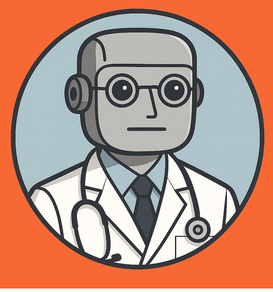


Dr. CaBot: The AI Expert Discussant

Thomas A Buckley, Riccardo Conci, Peter G Brodeur, Jason Gusdorf, Sourik Beltrán, Bita Behrouzi, Byron Crowe, Jacob Dockterman, Muzzammil Muhammad, Sarah Ohnigian, Andrew Sanchez, James A Diao, Aashna P Shah, Daniel Restrepo, Eric S Rosenberg, Andrew S Lea, Emily Glanton, Kimberly LeBlanc, Undiagnosed Diseases Network, Marinka Zitnik, Scott H Podolsky, Zahir Kanjee, Raja-Elie E Abdulnour, Jacob M Koshy, Adam Rodman, Arjun K Manrai

HARVARD MEDICAL SCHOOL | BLAVATNIK INSTITUTE BIOMEDICAL INFORMATICS

1. Background

Prior work has shown that LLMs can generate accurate diagnoses based on clinical vignettes [1, 2].

Unknown if LLMs can:

1. Generate complete clinical reasoning with supporting literature citations, in both **written and live-presentation form.**
2. Generate accurate diagnoses for **patients referred to a rare disease center**, such as the Undiagnosed Diseases Network (UDN)

Figure 1: Example Clinicopathologic Conference

Key Findings

1) CaBot generated the first AI diagnosis published in the 100+ year history of the *NEJM* Clinicopathologic Conferences
2) Physicians can't distinguish CaBot from *NEJM* experts (26% accuracy)
3) 69% accurate on real UDN rare disease referrals

2. Architecture

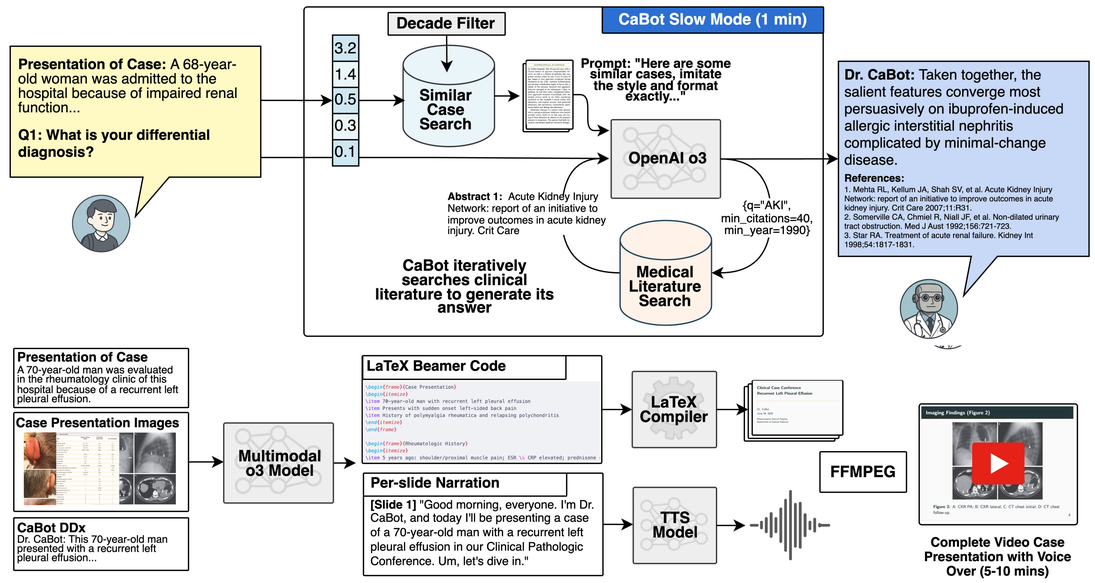


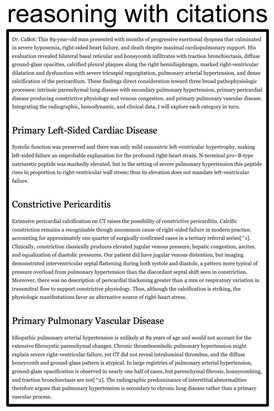
Figure 2: CaBot architecture figure.

3. CaBot Example

A. Prior AI: ranked list of diseases

1. Community-acquired pneumonia
2. Acute bronchitis
3. Pulmonary embolism
4. Congestive heart failure exacerbation
5. COVID-19 pneumonitis

B. CaBot: complete reasoning with citations



Dr. CaBot: This 89-year-old man presented with months of progressive exertional dyspnea that culminated in severe hypoxemia, right-sided heart failure, and death despite maximal cardiopulmonary support. His evaluation revealed bilateral basal reticular and honeycomb infiltrates with traction bronchiectasis, diffuse ground-glass opacities, calcified pleural plaques along the right hemidiaphragm, marked right-ventricular dilatation and dysfunction with severe tricuspid regurgitation, pulmonary arterial hypertension, and dense calcification of the pericardium. These findings direct consideration toward three broad pathophysiologic processes: intrinsic parenchymal lung disease with secondary pulmonary hypertension, primary pericardial disease producing constrictive physiology and venous congestion, and primary pulmonary vascular disease. Integrating the radiographic, hemodynamic, and clinical data, I will explore each category in turn.

Primary Left-Sided Cardiac Disease

Systolic function was preserved and there was only mild concentric left-ventricular hypertrophy, making left-sided failure an improbable explanation for the profound right-heart strain. N-terminal pro–B-type natriuretic peptide was markedly elevated, but in the setting of severe pulmonary hypertension this peptide rises in proportion to right-ventricular wall stress; thus its elevation does not mandate left-ventricular failure.

