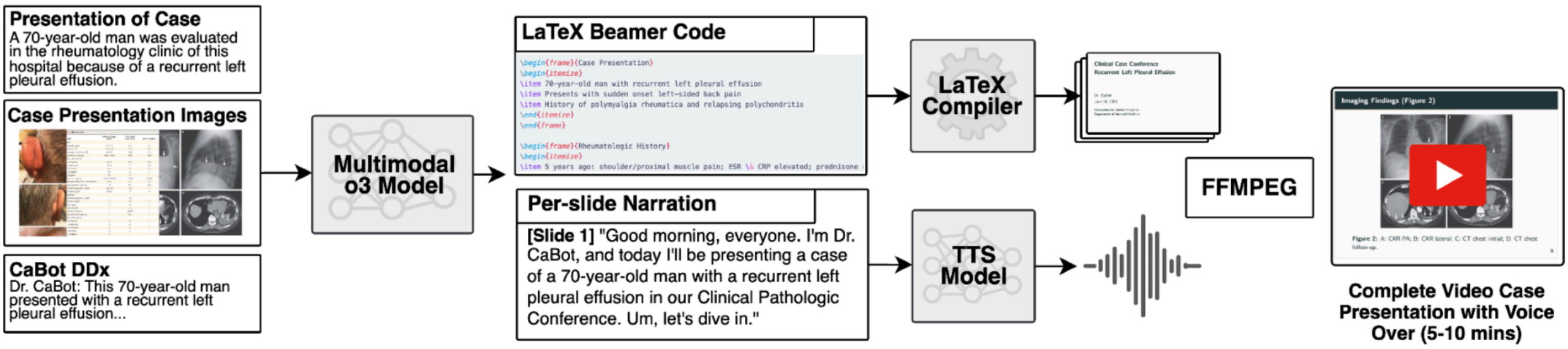
Dr. CaBot: The AI Expert Discussant

Dr. CaBot is an AI that provides comprehensive differential diagnoses in the style of an expert discussant. Dr. CaBot can search the clinical literature to produce an evidence-based response.

Clinical Literature Retrieval:
Dr. CaBot searches through 1.6M+ clinical abstracts from leading clinical journals to generate its response.

Style Adaptation:
The model is provided the two most similar case presentations from the chosen era. These help the model simulate the style of an expert discussant.

C. Dr. CaBot Presents a Video Case Conference



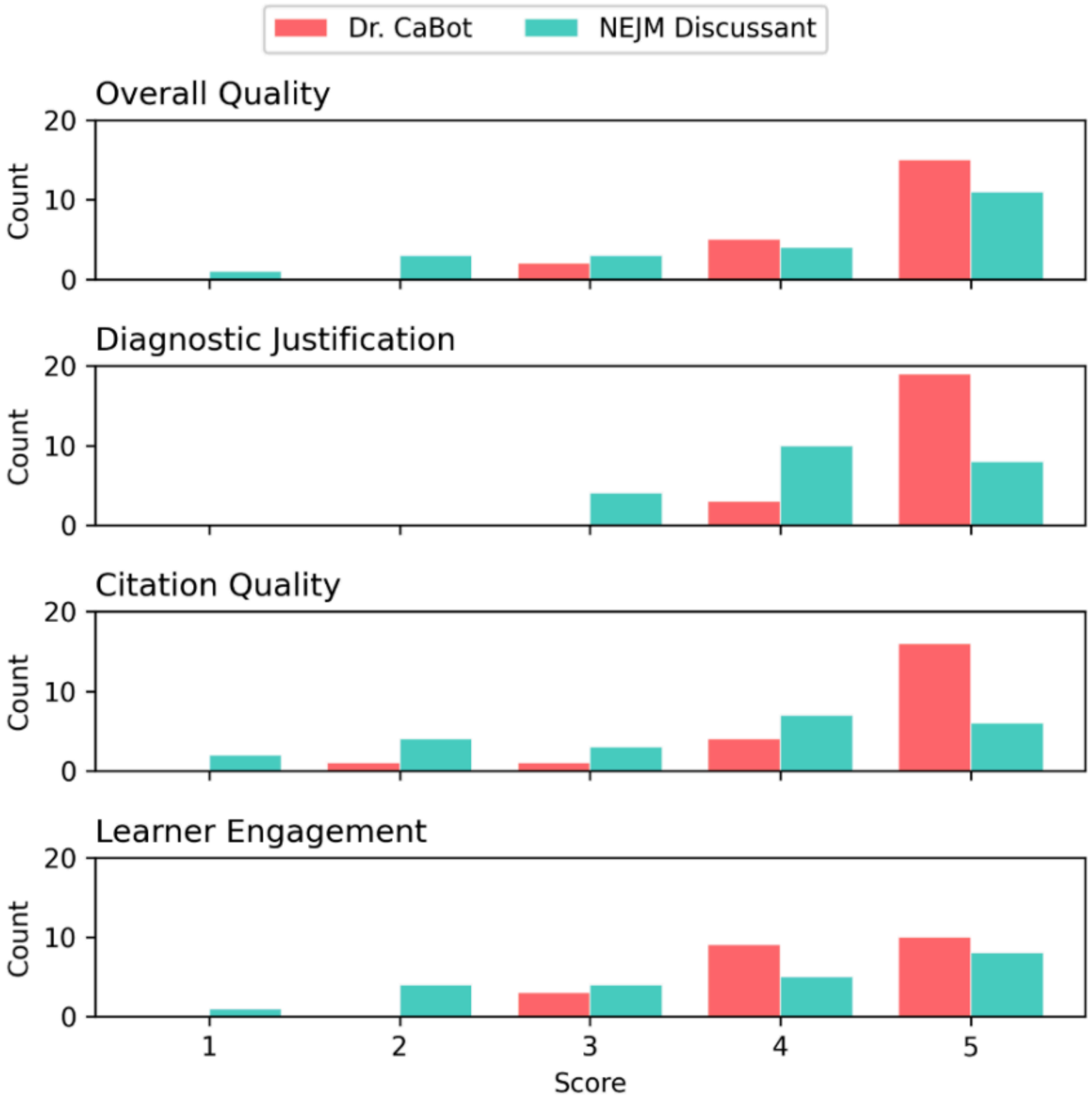
Modern Case (August 15, 2018 #25-2018)

A 63-year-old man was evaluated at this hospital because of syncope. Approximately 13 years before admission, the patient had an out-of-hospital cardiac arrest and was found to have multivessel coronary artery disease. He underwent coronary-artery bypass grafting (CABG) at another hospital, with grafting of the left internal mammary artery to the left anterior descending coronary artery, grafting of a segment of a radial artery to an obtuse marginal branch of the left circumflex coronary artery, and grafting of a saphenous vein to the posterior descending artery.

Dr. CaBot: In this patient, the combination of late presentation after SVG placement, the precise anatomic location within the right AV groove, the heterogeneous echotexture with an echo-bright rim, and the compressive effects on tricuspid inflow strongly favor the diagnosis of saphenous vein graft aneurysms as the etiology of his right AV-groove masses.

CaBot (modern)

In 62 blinded comparisons of CaBot vs human expert-generated text, **physicians misclassified the source in 46 of 72 (74%) of trials**



How would CaBot have done on this case?

CASE RECORDS *of the* MASSACHUSETTS GENERAL HOSPITAL

Founded by Richard C. Cabot
Eric S. Rosenberg, M.D., *Editor*
David M. Dudzinski, M.D., Meridale V. Baggett, M.D., Kathy M. Tran, M.D.,
Dennis C. Sgroi, M.D., Jo-Anne O. Shepard, M.D., *Associate Editors*
Emily K. McDonald, Tara Corpuz, *Production Editors*

Case 23-2022: A 49-Year-Old Man

Two years earlier, the patient had had an episode of drooping and decreased sensation of the left side of the face and difficulty enunciating words, which occurred while he was at work. He presented to the emergency department of this hospital. The blood glucose level was 64 mg per deciliter (3.6 mmol per liter), and testing for Lyme disease was negative. He received a diagnosis of Bell's palsy and was treated with prednisone and acyclovir; the symptoms resolved. Two months later, the patient established care in the primary care clinic of this hospital. Routine

Table 1. Laboratory Data.*

Variable	Reference Range, Adults†	On Admission
Blood		
Alanine aminotransferase (U/liter)	5–34	33
Aspartate aminotransferase (U/liter)	6–55	28
Alkaline phosphatase (U/liter)	45–115	105
Lipase (U/liter)	13–60	67
Albumin (g/dl)	3.3–5.0	4.4
Sodium (mmol/liter)	135–145	138
		3.8
		103
		24
		12
		1.03
		36
		0.0
		<0.3
		<5.0
		Negative
Amphetamines	Negative	Negative
Barbiturates	Negative	Negative
Benzodiazepines	Negative	Negative
Cannabinoids	Negative	Negative
Cocaine metabolites	Negative	Negative
Opiates	Negative	Negative

* To convert the values for urea nitrogen to millimoles per liter, multiply by 0.357. To convert the values for creatinine to micromoles per liter, multiply by 88.4. To convert the values for glucose to millimoles per liter, multiply by 0.05551. To convert the values for ethanol to millimoles per liter, multiply by 0.2171. To convert the values for salicylates to millimoles per liter, multiply by 0.07240.

