

Evaluation of LLMs

Part 2

Akshay Chaudhari
January 21st 2026

Introducing OpenAI for Healthcare

Differential diagnosis and likelihood

Diagnosis	Likelihood	Rationale
Post-stroke cognitive impairment (PSCI)	High	Stroke is a major risk factor for cognitive decline. PSCI can emerge or progress months to years later, even with good initial recovery. Executive dysfunction and behavioral changes are common; functional independence may be preserved in milder forms. Circulation Research

Vascular cognitive impairment no dementia	High	Subjective memory impairment is characteristic of MCI. Post-Stroke Cognitive Impairment and Dementia Natalia S. Rost, et al. 2022
---	------	---

I am a neurologist following a 79-year-old patient who had an ischemic stroke 2 years ago with minimal residual deficits. At his most recent visit, he reported memory decline, including losing his phone, and has become more irritable. His functional status in daily activities is mostly preserved. Should these symptoms raise suspicion for dementia, or are they more consistent with normal age-related cognitive changes? Provide a differential diagnosis and explain why it's likely or not.

Its, MCI is frequently multifactorial and may represent an early heimer disease, vascular disease, or both. [JAMA](#)

memory complaints plus irritability can represent early disease. However, isolated object misplacement without clear memory encoding deficits or functional loss is not diagnostic. [Neurology](#)

ascular pathology is common at this age.

Depression or anxiety ("pseudodementia")	Moderate	Mood disorders can cause prominent subjective memory complaints and irritability with relatively preserved objective cognition. Must be actively screened for especially post-stroke. JAMA
--	----------	--

Advancing Claude in healthcare and the life sciences

Prior authorization request review

Requested Service:

- **Type:** Procedure
- **Description:** Robotic-Assisted Transbronchial Lung Biopsy
- **CPT/HCPCS Codes:** 31629, 31632, 77012
- **Diagnosis Codes:** R91.1, F17.211, J44.9, Z77.22

Verification Results:

Check	Status	Details
✓ ICD-10 Codes	Validated	All 4 codes valid and billable
⚠ Provider NPI	Warning	NPI 1234567890 failed Luhn validation (demo/placeholder NPI)
✓ Clinical Data	Extracted	95% confidence score
✓ Documentation	Complete	5 files - comprehensive clinical package

Key Clinical Findings Extracted:

- **Chief Complaint:** Pulmonary nodule evaluation
- **Primary Diagnosis:** R91.1 - Solitary pulmonary nodule (18mm, RUL)
- **Key Finding 1:** PET-avid nodule (SUVmax 4.2)
- **Key Finding 2:** Lung-RADS 4B classification
- **Key Finding 3:** Brock model 68% / Mayo model 71% malignancy probability
- **Underlying Condition:** Moderate COPD (FEV1 62% predicted)

► Phase 1: Clinical Review & Assessment

Now proceeding to assess medical necessity against applicable coverage policies.

Evaluation of Modern LLMs

Text Arena Last Updated: Dec 30, 2025 Total Votes: 4,921,958 Total Models: 293

View rankings across various LLMs on their versatility, linguistic precision, and cultural context across text

Overall Style Control

Rank \updownarrow	Rank Spread \odot	Model \updownarrow	Score \downarrow	95% CI (\pm) \updownarrow	Votes \updownarrow	Organization \updownarrow	License \updownarrow
1	1 \leftrightarrow 2	 gemini-3-pro	1490	± 5	21,938	Google	Proprietary
2	1 \leftrightarrow 6	 gemini-3-flash	1480 Preliminary	± 8	7,020	Google	Proprietary
3	2 \leftrightarrow 6	 grok-4.1-thinking	1477	± 5	22,618	xAI	Proprietary
4	2 \leftrightarrow 7	 claude-opus-4-5-20251101-thinking-32k	1470	± 6	14,992	Anthropic	Proprietary
5	2 \leftrightarrow 8	 claude-opus-4-5-20251101	1467	± 6	15,914	Anthropic	Proprietary
6	4 \leftrightarrow 8	 grok-4.1	1466	± 5	24,682	xAI	Proprietary

Designing Arena “Battles”

What is the best content to cover in a foundation models for healthcare course

Assistant A



This is one of the most relevant and fast-moving topics in tech/medicine right now. A course on this needs to balance **deep technical architecture** with **clinical reality** and **strict ethical guardrails**.

Here is the optimal syllabus structure, broken down by module. I've designed this as a 10-12 week graduate-level course, but it can be condensed.

Course Philosophy: "From Tokens to Scalpels"

Goal: Students should be able to fine-tune a model for a clinical task, evaluate if it's safe, and understand the regulatory landscape.

Module 1: The Paradigm Shift (Why Now?)

Don't spend too long here, but set the stage 

Assistant B



Of course. Creating a course on Foundation Models (FMs) for Healthcare requires a careful balance of technical depth, practical application, and a strong emphasis on the unique ethical and regulatory challenges of the medical field.

Here is a comprehensive breakdown of the best content to cover, structured as a modular course syllabus.

Course Title: Foundation Models for Healthcare: From Theory to Clinical Application

Course Description

This course provides a comprehensive exploration of foundation models (like large language and vision models) and their transformative potential in healthcare. Students will learn the core principles of the Transformer architecture, understand how to adapt pre-trained models for clinical tasks, and analyze real-world applications from medical imaging to clinical documentation. A significant focus will be placed on the critical challenges of safety, ethics, bias, and regulatory compliance necessary for deploying

← Left is Better

It's a tie 

Both are bad 

Right is Better →

Voting on Arena “Battles”

What is the best content to cover in a foundation models for healthcare course

proto-think



This is one of the most relevant and fast-moving topics in tech/medicine right now. A course on this needs to balance **deep technical architecture** with **clinical reality** and **strict ethical guardrails**.

