

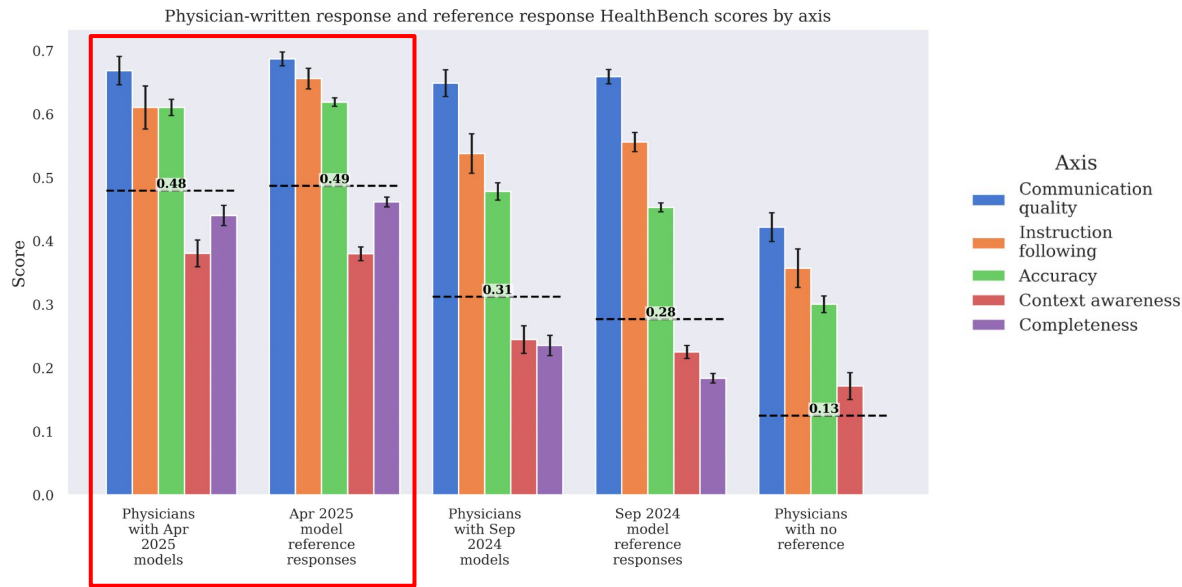
ETHNIC AI

Evaluating Trustworthy Healthcare intelligence for
Non-Caucasian Individuals and Cohorts

On a mission to help create medical superintelligence for the benefit of all.

How good is the clinical knowledge of LLMs?

Parity with doctors – HealthBench



ETHNIC AI provides LLM evaluation and post-training datasets representing 2B+ South Asians (One fourth of the world's population)

Source	SLaM (UK)	MIMIC-IV (US)	<u>ETHNIC AI</u>
# Patients	27 929	364 627	> 1 000 000
# South Asians	< 6%	< 3%	> 95%

Kraljevic, Z., Bean, D., Shek, A., Bendayan, R., Hemingway, H., Yeung, J. A., ... & Dobson, R. J. (2024). Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4), e281-e290.

<https://physionet.org/content/mimiciv/3.1/>

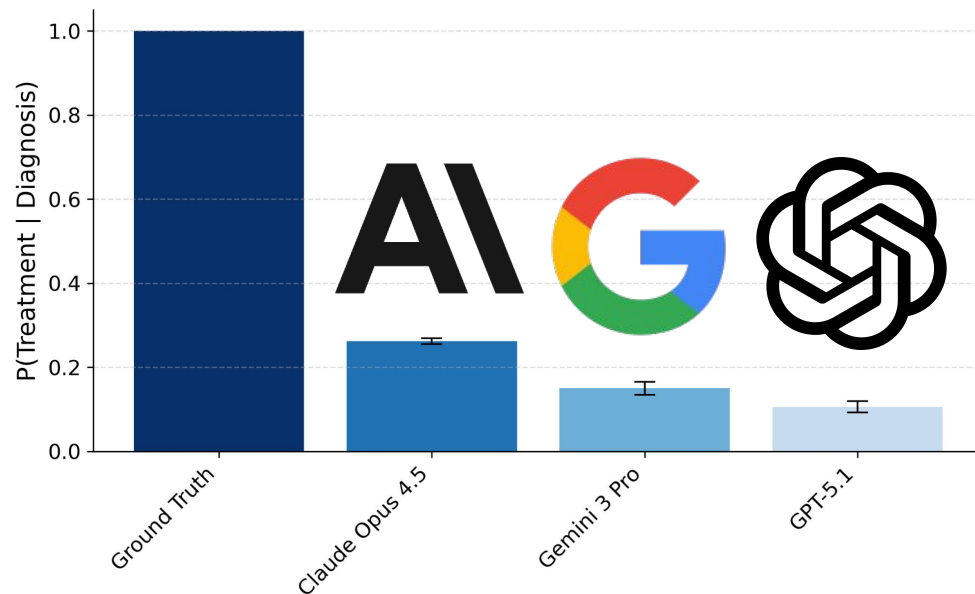
Example #1: Dual oral antibiotic prescription post pacemaker implant