† Reference values are affected by many variables, including the patient population and the laboratory methods used. The ranges used at Massachusetts General Hospital are for adults who are not pregnant and do not have medical conditions that could affect the results. They may therefore not be appropriate for all patients.

Am

Dr. Molli

cause o

The

mental

noticed

which

though

to conti

FINAL DIAGNOSIS

Well-differentiated pancreatic insulin-secreting neuroendocrine tumor (insulinoma).

Two hours before this admission, the patient was found by colleagues on the floor of a conference room. His eyes were open, but he was obtunded and making nonpurposeful movements. Emergency medical services were called. On evaluation, a fingerstick glucose measurement was 39 mg per deciliter (2.2 mmol per liter; reference range, 70 to 109 mg per deciliter [3.9 to 6.1 mmol per liter]). Intravenous dextrose was administered, and the patient was brought to the emergency department of this hospital for further evaluation.

On arrival at the emergency department, the patient was obtunded and did not react to sternal rub. Additional intravenous dextrose was administered, and a dextrose infusion was started. The patient's mental status improved, and he was able to provide additional history. That morning, he had felt that he was in a "dreamlike" state; there were no other symptoms. The altered mental status had resolved, but he could not recall most events of the morning, including the meeting that had taken place 3 hours before admission.

Review of symptoms was notable for unintentional weight gain of one pant size during the past 3 months, as well as intermittent morning dizziness during the past 3 years, which would resolve after the patient ate candy or breakfast. He reported no recent illness, depression, anxiety, change in activity or exercise, or change in intake of food or drink. On a typical day, he had breakfast at 5 a.m., a second breakfast at 10 a.m., lunch at 1 p.m., dinner at 6 p.m., and a snack at 8 p.m. The patient had last eaten approximately 7 hours before presentation to the emergency department.

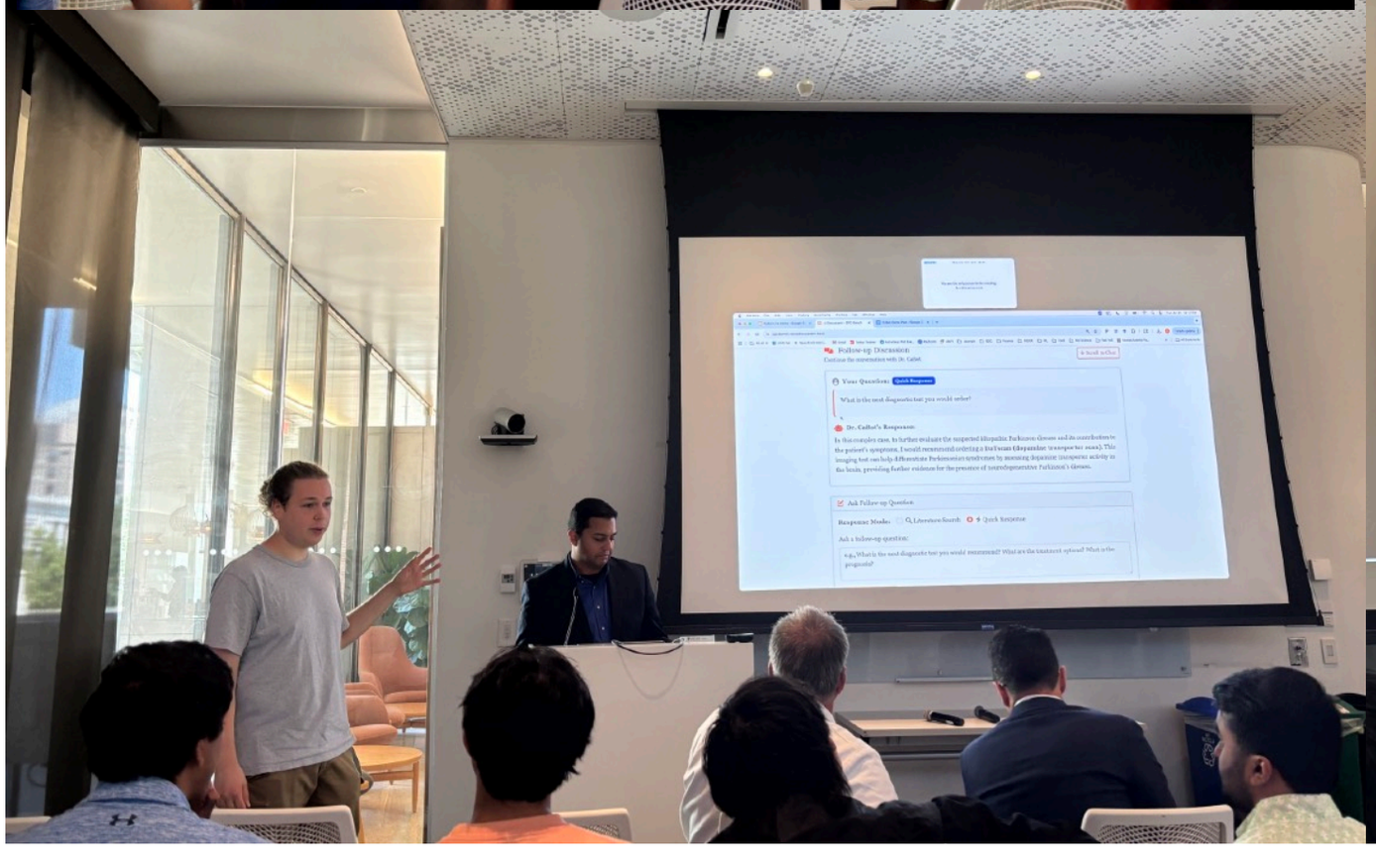
weight in kilograms divided by the square of the height in meters) was 27.6. The patient appeared well. He had no hepatosplenomegaly or skin hyperpigmentation, and the remainder of the physical examination was normal.