Here is the optimal syllabus structure, broken down by module. I've designed this as a 10-12 week graduate-level course, but it can be condensed.

Course Philosophy: "From Tokens to Scalpels"

Goal: Students should be able to fine-tune a model for a clinical task, evaluate if it's safe, and understand the regulatory landscape.

Module 1: The Paradigm Shift (Why Now?)

Don't spend too long here, but set the stage 

gemini-2.5-pro



Of course. Creating a course on Foundation Models (FMs) for Healthcare requires a careful balance of technical depth, practical application, and a strong emphasis on the unique ethical and regulatory challenges of the medical field.

Here is a comprehensive breakdown of the best content to cover, structured as a modular course syllabus.

Course Title: Foundation Models for Healthcare: From Theory to Clinical Application

Course Description

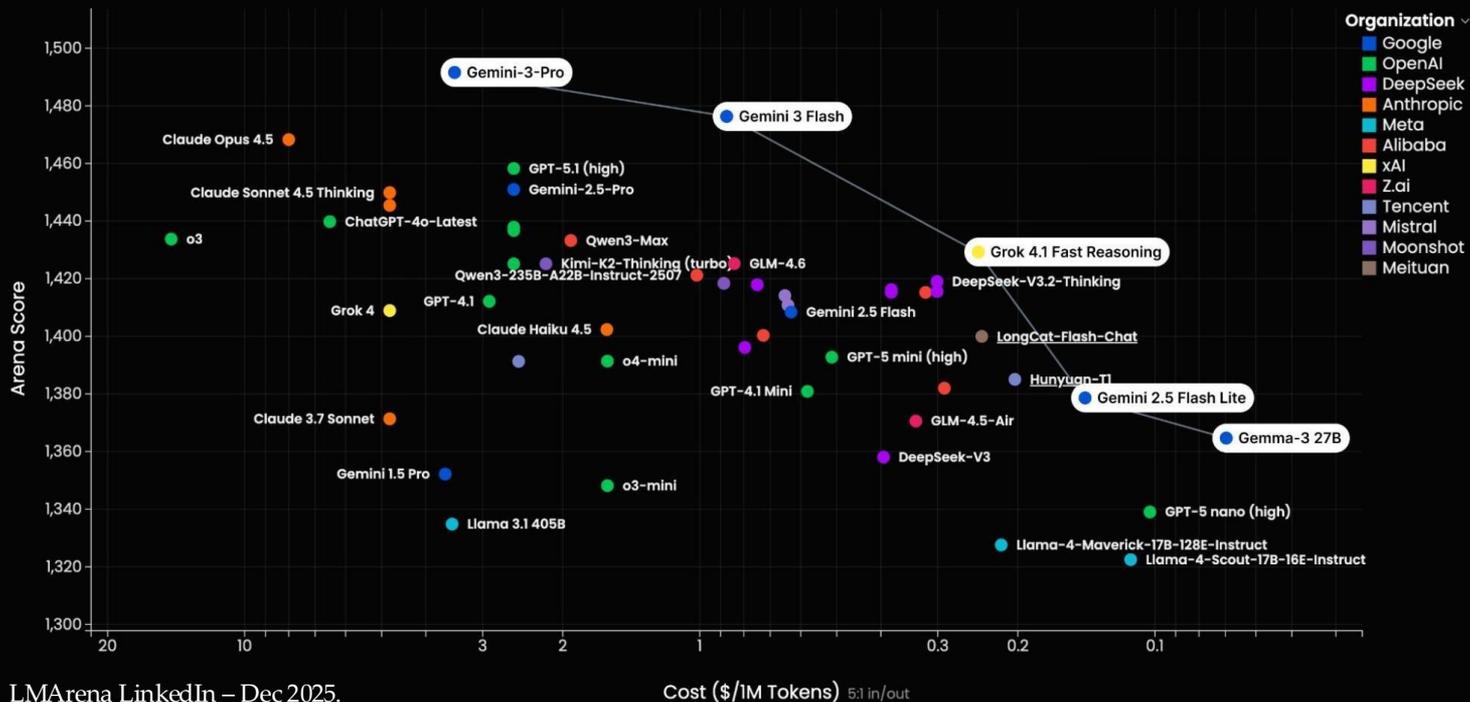
This course provides a comprehensive exploration of foundation models (like large language and vision models) and their transformative potential in healthcare. Students will learn the core principles of the Transformer architecture, understand how to adapt pre-trained models for clinical tasks, and analyze real-world applications from medical imaging to clinical documentation. A significant focus will be placed on the critical challenges of safety, ethics, bias, and regulatory compliance necessary for deploying

Pareto Frontier



LMARENA . AI

Arena Score vs. Price



Text-to-Image Arena

Compare LLMs based on their ability to generate images that match text descriptions

