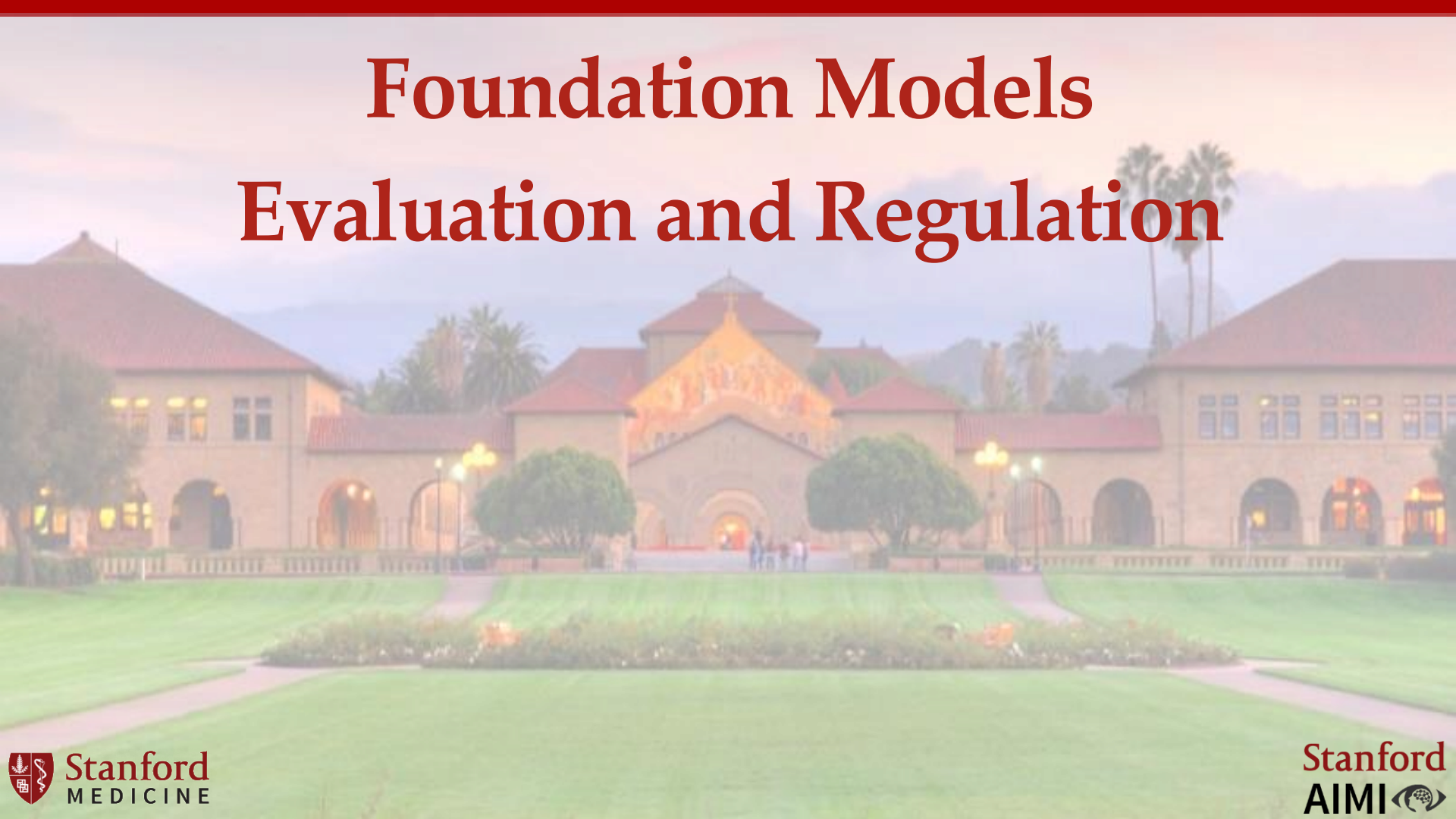
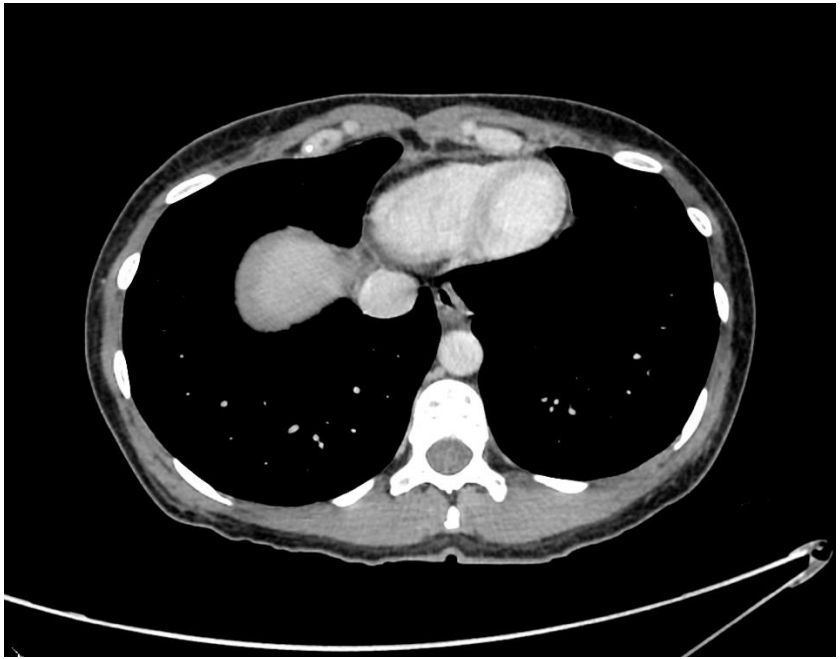
Foundation Models Evaluation and Regulation