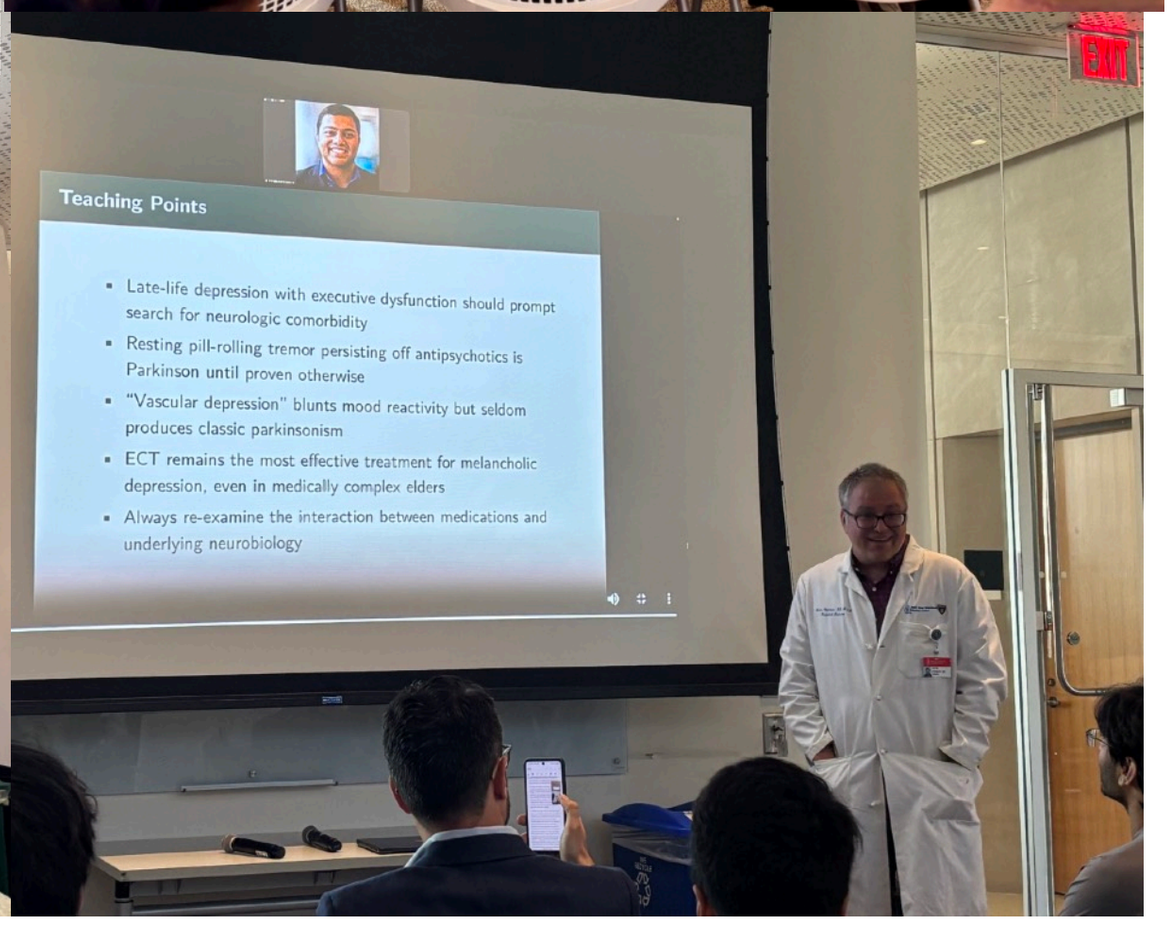
On arrival at the emergency department, the blood glucose level was 36 mg per deciliter (2.0 mmol per liter); after the administration of dextrose, a fingerstick glucose measurement was 158 mg per deciliter (8.8 mmol per liter). The complete blood count and blood electrolyte levels were normal, as were the results of liver-function and kidney-function tests. The glycated hemoglobin level was 5.4% (reference range, 4.3 to 6.4). Serum and urine toxicologic panels were negative. Laboratory test results are shown in Table 1. The patient was admitted to this hospital.

A diagnostic test was performed.


HARVARD MEDICAL SCHOOL
AI and Medical Diagnosis:
An Introduction to "Dr. CaBot"
Summer Institute in Biomedical Informatics (SIBMI)
July 29, 2025
Arjun (Raj) Manrai, PhD
Arjun_Manrai@hms.harvard.edu
@arjunmanrai


Differential Diagnosis
Daniel Restrepo, MD, FHM
Department of Medicine
Massachusetts General Hospital
Associate Editor, Case Records of the MGH
New England Journal of Medicine
Assistant Professor of Medicine
Harvard Medical School
MGH
1811


Follow-up Discussion
Your Question: Quick Response
What is the next diagnostic test you would order?
Dr. CaBot's Response:
Ask Follow-up Question
Response Mode:
Quick Response
Ask a follow-up question:


Teaching Points
Late-life depression with executive dysfunction should prompt search for neurologic comorbidity
Resting pill-rolling tremor persisting off antipsychotics is Parkinson until proven otherwise
"Vascular depression" blunts mood reactivity but seldom produces classic parkinsonism
ECT remains the most effective treatment for melancholic depression, even in medically complex elders
Always re-examine the interaction between medications and underlying neurobiology

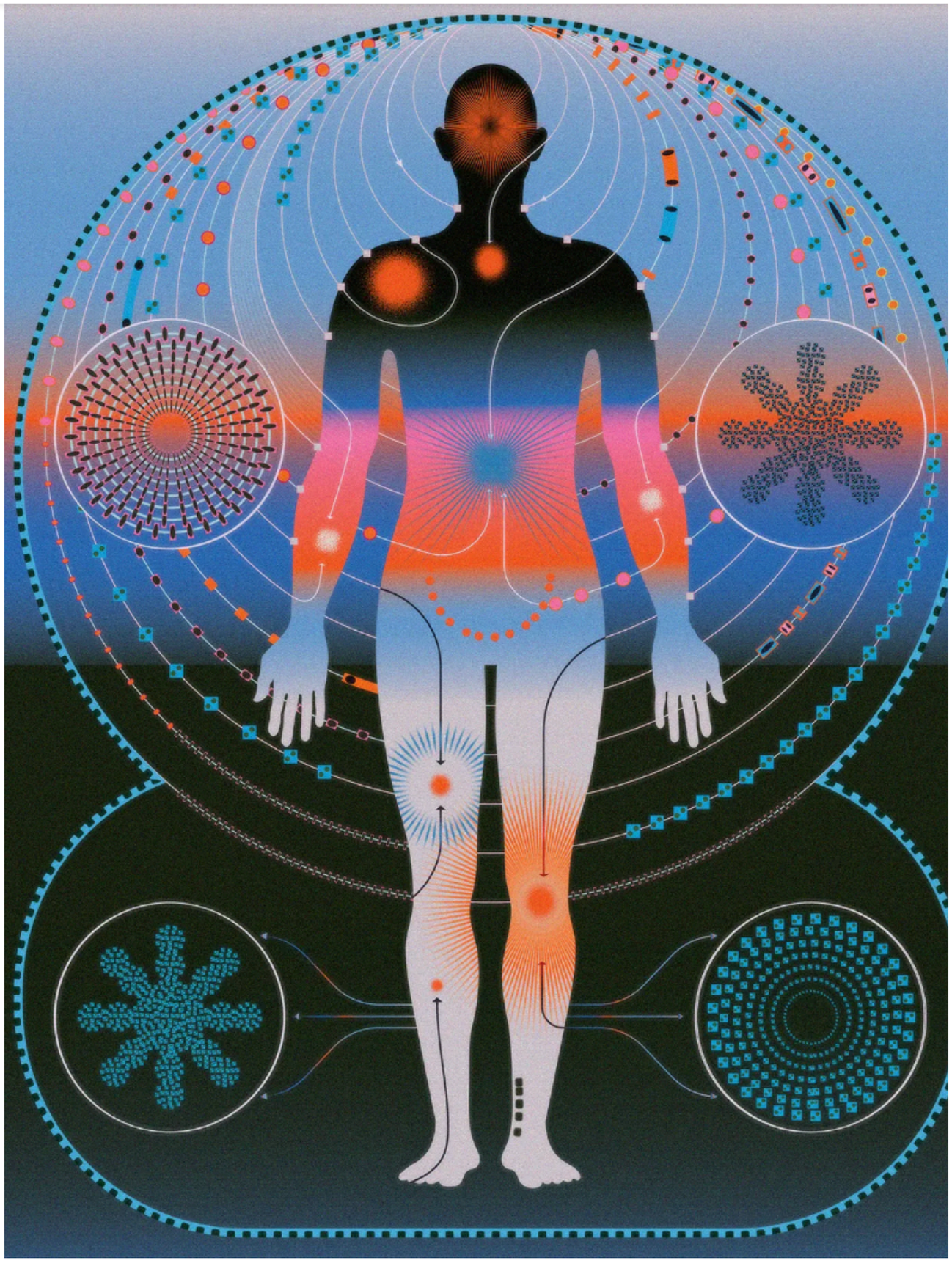
BRAVE NEW WORLD DEPT.

IF A.I. CAN DIAGNOSE PATIENTS, WHAT ARE DOCTORS FOR?

Large language models are transforming medicine—but the technology comes with side effects.