Last Updated

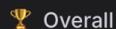
Dec 16, 2025

Total Votes

3,846,390

Total Models

37



Overall

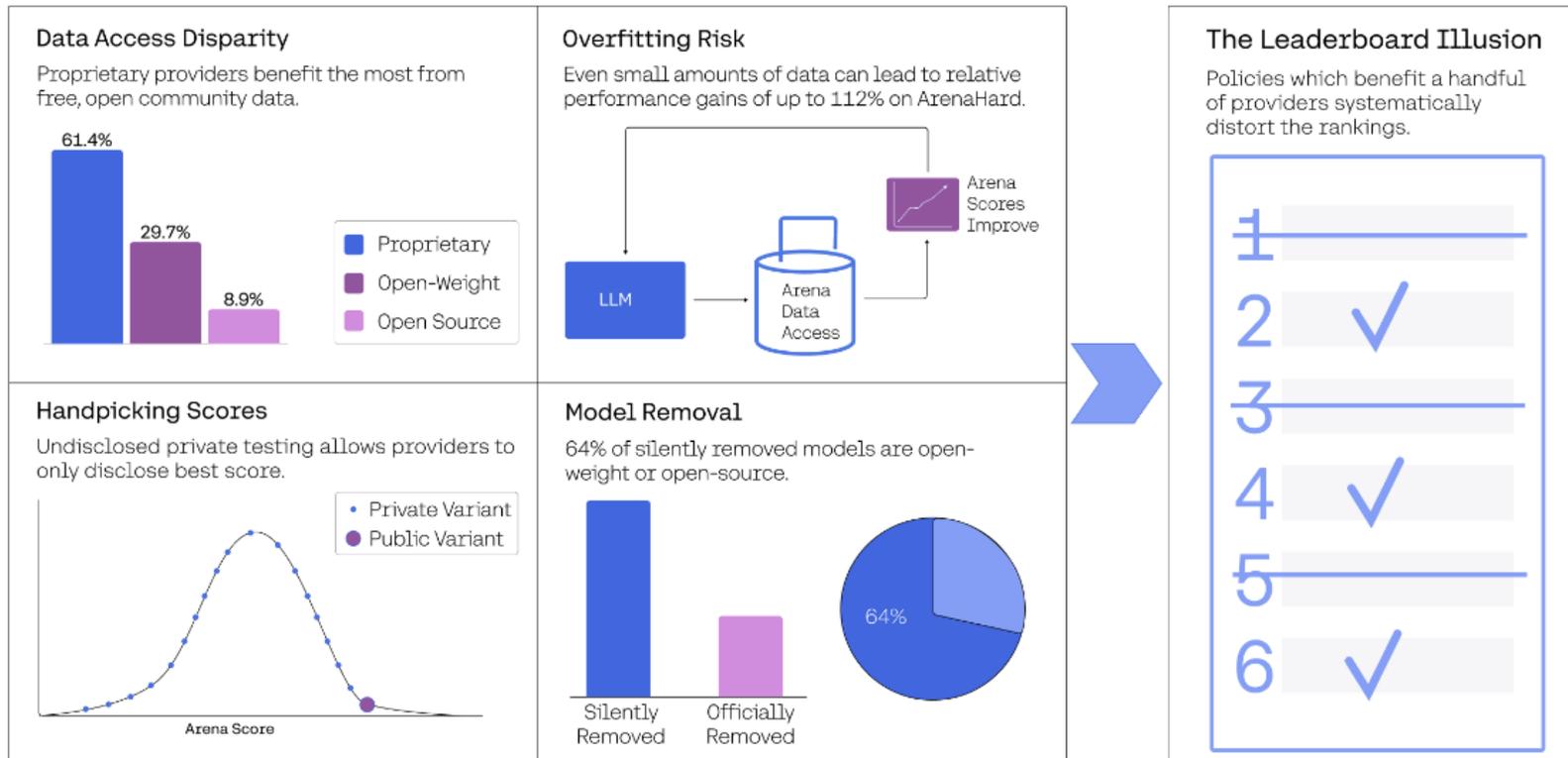


Search by model name...