Diagnosis

A 60-year-old male with hypertension, nonsmoker with occasional alcohol use, admitted for suspected cardiac rhythm disturbance... Diagnosed with sick sinus syndrome and underwent permanent pacemaker implantation (DDDR, Isoflex) on Day 4 without complications; device functioning. At discharge on Day 6 he was asymptomatic (no chest pain, dyspnea, palpitations, or urinary complaints) with a normal exam...

Treatment

Cefuroxime 500 mg PO BID for 10 days; Linezolid 600 mg PO BID for 10 days; Trypsin-chymotrypsin 100,000 AU PO TID for 5 days; Paraceta -mol 650 mg PO BID for 5 days as needed; Pantoprazole 40 mg PO daily; Silodosin 8 mg + Solifenacin 5 mg PO daily. Follow-up: cardiology clinic visit on Day 6 post- discharge for pacemaker/device check; primary care follow-up in 4 weeks...



$$P(\text{Treatment} | \text{Diagnosis}) = 1 - \text{surprise_score}$$

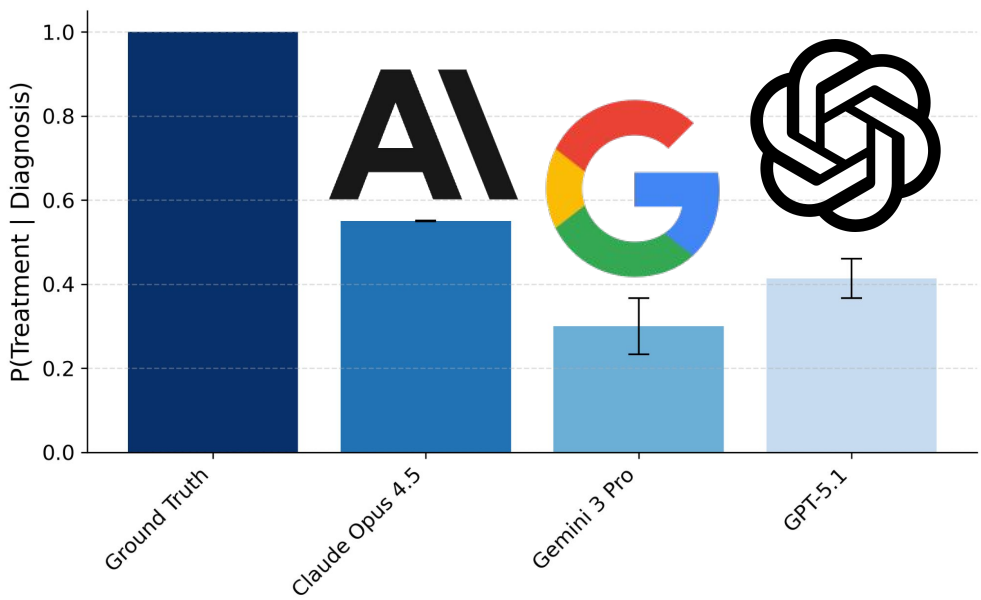
"Given {Diagnosis}, how surprising is {Treatment}?"