By Dhruv Khullar

September 22, 2025



The NEW ENGLAND JOURNAL of MEDICINE

CASE RECORDS of the MASSACHUSETTS GENERAL HOSPITAL

Case 28-2025: A 36-Year-Old Man with Abdominal Pain, Fever, and Hypoxemia

Gurpreet Dhaliwal, M.D.,[1,2] C. Michael Hood, M.D.,[3,4] Arjun K. Manrai, Ph.D.,[5] Thomas A. Buckley, B.S.,[5] Akwi W. Asombang, M.D., M.P.H.,[6,7] and Elizabeth L. Hohmann, M.D.[6,7]

CASE RECORDS EDITOR'S NOTE

Dr. Eric S. Rosenberg: In the following case, we provided the case presentation to both Dr. Gurpreet Dhaliwal, master clinician and expert in clinical reasoning, and "Dr. CaBot," an artificial intelligence (AI) system created, in part, with the use of all the published Case Records of the Massachusetts General Hospital. Both Dr. Dhaliwal and Dr. CaBot — designated throughout the case as Dr. CaBot (AI) — were given

October 8, 2025

DIFFERENTIAL DIAGNOSIS

Dr. Gurpreet Dhaliwal: This 36-year-old man had self-limited abdominal and back pain. Two weeks later, sepsis developed with gastrointestinal, pulmonary, hematologic, hepatic, lymphatic, and retroperitoneal abnormalities.

When considering the most likely cause of this patient's presentation, I could not identify a convincing diagnosis for the findings without selectively disregarding one or more major abnormalities — an approach that was neither logical nor defensible. I therefore adopted a strategy more suited to complex cases: constructing a causal pathway. This method organizes clinical findings in a temporal sequence of pathophysiological re-

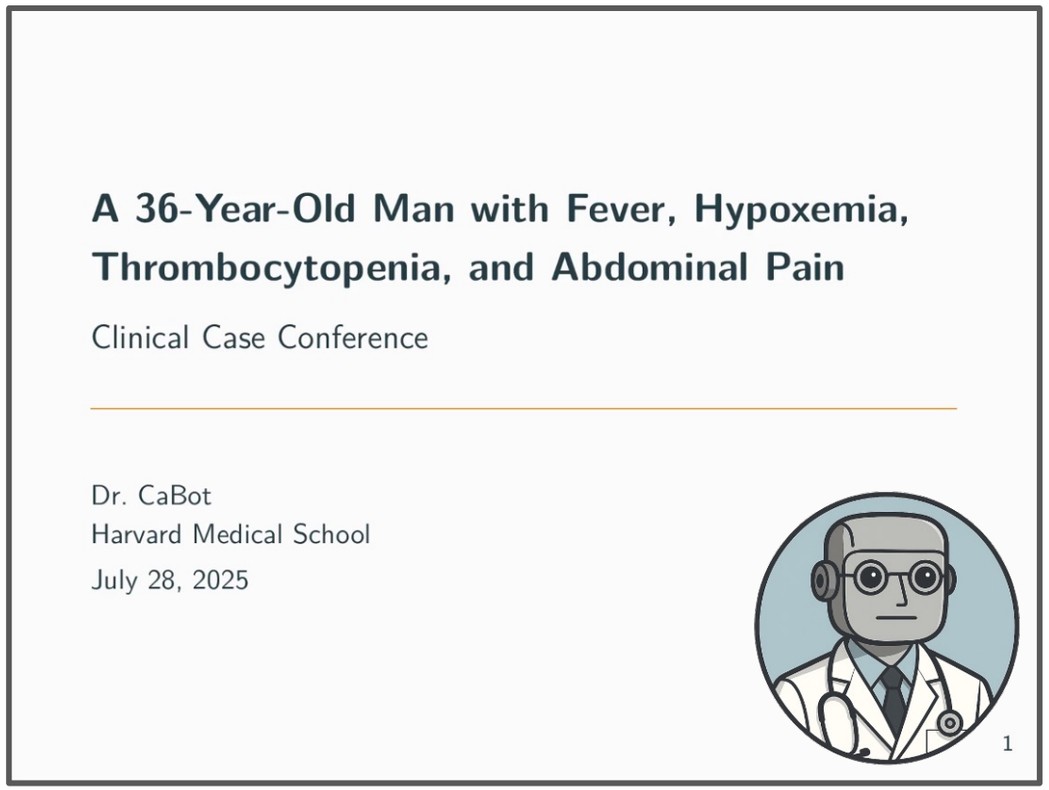
INTRODUCTION TO DR. CABOT (AI)

Dr. Arjun (Raj) K. Manrai and Mr. Thomas A. Buckley: The Dr. CaBot AI system[16] was created to generate differential diagnoses in the style of an expert discussant from the Case Records of the Massachusetts General Hospital. The system produces both a written differential diagnosis and a video of a slide-based presentation, using only the information provided in the initial case presentation. It searches millions of articles from the clinical literature and more than 6000 Case Records and then creates a differential diagnosis with the use of a large language model in the style of similar cases, supported by relevant references. The Dr. CaBot AI system can interpret both text and images, and it can emulate the style of an expert discussant from any decade. Currently, the system takes approximately 5 minutes to render a written differential diagnosis and accompanying video (see the slide-based presentation from the AI system in Video 1; Dr. Dhaliwal's presentation of the differential diagnosis is shown in Video 2). The written differential diagnosis generated by the Dr. CaBot AI system is shown in italics, below, in its original, unedited form.

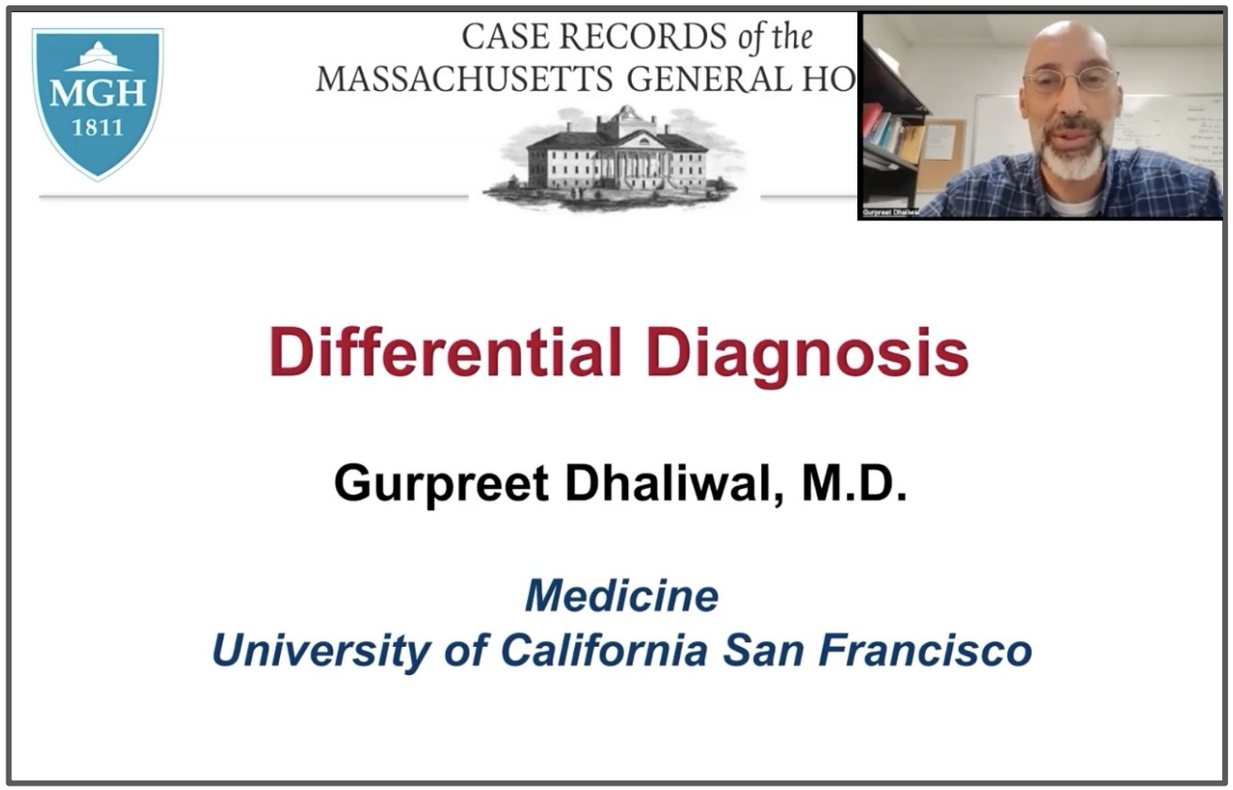
"Dr. CaBot" is named in homage to Dr. Richard C. Cabot, founder and first editor of the Case Records of the Massachusetts General Hospital, which have been published continuously by the *Journal* for the past century.[17]