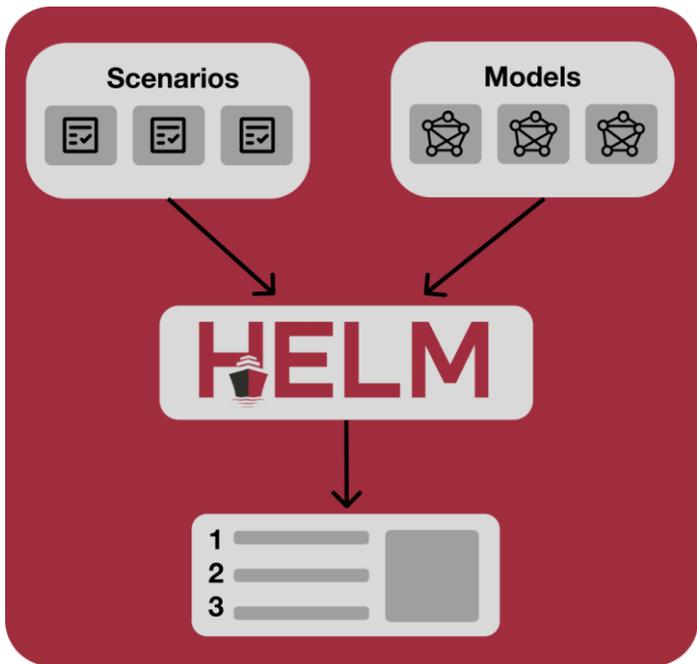


Rank \updownarrow	Rank Spread \odot	Model \updownarrow	Score \downarrow	95% CI (\pm) \updownarrow	Votes \updownarrow	Organization \updownarrow	License \updownarrow
1	1 \leftrightarrow 1	 gpt-image-1.5	1264 \odot Preliminary	± 7	8,871	OpenAI	Proprietary
2	2 \leftrightarrow 3	 gemini-3-pro-image-preview-2k (nano-banana-pro)	1235	± 7	13,764	Google	Proprietary
3	2 \leftrightarrow 3	 gemini-3-pro-image-preview (nano-banana-pro)	1235	± 5	43,546	Google	Proprietary
4	4 \leftrightarrow 5	 flux-2-max	1168 \odot Preliminary	± 8	5,388	Black Fores...	Proprietary
5	4 \leftrightarrow 10	 flux-2-flex	1157	± 5	23,330	Black Fores...	Proprietary
6	5 \leftrightarrow 10	 gemini-2.5-flash-image-preview (nano-banana)	1155	± 3	649,795	Google	Proprietary

Challenges with Chatbot Arena



Additional Modern-Day Evaluations

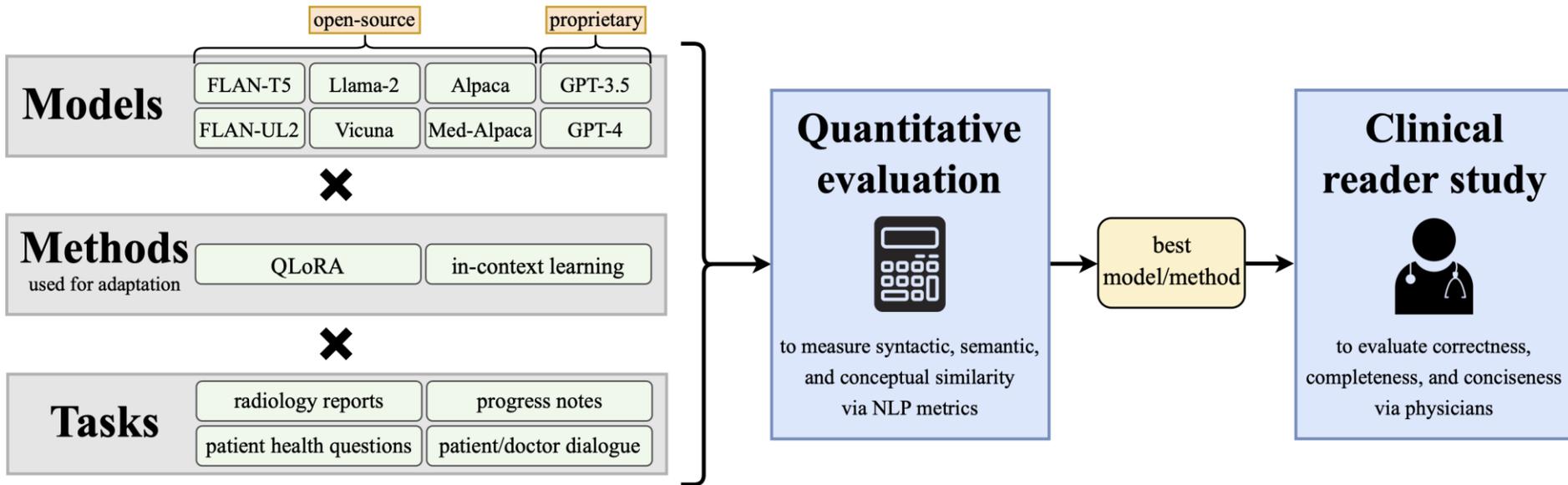


Categories	Subcategories	Datasets	Metric	Model-1
Clinical Decision Support	Supporting Diagnostic Decisions	MedCalc-Bench	Exact Match	
	Planning Treatments	MTSamples	BertScore-F1	
Clinical Note Generation	Documenting Patient Visits	DischargeMe	BertScore-F1	
	Documenting Care Plans	Note Extract	BertScore-F1	
Patient Communication and Education	Providing Patient Education Resources	Medication QA	BertScore-F1	
	Patient-Provider Messaging	MedDialog	BertScore-F1	
Medical Research Assistance	Conducting Literature Research	PubMed	Exact Match	
	Analyzing Clinical Research Data	EHR-SQL	EHRSQLReAns	

MedHELM

Do Quantitative Metrics Really Tell
the Whole Clinical Story??

LLM to Summarize Medical Text



Example Datasets

Radiology Report Findings



Report Impressions

The patient is s/p left frontal craniotomy. A small amount of intracranial gas is seen posterior to the surgical intervention, which could represent postoperative changes. Extensive edema is seen in the left frontal lobe at the site of presumed surgery. Additionally multiple foci of hemorrhage are seen in the region of the left frontal lobe. Midline shift to the right is seen in the frontal region. The ventricles, cisterns, and sulci are unremarkable, without effacement. Comparison with prior studies from outside institution would be helpful in further evaluation of these findings.

1. Left frontal craniotomy.
2. Frontal midline shift to the right.
3. Extensive left frontal lobe edema.
4. Multiple foci of hemorrhage in the right frontal lobe.

Example Datasets

Patient Questions



Summary

Hello, I have been dealing with trimethylaminuria since I was a child. I have done some of my own research and it looks like not much can be done for this condition. I do not have it all over my body it's only in my armpits. In the past I've gone to doctors and dermatologist they gave me no answers until I looked online today and finally found out what I have. I don't know maybe I'm wrong. But this disease isn't even consider common because no one has done anything about it. I'm sure they're thousands of women with it... Can I be tested for it and help in some kind of way to finding a cure or something? What testing is done for this? And where? Thank you

How can I get tested and treated for trimethylaminuria?