Constrictive Pericarditis

Extensive pericardial calcification on CT raises the possibility of constrictive pericarditis. Calcific constriction remains a recognizable though uncommon cause of right-sided failure in modern practice, accounting for approximately one quarter of surgically confirmed cases in a tertiary referral series[4]. Clinically, constriction classically produces elevated jugular venous pressure, hepatic congestion, ascites, and equalization of diastolic pressures. Our patient did have jugular venous distention, but imaging demonstrated interventricular septal flattening during both systole and diastole, a pattern more typical of pressure overload from pulmonary hypertension than the discordant septal shift seen in constriction. Moreover, there was no description of pericardial thickening greater than 4 mm or respiratory variation in transmitral flow to support constrictive physiology. Thus, although the calcification is striking, the physiologic manifestations favor an alternative source of right-heart stress.

Primary Pulmonary Vascular Disease

Idiopathic pulmonary arterial hypertension is unlikely at 89 years of age and would not account for the extensive fibrocystic parenchymal changes. Chronic thromboembolic pulmonary hypertension might explain severe right-ventricular failure, yet CT did not reveal intraluminal thrombus, and the diffuse honeycomb and ground-glass pattern is atypical. In large registries of pulmonary arterial hypertension, ground-glass opacification is observed in nearly one half of cases, but parenchymal fibrosis, honeycombing, and traction bronchiectasis are not[5,6]. The radiographic predominance of interstitial abnormalities therefore argues that pulmonary hypertension is secondary to chronic lung disease rather than a primary vascular process.

4. CaBot Diagnosis Published in *NEJM*

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

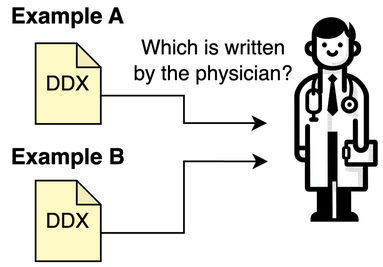
Gurpreet Dhaliwal, M.D.,[1,2] C. Michael Hood, M.D.,[3,4] Arjun K. Manrai, Ph.D.,[5] Thomas A. Buckley, B.S.,[5] Akwi W. Asombang, M.D., M.P.H.,[6,7] and Elizabeth L. Hohmann, M.D.[6,7]

CASE RECORDS EDITOR'S NOTE

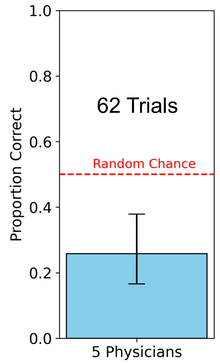
Dr. Eric S. Rosenberg: In the following case, we provided the case presentation to both Dr. Gurpreet Dhaliwal, master clinician and expert in clinical reasoning, and "Dr. CaBot," an artificial intelligence (AI) system created, in part, with the use of all the published Case Records of the Massachusetts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

5. Physician Evaluation

A. 5 physicians were assigned sets of (1) CaBot discussion and (2) the original discussion published in *NEJM*



B. Physician accuracy in classifying AI-generated vs physician-generated discussion



C. Physician quality ratings

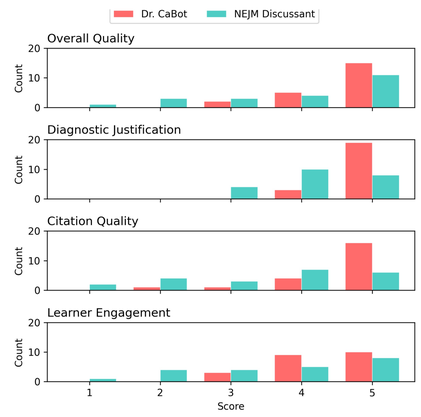
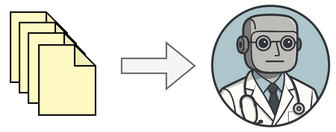


Figure 3: A. Schematic of experiment. **B.** Mean accuracy of 5 internal medicine physicians in identifying the source of a differential diagnosis as either AI or physician generated for 62 trials. **C.** Scores were provided for 44 total differential diagnoses by five internal medicine physicians who were blinded to the source of the differential.

6. Rare Disease



72 de-identified referral letters submitted to the Undiagnosed Diseases Network (UDN)

	All (n=72)	Diagnosed by UDN clinical/research evaluation (n=26)	Suggested diagnosis, not fully evaluated (n=46)
Correct disease category	50/72 (69.4%)	16/26 (61.5%)	34/46 (73.9%)
Specific diagnosis or gene named	46/72 (63.9%)	13/26 (50.0%)	33/46 (71.7%)

Figure 4: Summary of reviewer assessments for the 72 Undiagnosed Diseases Network (UDN) cases retained for analysis. Each CaBot response was reviewed by certified genetic counselors (K.L., E.G.).

References

1. Z. Kanjee, B. Crowe, and A. Rodman, "Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge," JAMA, vol. 330, no. 1, pp. 78–80, Jul. 2023.
2. P. G. Brodeur et al., "Performance of a large language model on the reasoning tasks of a physician," Science, vol. 392, no. 6797, pp. 524–527, Apr. 2026.
3. T. A. Buckley et al., "Advancing medical artificial intelligence using a century of cases," arXiv [cs.AI], Sep. 15, 2025. [Online]. Available: http://arxiv.org/abs/2509.12194
4. D. Khullar, "If A.I. Can Diagnose Patients, What Are Doctors For?," The New Yorker, The New Yorker, Sep. 22, 2025. Accessed: May 02, 2026. [Online].

Try CaBot at cpcbench.com