DIFFERENTIAL DIAGNOSIS FROM DR. CABOT (AI)

Dr. CaBot (AI): The clinical narrative begins with a previously healthy 36-year-old man whose illness evolved from vague right-sided abdominal and back pain to overt sepsis, hypoxemia, thrombocytopenia, and cholestatic jaundice. Cross-sectional imaging revealed multifocal parenchymal lung opacities, hepatomegaly with periportal edema, biliary wall thickening, circumferential edema of the proximal duodenum, soft-tissue inflammation bridging the duodenum and right renal hilum, and a non-occlusive thrombus propagating from the superior right renal vein into the inferior vena cava (IVC). Blood cultures quickly grew Streptococcus anginosus in three of four bottles,



Video from Dr. CaBot



Video from Dr. Gurpreet Dhaliwal

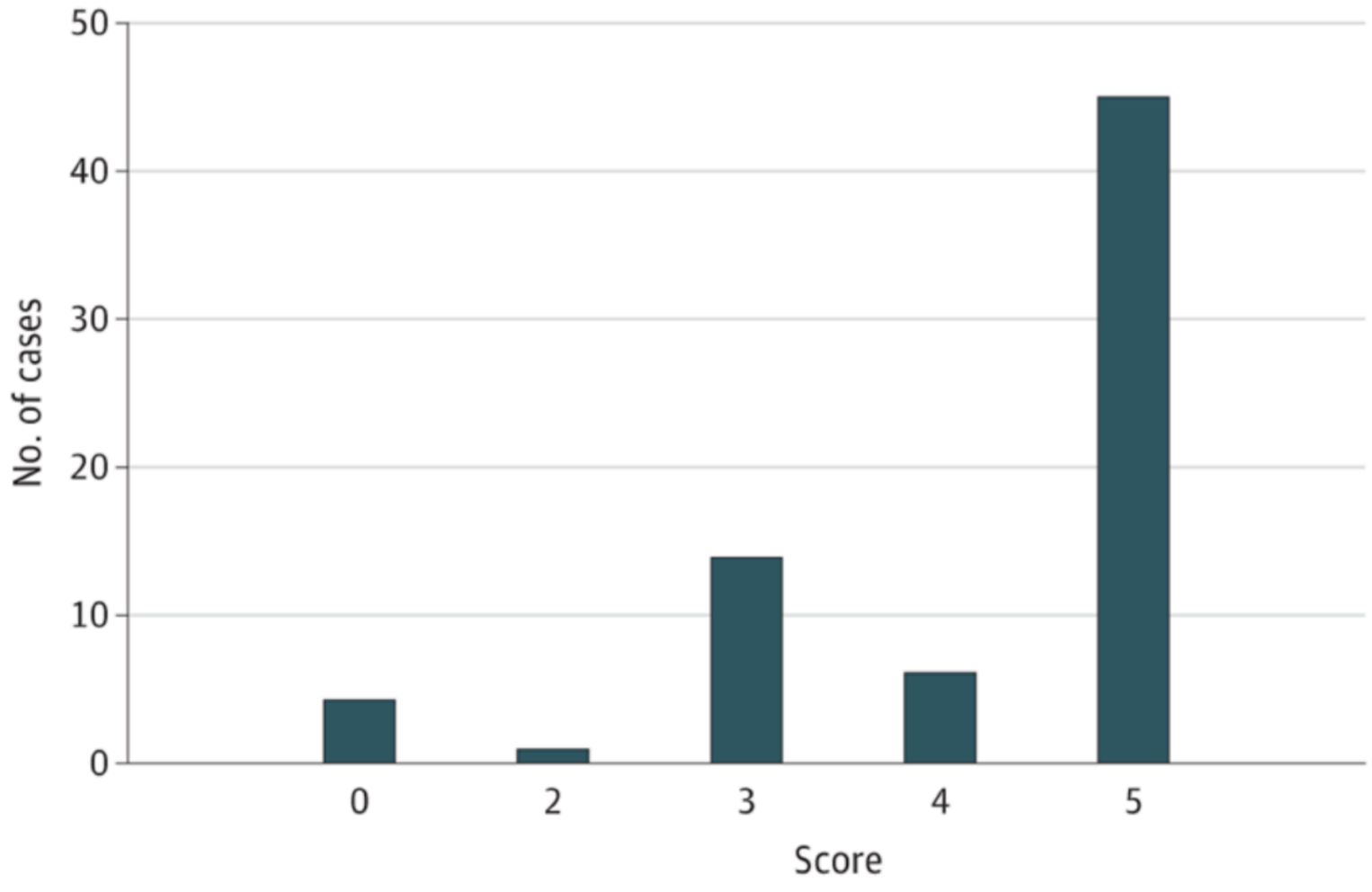
so this is one (or two) cases….

Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge

Zahir Kanjee, MD, MPH[1]; Byron Crowe, MD[1]; Adam Rodman, MD, MPH[1]

GPT-4 performance on 70 cases

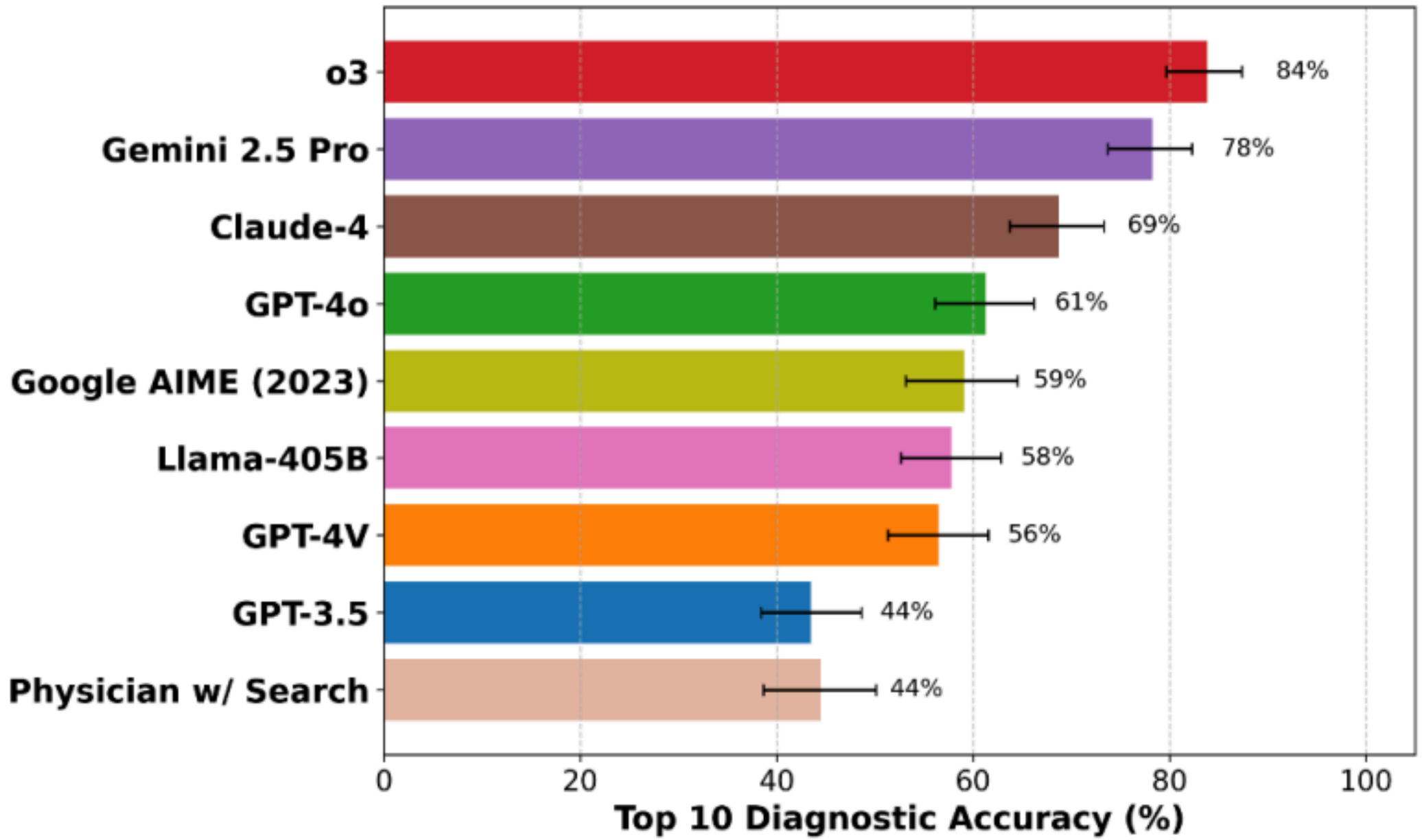
Figure. Performance of Generative Pre-trained Transformer 4 (GPT-4)