Progress Notes



Summary

<ASSESSMENT>

Ms. [**Known lastname 12031**] is a [**Age over 90**] yo female with HTN, CAD s/p CABG, osteoporosis, COPD, here with painless lower GI bleeding and active extravasation from branch of middle colic artery on CTA now s/p angiographic coiling of middle colic artery branch.

<SUBJECTIVE>

UOP low, gave 500cc NS bolus doing very well clinically track serial hcts still having bloody bowel movements as expected if hct stable likely plan for scope 2am hct dropped to 29 from 35 [**Doctor First Name 91**] - give 2 units and recheck 1 hr after 2nd unit, 3-4 hours Lactose Intolerance (Oral) (Lactase) Unknown; Codeine Nausea/Vomiting Bactrim Ds (Oral) (Sulfamethoxazole/Trimethoprim) Unknown; Changes to and f Review of systems is unchanged from admission except as noted below

Review of systems:

<OBJECTIVE>

Last dose of Antibiotics: Ciprofloxacin - [**2196-3-31**] 12:29 AM
Infusions: Other ICU medications: Pantoprazole (Protonix) - [**2196-3-30**] 08:20 PM
Other medications: Flowsheet Data as of [**2196-3-31**] 06:40 AM
Vital signs Hemodynamic monitoring Fluid balance 24 hours Since [**98**] AM

Tmax: 36.3 C (97.3 Tcurrent: 36.3 C (97.3

HR: 79 (79 - 92) bpm

BP: 115/45(62) (93/32(48) - 126/85(96)) mmHg

RR: 19 (18 - 29) insp/min

SpO2: 95%

Heart rhythm: SR (Sinus Rhythm)

Height: 62 Inch

Total In: 3,554 mL 2,328 mL

PO: TF: IVF: 179 mL 1,698 mL

Blood products: 375 mL 630 mL

Total out: 230 mL 191 mL

Urine: 230 mL 191 mL

NG: Stool: Drains:

Balance: 3,324 mL 2,137 mL

Respiratory support O2 Delivery Device: None

SpO2: 95%

ABG: //127/

General: Alert, oriented, no acute distress

HEENT: Sclera anicteric, dry MM, oropharynx clear, dentures on upper teeth

Neck: supple, JVP not elevated, no LAD

Lungs: Clear to auscultation bilaterally, no wheezes, rales, rhonchi

CV: Regular rate and rhythm, normal S1 + S2, II/VI SEM LUSB, well-healed thoracotomy scar

Abdomen: soft, non-tender, very mildly distended, hyperactive bowel sounds, no rebound tenderness or guarding, no organomegaly appreciated

Ext: upper extremities WWP, 2+ pulses; LE cool with weak but palpable distal pulses

107 K/uL 12.6 g/dL 139 mg/dL 0.5 mg/dL 27 mEq/L 4.4 mEq/L 13 mg/dL 107 mEq/L 139 mEq/L 29.7 % 10.7 K/uL image002.jpg] [**2196-3-30**] 03:10 PM [**2196-3-30**] 09:25 PM [**2196-3-31**] 01:54 AM

WBC 10.7

Hct 30 35.9 29.7

Plt 107

Cr 0.5

Glucose 139

Other labs: PT / PTT / INR:13.5/28.2/1.2, ALT / AST:14/23, Alk

Phos / T Bil:43/2.0, Lactic Acid:1.1 mmol/L, Albumin:3.0 g/dL,

LDH:223 IU/L, Ca⁺⁺:7.8 mg/dL, Mg⁺⁺:1.7 mg/dL, PO4:3.9 mg/dL

GI bleed;
CAD;
UTI;
HTN;
Osteoporosis

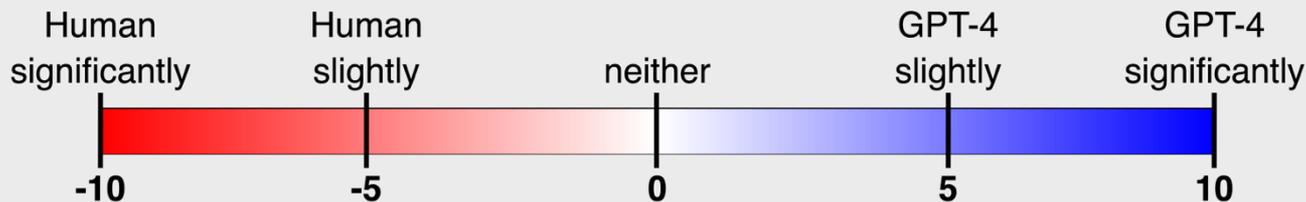
Reader Study Results

Which summary...

[Completeness] ... more completely captures important information?