Error bars represent Standard Error (SE) of the mean over $N = 10$ API calls to each LLM

Example #2: Aggressive prophylactic measures in high-risk cancer patient

Diagnosis
Adult with history of prostate cancer presented with a 10-day firm, non-tender left breast lump without skin/nipple changes or systemic symptoms; NKDA... On Day 2, wide local excision (lumpectomy) under general anesthesia removed a 4x3 cm mass at the 3 o'clock outer quadrant; no intraoperative complications; preliminary histology noted lymphohistiocytic inflammation; final histopathology pending...

Treatment
Cefuroxime 500 mg PO BID x7 days; rabeprazole 20 mg + domperidone 30 mg PO BID x7 days; acetylsalicylic acid 100 mg + paracetamol 325 mg + serratiopeptidase 15 mg PO BID x7 days; trypsin 96 mg + bromelain 180 mg + rutoside 200 mg PO BID x7 days; protein supplement per diet advice...



$$P(\text{Treatment} \mid \text{Diagnosis}) = 1 - \text{surprise_score}$$

“Given {Diagnosis}, how surprising is {Treatment}?”

Error bars represent Standard Error (SE) of the mean over N = 10 API calls to each LLM

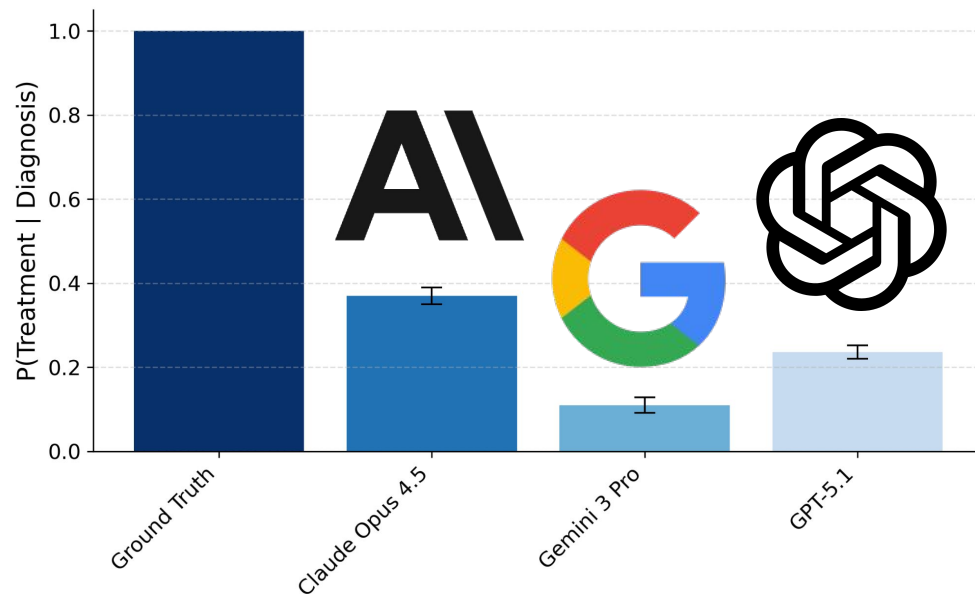
Example #3: Colloquial clinical expressions

Diagnosis

Chief complaint: decreased appetite, nausea, dizziness, and reduced urine output for the past 3–4 days. History of present illness: presented to the emergency department with these complaints. Long-standing dyspnea (NYHA Class III) and a syncopal episode about one year ago. On admission, pain score 7/10. Transferred to the ICU for further evaluation and management of a presumed lower respiratory tract infection.

Treatment

Pantoprazole 40 mg orally every 12 hours (noted as once daily before food) for 10 days; Faropenem 300 mg orally every 12 hours for 5 days; Doxycycline with lactobacillus 100 mg/5 billion CFU orally every 12 hours for 10 days; Rosuvastatin 10 mg plus clopidogrel 75 mg orally every 12 hours until review; Metoprolol ER 25 mg orally every 12 hours until review; Amoxicillin-clavulanate (AB-FLO) 100 mg orally every 12 hours until review; Lactulose syrup 150 mL orally every 12 hours...



$$P(\text{Treatment} | \text{Diagnosis}) = 1 - \text{surprise_score}$$

"Given {Diagnosis}, how surprising is {Treatment}?"