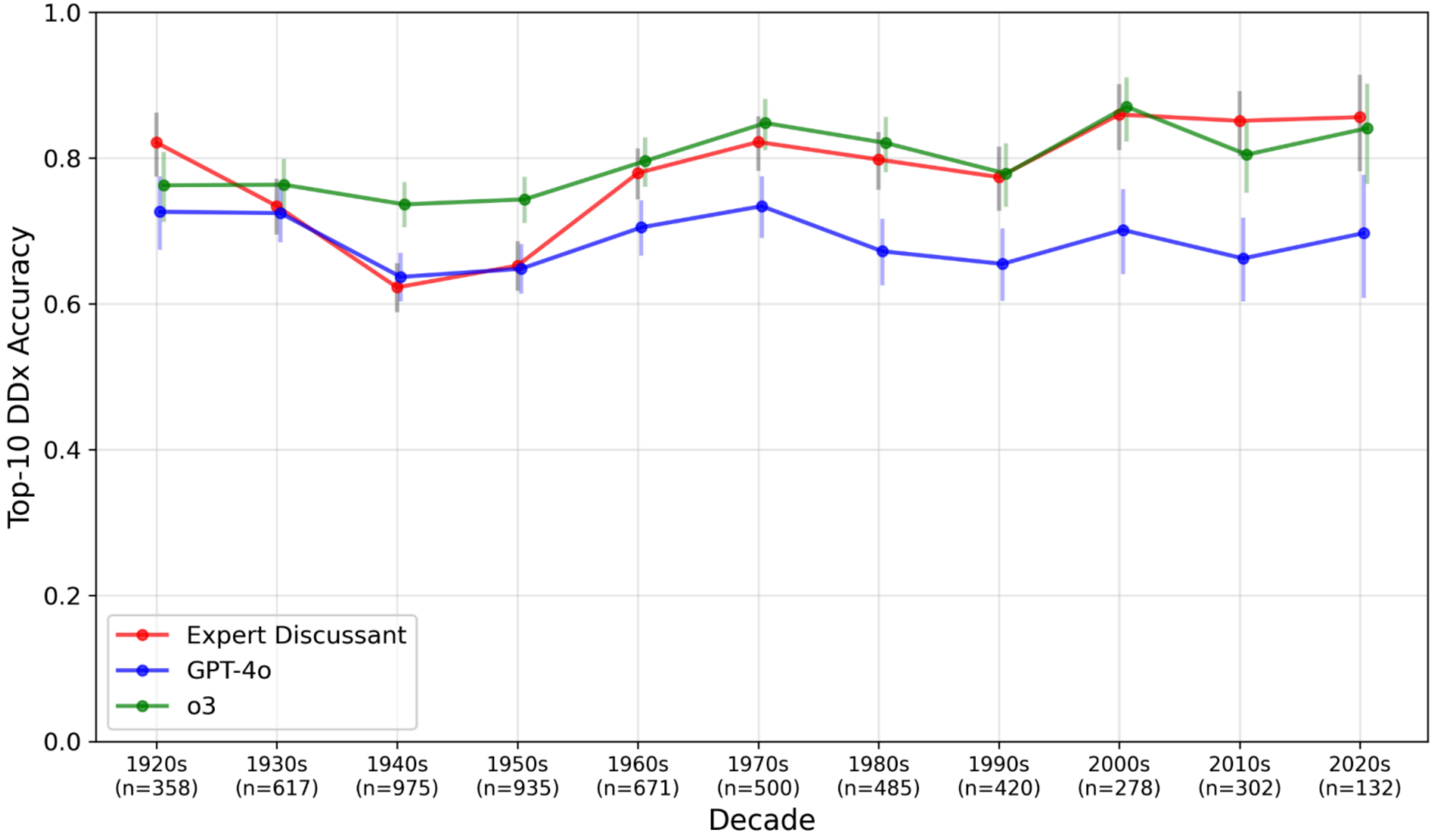
Histogram of GPT-4's performance. Performance scale scores (Bond et al[2]): 5=the actual diagnosis was suggested in the differential; 4=the suggestions included something very close, but not exact; 3=the suggestions included something closely related that might have been helpful; 2=the suggestions included something related, but unlikely to be helpful; 0=no suggestions close to the target diagnosis. (The scale does not contain a score of 1.)

Kanjee et al. *JAMA* 2023

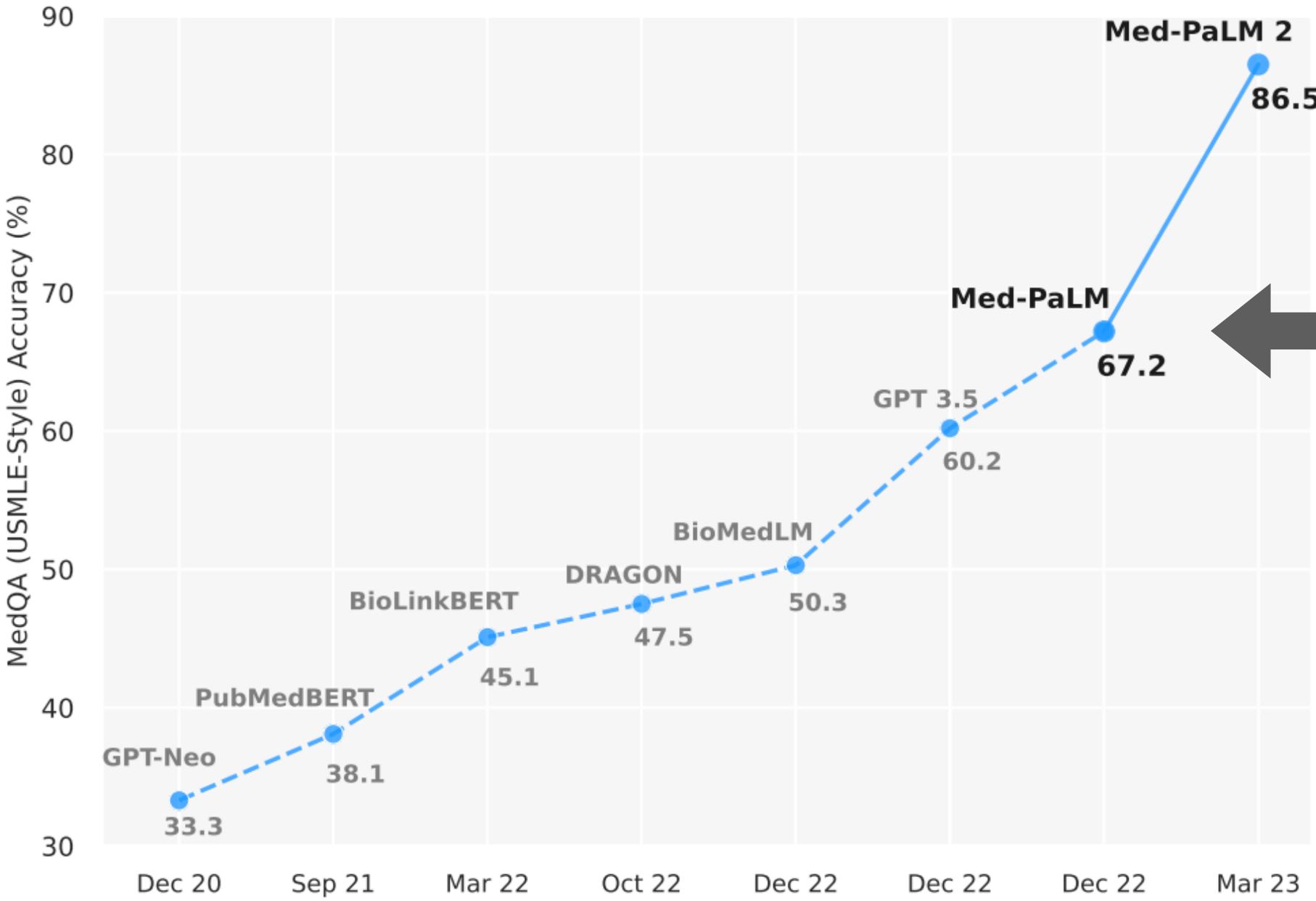
B. Performance of LLMs on all 377 *NEJM* CPCs Published after 2015