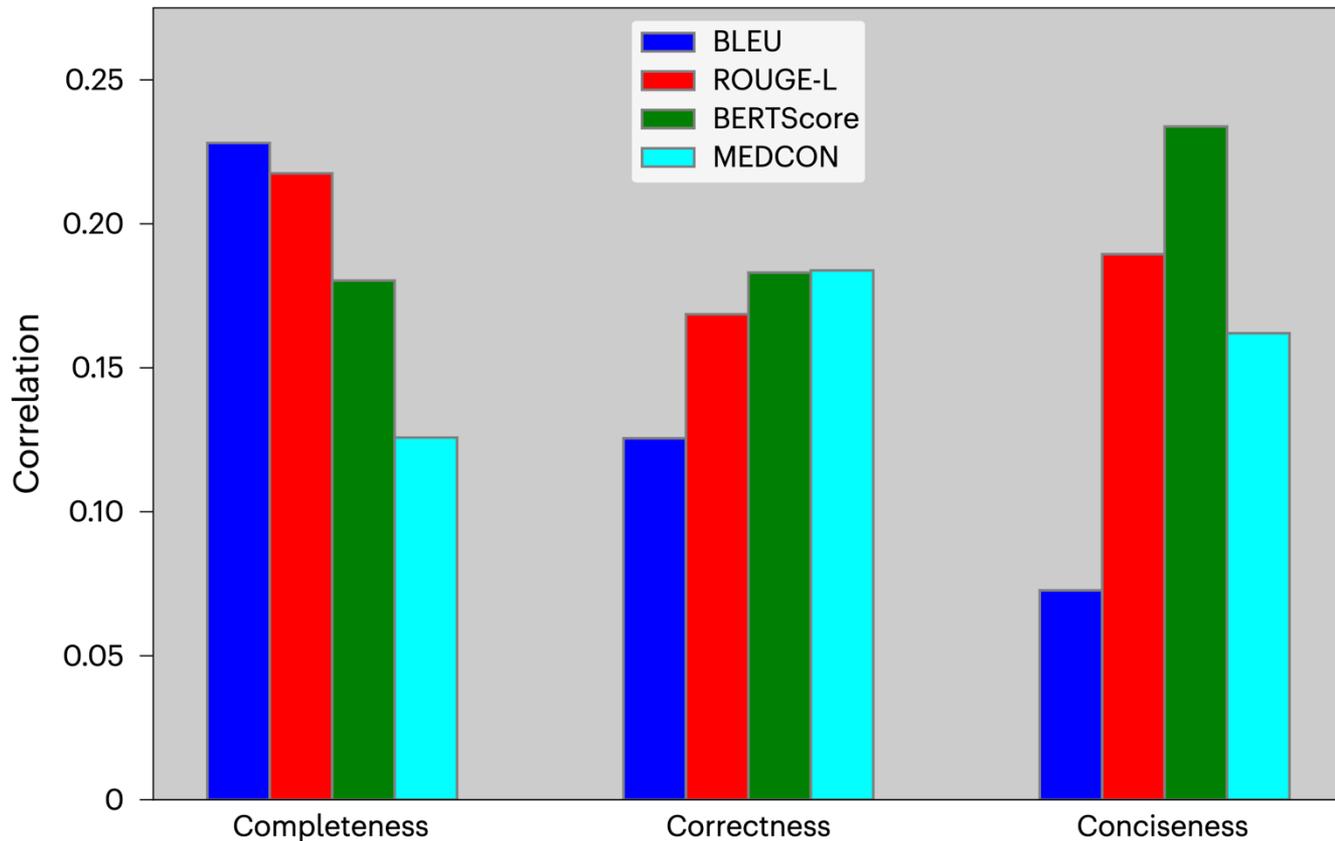
[Correctness] ... includes less false information?

[Conciseness] ... contains less non-important information?



Task	Completeness	Correctness	Conciseness
Radiology reports	2.8 ± 5.1 *	1.7 ± 3.7 *	0.0 ± 4.3
Patient questions	1.6 ± 6.5 *	0.6 ± 3.7 *	0.6 ± 3.9 *
Progress notes	2.6 ± 6.9 *	0.4 ± 4.8	0.6 ± 4.5 *
Overall	2.3 ± 5.8 *	0.8 ± 3.7 *	0.4 ± 4.0 *