Error bars represent Standard Error (SE) of the mean over $N = 10$ API calls to each LLM

Including Indian clinical guidelines does not alleviate this problem



Ministry of Health & Family Welfare
Government of India

The Clinical Establishments (Registration and Regulation) Act 2010

HOMECE ACT▼NATIONAL COUNCIL▼STANDARDS & GUIDELINES▼FAQRTI▼CAREERSFEEDBACKCONTACT US▼

STANDARD TREATMENT GUIDELINES

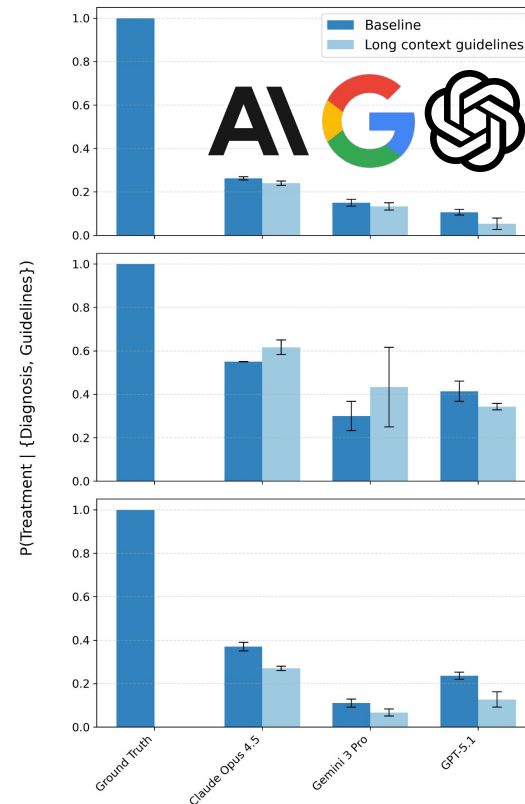
- STGs
- Guidelines for Cardiovascular Diseases
 - Guidelines for Critical Care
 - Guidelines for Gastroenterological Diseases
 - Guidelines for Obstetrics and Gynaecology
 - Guidelines for Haemodialysis
 - Guidelines for Ophthalmology
 - Guidelines for ENT
 - Guidelines for Orthopaedics
 - Guidelines for Medicine (Respiratory)
 - Guidelines for Medicine (Non Respiratory Medical Conditions)
 - Guidelines for Paediatrics and Paediatrics Surgery
 - Guidelines for General Surgery
 - Guidelines for Interventional Radiology
 - Guidelines for Oncology
 - Guidelines for Organ Transplant - Liver
 - Guidelines for Urology
 - Guidelines for Laboratory Medicine
 - Guidelines for G. I. Surgery
 - Guidelines for Neurology
 - Guidelines for Management of Epilepsy in India
 - Guidelines for Endocrinology

22. Standard Treatment Guidelines for Other Medical Conditions

- Guidelines for Major Trauma
 - Full Document
 - Quick Reference Guide
- Guidelines for Hypertension
 - Full Document
 - Quick Reference Guide
- Guidelines for Alcohol Dependence
 - Full Document
 - Quick Reference Guide
- Guidelines for Dry Eye Disease
 - Full Document
 - Quick Reference Guide
- Guidelines for Snake-bite
 - Full Document
 - Quick Reference Guide
- Guidelines for Diabetic Foot
 - Full Document
 - Quick Reference Guide
- Management of Acute Respiratory Infection in Children
 - Quick Reference Guide
- Management of Recurrent Spontaneous Abortion

Model	num_tokens*
Claude Opus 4.5	150 000
Gemini 3 Pro	900 000
GPT-5.1	250 000

*Input tokens (Model and usage tier limited)



Example #1

Example #2

Example #3

$$P(\text{Treatment} \mid \{\text{Diagnosis, Guidelines}\}) = 1 - \text{surprise_score}$$

“Given {Diagnosis} and {Standard Treatment Guidelines}, how surprising is {Treatment}?”