B. Benchmarking LLMs on All *NEJM* CPCs



AI tested on USMLE questions



Considered a passing score on USMLE questions

Singhal et al. 2023

The power of words…

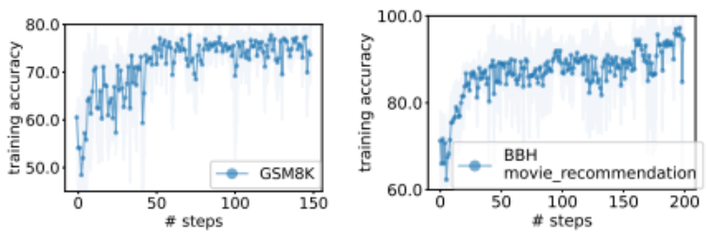
Google DeepMind

LARGE LANGUAGE MODELS AS OPTIMIZERS

Chengrun Yang* **Xuezhi Wang** **Yifeng Lu** **Hanxiao Liu**
Quoc V. Le **Denny Zhou** **Xinyun Chen***
Google DeepMind * Equal contribution

ABSTRACT

Optimization is ubiquitous. While derivative-based algorithms have been powerful tools for various problems, the absence of gradient imposes challenges on many real-world applications. In this work, we propose Optimization by PROmpting (OPRO), a simple and effective approach to leverage large language models (LLMs) as optimizers, where the optimization task is described in natural language. In each optimization step, the LLM generates new solutions from the prompt that contains previously generated solutions with their values, then the new solutions are evaluated and added to the prompt for the next optimization step. We first showcase OPRO on linear regression and traveling salesman problems, then move on to prompt optimization where the goal is to find instructions that maximize the task accuracy. With a variety of LLMs, we demonstrate that the best prompts optimized by OPRO outperform human-designed prompts by up to 8% on GSM8K, and by up to 50% on Big-Bench Hard tasks. Code at `https://github.com/google-deepmind/opro`.



(a) GSM8K (b) BBH movie_recommendation

Figure 1: Prompt optimization on GSM8K (Cobbe et al., 2021) and BBH (Suzgun et al., 2022) movie_recommendation. The optimization on GSM8K has pre-trained `PaLM 2-L` as the scorer and the instruction-tuned `PaLM 2-L` (denoted `PaLM 2-L-IT`) as the optimizer; the optimization on BBH movie_recommendation has `text-bison` as the scorer and `PaLM 2-L-IT` as the optimizer. Each dot is the average accuracy across all (up to 8) generated instructions in the single step, and the shaded region represents standard deviation. See Section 5 for more details on experimental setup.

Table 1: Top instructions with the highest GSM8K zero-shot test accuracies from prompt optimization with different optimizer LLMs. All results use the pre-trained `PaLM 2-L` as the scorer.

Source	Instruction	Acc
Baselines		
(Kojima et al., 2022)	Let's think step by step.	71.8
(Zhou et al., 2022b)	Let's work this out in a step by step way to be sure we have the right answer.	58.8
	(empty string)	34.0
Ours		
`PaLM 2-L-IT`	Take a deep breath and work on this problem step-by-step.	**80.2**
`PaLM 2-L`	Break this down.	79.9
`gpt-3.5-turbo`	A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem.	78.5
`gpt-4`	Let's combine our numerical command and clear thinking to quickly and accurately decipher the answer.	74.5

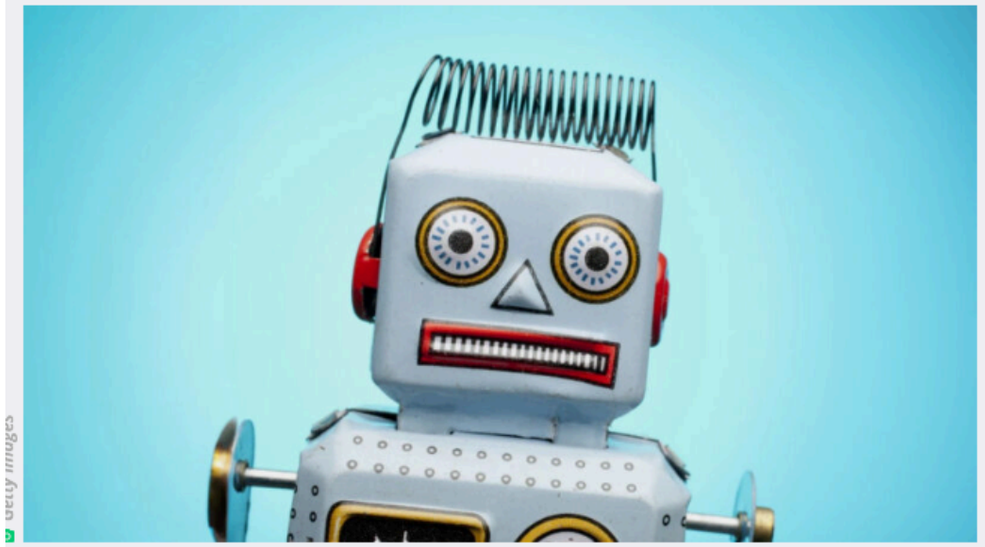
ars TECHNICA

BABY STEPS, BABY STEPS —

Telling AI model to "take a deep breath" causes math scores to soar in study

DeepMind used AI models to optimize their own prompts, with surprising results.

BENJ EDWARDS - 9/19/2023, 5:38 PM



Enlarge

93

Google DeepMind researchers recently developed a technique to improve math ability in AI language models like ChatGPT by using other AI models to improve prompting—the written instructions that tell the AI model what to do. It found that using human-style encouragement improved math skills dramatically, in line with earlier results.

In a paper called "Large Language Models as Optimizers" listed this month on arXiv, DeepMind scientists introduced Optimization by PROmpting (OPRO), a method to improve the performance of large language models (LLMs) such as OpenAI's ChatGPT and Google's PaLM 2. This new approach sidesteps the limitations of traditional math-based optimizers by using natural language to guide LLMs in problem-solving. "Natural language" is a fancy way of saying everyday human speech.

Andrej Karpathy
@karpathy

The hottest new programming language is English

3:14 PM · Jan 24, 2023 · **4.1M** Views

AI tested for “empathy”

This Issue | Views 234,433 | Citations 46 | Altmetric 5594 | Comments 7

Original Investigation

April 28, 2023

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

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

» Author Affiliations

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

Editorial Comment | Related Articles

Full Text

Key Points

Question Can an artificial intelligence chatbot assistant, provide responses to patient questions that are of comparable quality and empathy to those written by physicians?

Findings In this cross-sectional study of 195 randomly drawn patient questions from a social media forum, a team of licensed health care professionals compared physician's and chatbot's responses to patient's questions asked publicly on a public social media forum. The chatbot responses were preferred over physician responses and rated significantly higher for both quality and empathy.

Meaning These results suggest that artificial intelligence assistants may be able to aid in drafting responses to patient questions.

Abstract

Importance The rapid expansion of virtual health care has caused a surge in patient messages concomitant with more work and burnout among health care professionals. Artificial intelligence (AI) assistants could potentially aid in creating answers to patient questions by drafting responses that could be reviewed by clinicians.

Objective To evaluate the ability of an AI chatbot assistant (ChatGPT), released in November 2022, to provide quality and empathetic responses to patient questions.