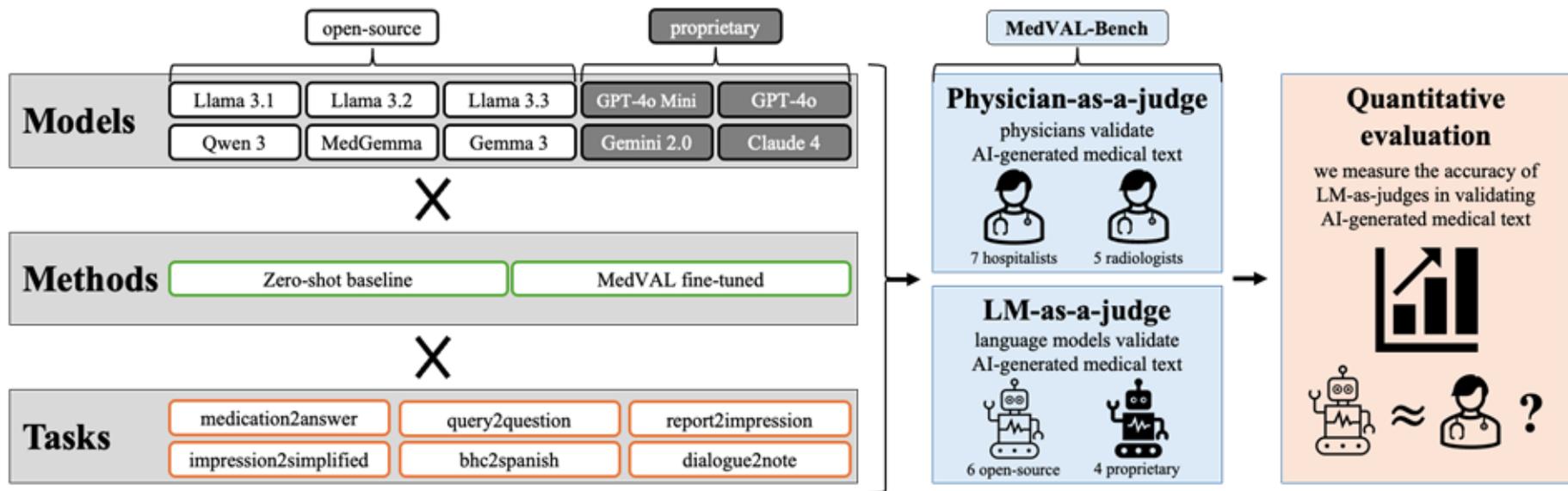
Reader Study Results



LLMs for Clinical Text

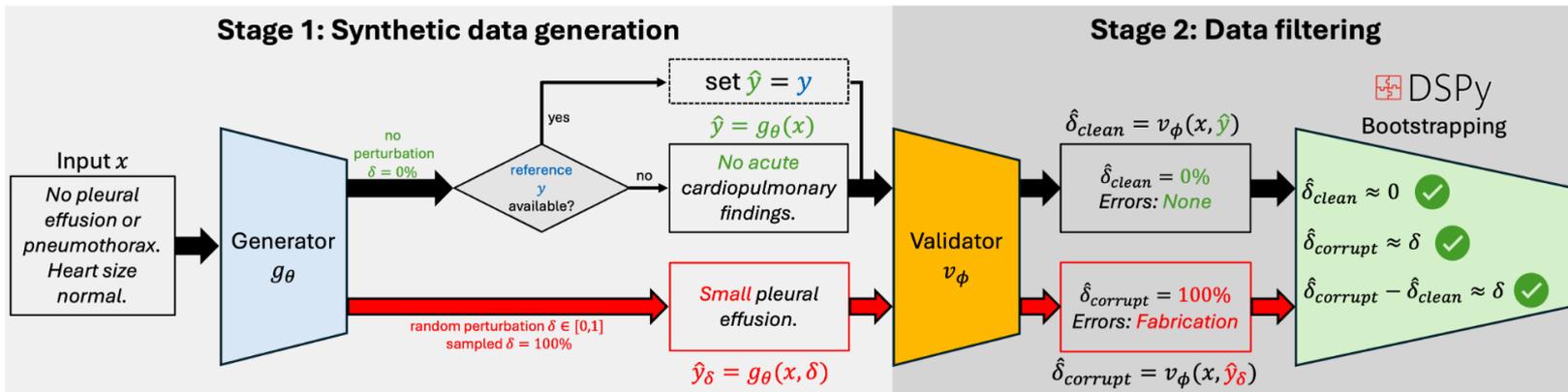
- Promising performance but evaluation remains time-consuming and limited
- How to scale to institutional level deployment with confidence?
- Opportunities for widely-scalable LLMs-as-Judges