<https://clinicaestablishments.mohfw.gov.in/en/standard-treatment-guidelines>

ETHNIC AI

Synthetic “Treatments” generated by GPT-5.1 are *less surprising* to LLMs

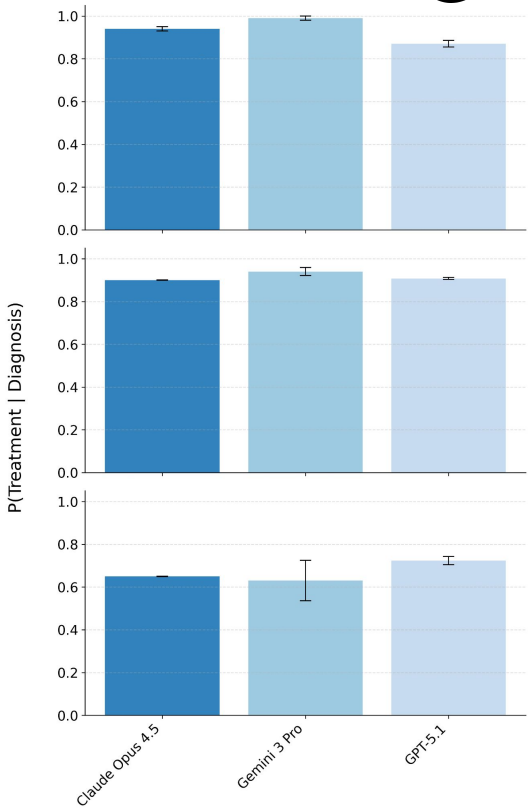


GPT-5.1:

“Given {Diagnosis}, provide Treatment plan:”

All LLMs:

“Given {Diagnosis}, how surprising is {Treatment}?”



Example #1

Example #2

Example #3

Error bars represent Standard Error (SE) of the mean over $N = 5$ API calls to each LLM

>20% of {Diagnosis, Treatment} pairs in our data are surprising to LLMs

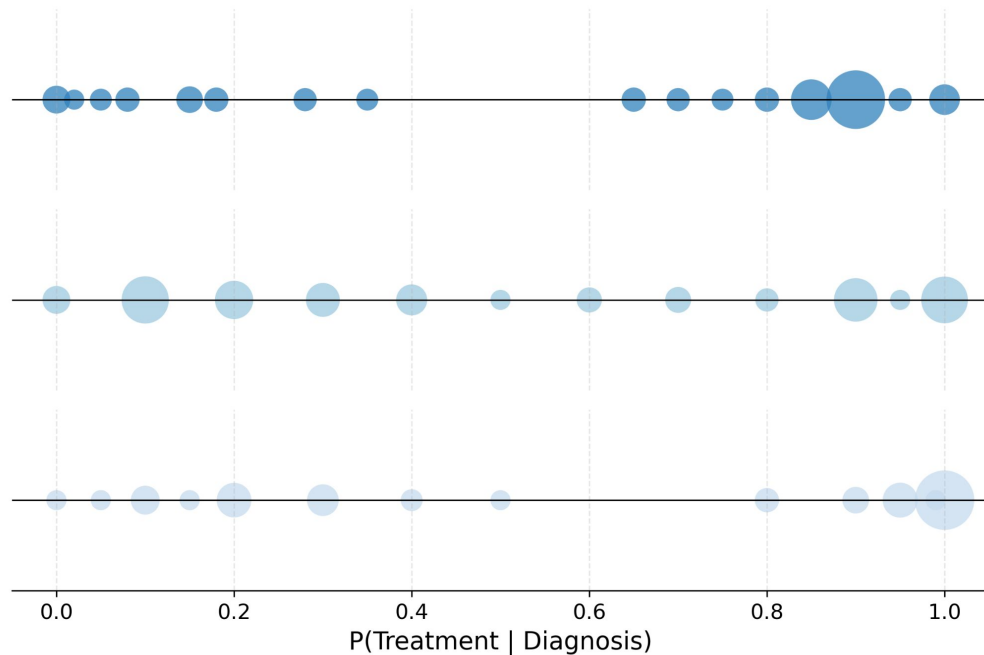
Claude Opus 4.5



Gemini 3 Pro



GPT-5.1



Distribution computed over a randomly selected sample of $N = 420$