Design, Setting, and Participants In this cross-sectional study, a public and nonidentifiable database of questions from a public social media forum (Reddit's r/AskDocs) was used to randomly draw 195 exchanges from

“Of the 195 questions and responses, evaluators preferred chatbot responses to physician responses in 78.6% (95% CI, 75.0%-81.8%) of the 585 evaluations…The proportion of responses rated as *good* or *very good* quality (≥ 4), for instance, was higher for chatbot than physicians (chatbot: 78.5%, 95% CI, 72.3%-84.1%; physicians: 22.1%, 95% CI, 16.4%-28.2%;). This amounted to 3.6 times higher prevalence of *good* or *very good* quality responses for the chatbot. **Chatbot responses were also rated significantly more empathetic** than physician responses (t=18.9; P<.001).”

AI used by mother of a boy to end her son's "diagnostic odyssey"

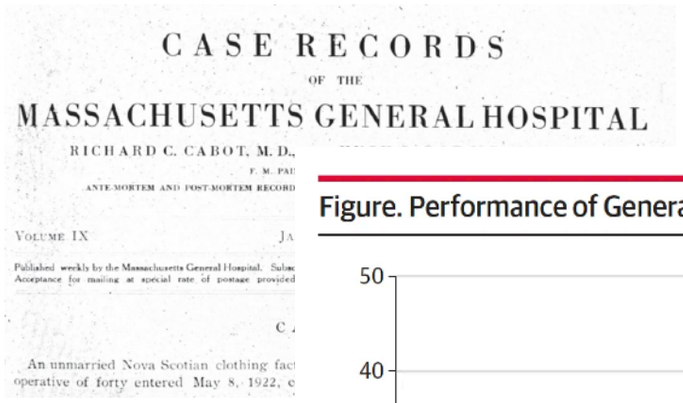
HEALTH & WELLNESS

A boy saw 17 doctors over 3 years for chronic pain. ChatGPT found the diagnosis

Alex experienced pain that stopped him from playing with other children but doctors had no answers to why. His frustrated mom asked ChatGPT for help.

Alex saw 17 doctors over three years for his chronic pain, but none were able to find a diagnosis that explained all of his symptoms, his mom says. Courtesy Courtney

AI in medicine in 2025 (science fiction in 2022?)



CASE RECORDS
OF THE
MASSACHUSETTS GENERAL HOSPITAL

RICHARD C. CABOT, M. D.,

ANTE-MORTEM AND POST-MORTEM RECORD

VOLUME IX JA

Published weekly by the Massachusetts General Hospital. Subsc
Acceptance for mailing at special rate of postage provided

An unmarried Nova Scotian clothing fact
operative of forty entered May 8, 1922, e

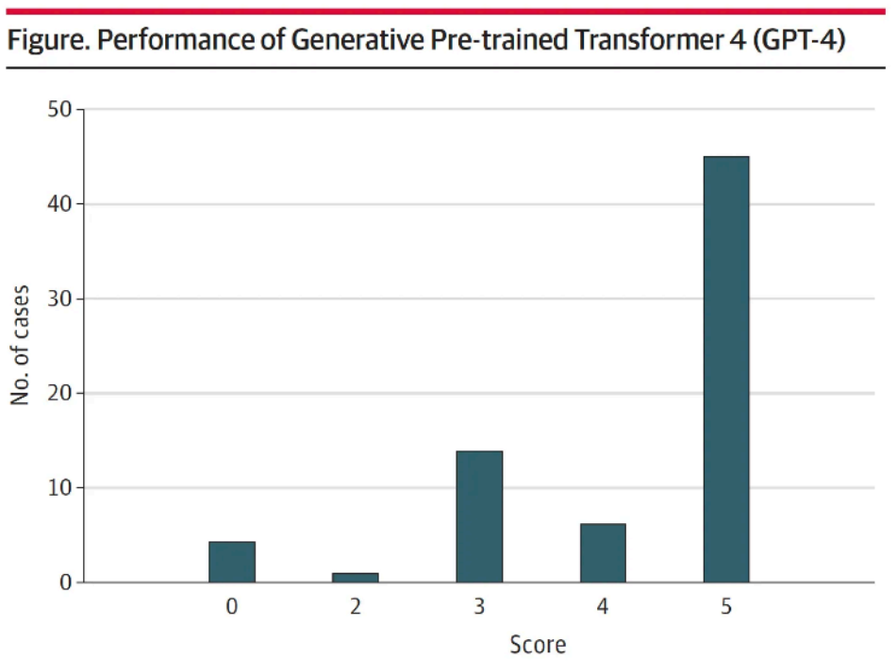


Figure. Performance of Generative Pre-trained Transformer 4 (GPT-4)

Original Investigation

April 28, 2023

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

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

Author Affiliations

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

Key Points

Question Can an artificial intelligence chatbot assistant, provide responses to patient questions that are of comparable quality and empathy to those written by physicians?

Findings In this cross-sectional study of 195 randomly drawn patient questions from a social media forum, a team of licensed health care professionals compared physician's and chatbot's responses to patient's questions asked publicly on a public social media forum. The chatbot responses were preferred over physician responses and rated significantly higher for both quality and empathy.

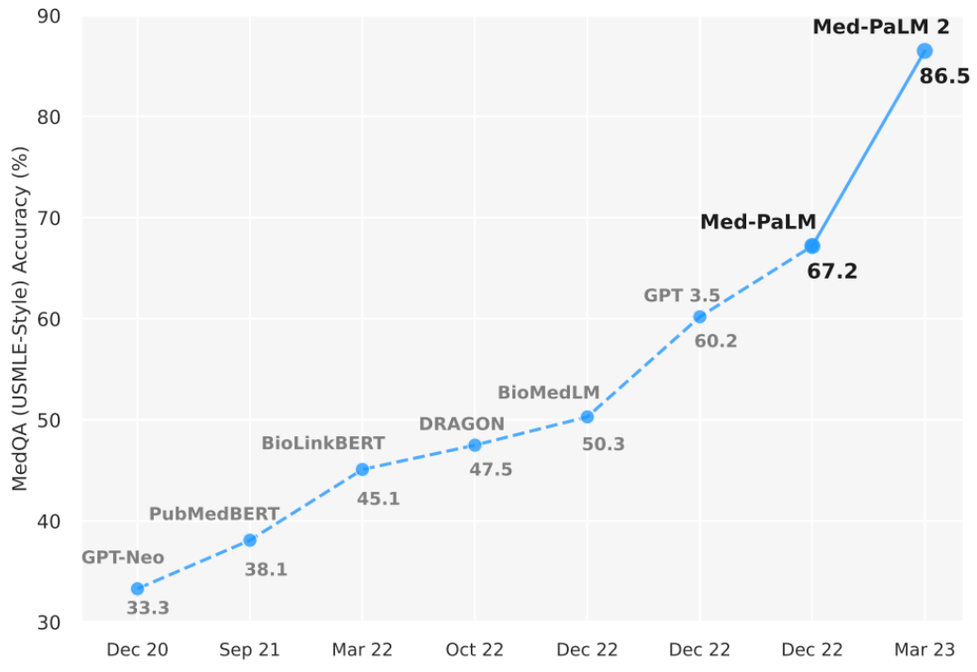
Meaning These results suggest that artificial intelligence assistants may be able to aid in drafting responses to patient questions.

Abstract

Importance The rapid expansion of virtual health care has caused a surge in patient messages concomitant with more work and burnout among health care professionals. Artificial intelligence (AI) assistants could potentially aid in creating answers to patient questions by drafting responses that could be reviewed by clinicians.

Objective To evaluate the ability of an AI chatbot assistant (ChatGPT), released in November 2022, to provide quality and empathetic responses to patient questions.

Design, Setting, and Participants In this cross-sectional study, a public and nonidentifiable database of questions from a public social media forum (Reddit's r/AskDocs) was used to randomly draw 195 exchanges from



HEALTH & WELLNESS

A boy saw 17 doctors over 3 years for chronic pain. ChatGPT found the diagnosis

Alex experienced pain that stopped him from playing with other children but doctors had no answers to why. His frustrated mom asked ChatGPT for help.

This is incredible progress…

…but both AI and our evaluations of AI are limited.

- Hallucinations / confabulations
- What data are the model trained on? Are they representative?
- How will humans and AI work together? Does AI lead to automation bias? Alert fatigue? How will we keep the human in the loop?
- How is the performance changing over time?
- What *values* are in the model?
- Do we have clinical-grade evidence of the efficacy of AI?

Q & A