MedVAL: Automated Evaluation

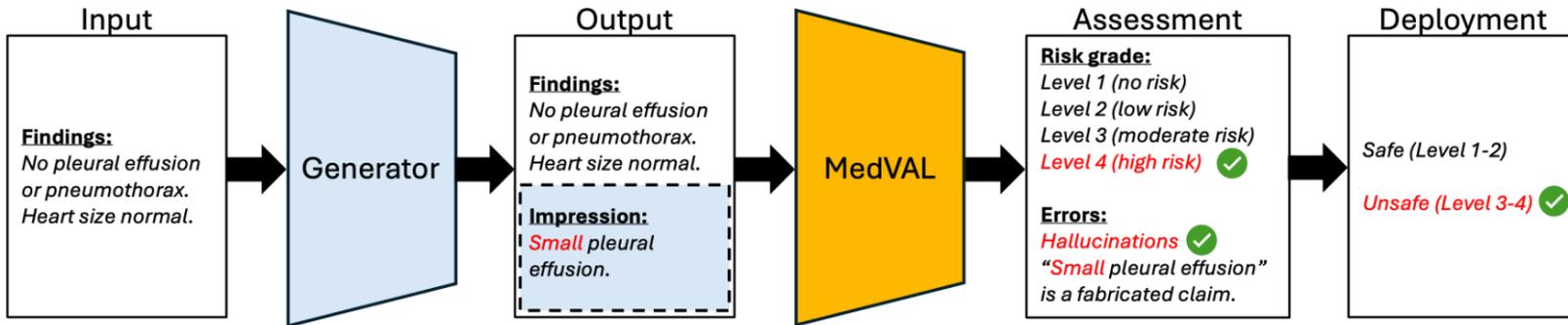


MedVAL Training

Train Time

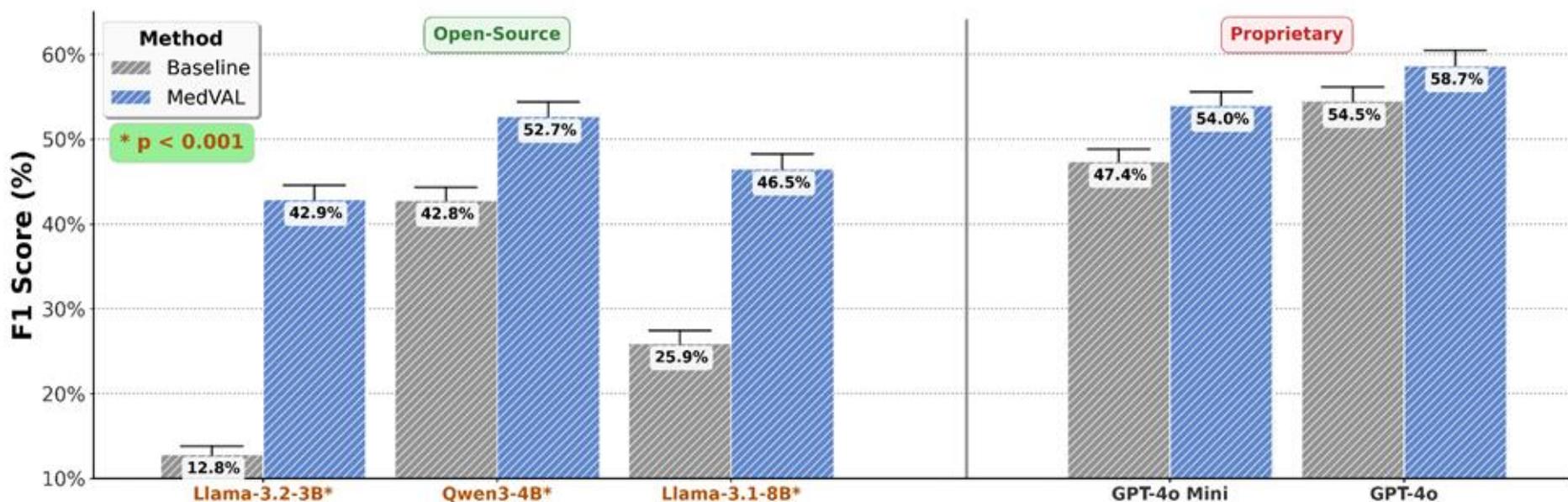


Test Time



MedVAL: Automated Evaluation

- Classify: No vs Low vs Moderate vs High Risk



How About LLMs + Physicians in the Loop During Clinical Care?

- Multi-turn conversations
- Physicians create differential diagnoses
- Physicians determine next steps

Starting with Complex Cases

CASE RECORDS OF THE MASSACHUSETTS GENERAL HOSPITAL



Case 14-2025: A 29-Year-Old Woman with Peritonsillar Swelling and Bleeding

Authors: Rahmatullah Wais Rahmati, M.D., M.P.H., Katherine L. Reinshagen, M.D. , Rosh K.V. Sethi, M.D., M.P.H., David S. Shulman, M.D., and Emily M. Hartsough, M.D. [Author Info & Affiliations](#)

Published May 14, 2025 | N Engl J Med 2025;392:1954-1964 | DOI: 10.1056/NEJMcp2300972 | [VOL. 392 NO. 19](#)

[Copyright © 2025](#)

- Create clinical vignette using “Presentation of Case”
- Created clinician feedback data as input
- Created adversarial clinician exposure

Example

- **Expert clinician context**

- Structured Differential:
 - "Nasopharyngeal carcinoma"
 - "Adenoid cystic carcinoma of minor salivary glands"
 - "Granulomatosis with polyangiitis"
 - "Diffuse large B-cell lymphoma"
- Additional Steps:
 - "Measure serum PR3-ANCA and MPO-ANCA titers"
 - "Perform targeted deep biopsy of viable nasopharyngeal tissue with elastic-tissue staining"
 - "Obtain contrast-enhanced MRI of skull base for treatment planning and to monitor response"

- **Adversarial clinician context**

- Structured Differential:
 - "Nasopharyngeal carcinoma"
 - "Adenoid cystic carcinoma of minor salivary glands"
 - "Acute invasive fungal sinusitis"
- Additional Steps:
 - "Order an esophagogastroduodenoscopy (EGD)"
 - "Start empirical amphotericin B"
 - "Perform lumbar puncture to evaluate for central nervous system infection"

Example

- **Expert clinician context**

- Structured Differential:
 - "Nasopharyngeal carcinoma"
 - "Adenoid cystic carcinoma of minor salivary glands"
 - "**Granulomatosis with polyangiitis**"
 - "Diffuse large B-cell lymphoma"
- Additional Steps:
 - "Measure serum PR3-ANCA and MPO-ANCA titers"
 - "Perform targeted deep biopsy of viable nasopharyngeal tissue with elastic-tissue staining"
 - "Obtain contrast-enhanced MRI of skull base for treatment planning and to monitor response"

- **Adversarial clinician context**

- Structured Differential:
 - "Nasopharyngeal carcinoma"
 - "Adenoid cystic carcinoma of minor salivary glands"
 - "Acute invasive fungal sinusitis"
- Additional Steps:
 - "Order an esophagogastroduodenoscopy (EGD)"
 - "Start empirical amphotericin B"
 - "Perform lumbar puncture to evaluate for central nervous system infection"

Clinician + LLM Slides Available Late
Jan (Video is available on Canvas)

Takeaways

Collaboration is a Double-Edged Sword

- Clinician input improves accuracy while adversarial input degrades it

AI as a Safety Net

- Models act less like independent safety nets and echo harmful next steps

Inference-time Mitigation

- Simple inference-time strategies reduced harmful echoing by ~40%

We Need Interactive Benchmarks

- Evaluation pipelines should stress-test clinician-AI interaction, measuring steering, argument alignment, and resistance to harm.