Stanford MEDICINE

Stanford AIMI

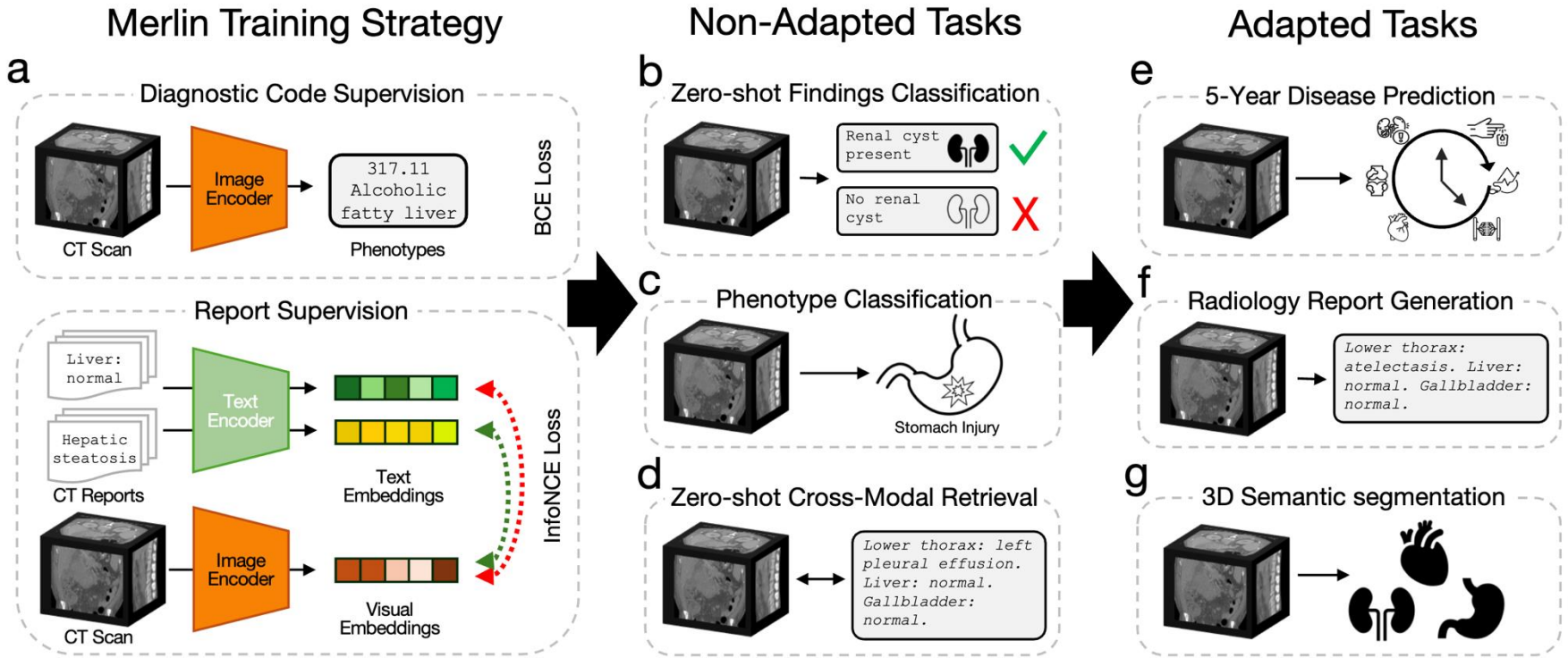
Merlin Abdominal CT FM



- Trained on 15.5k CT scans and corresponding radiology reports (6M tokens)
- Pre-trained using ICD diagnosis codes
- Evaluated on 5k internal and 5k external studies

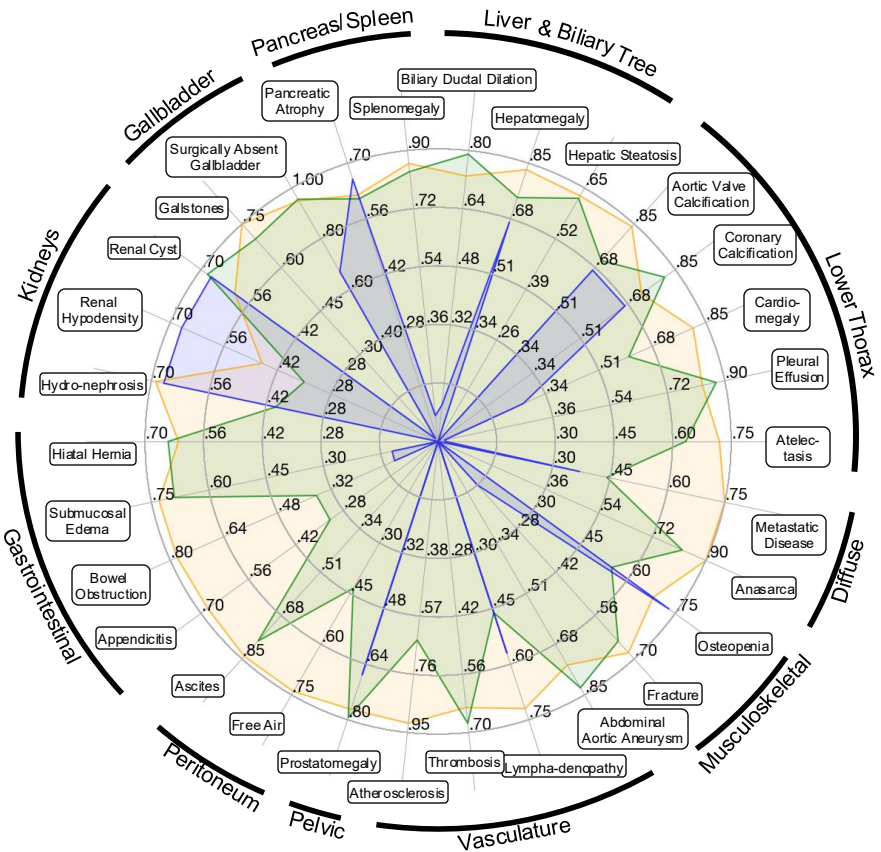
Blankemeier et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Merlin: 3D Abdominal CT FM



Blankemeier et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Merlin Capabilities: Zero Shot Classification



Prompt with natural language

F1 Scores

(not AUROC!)

Merlin (External)

Merlin (Internal)

BioMedCLIP (Internal)

Blankemeier et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

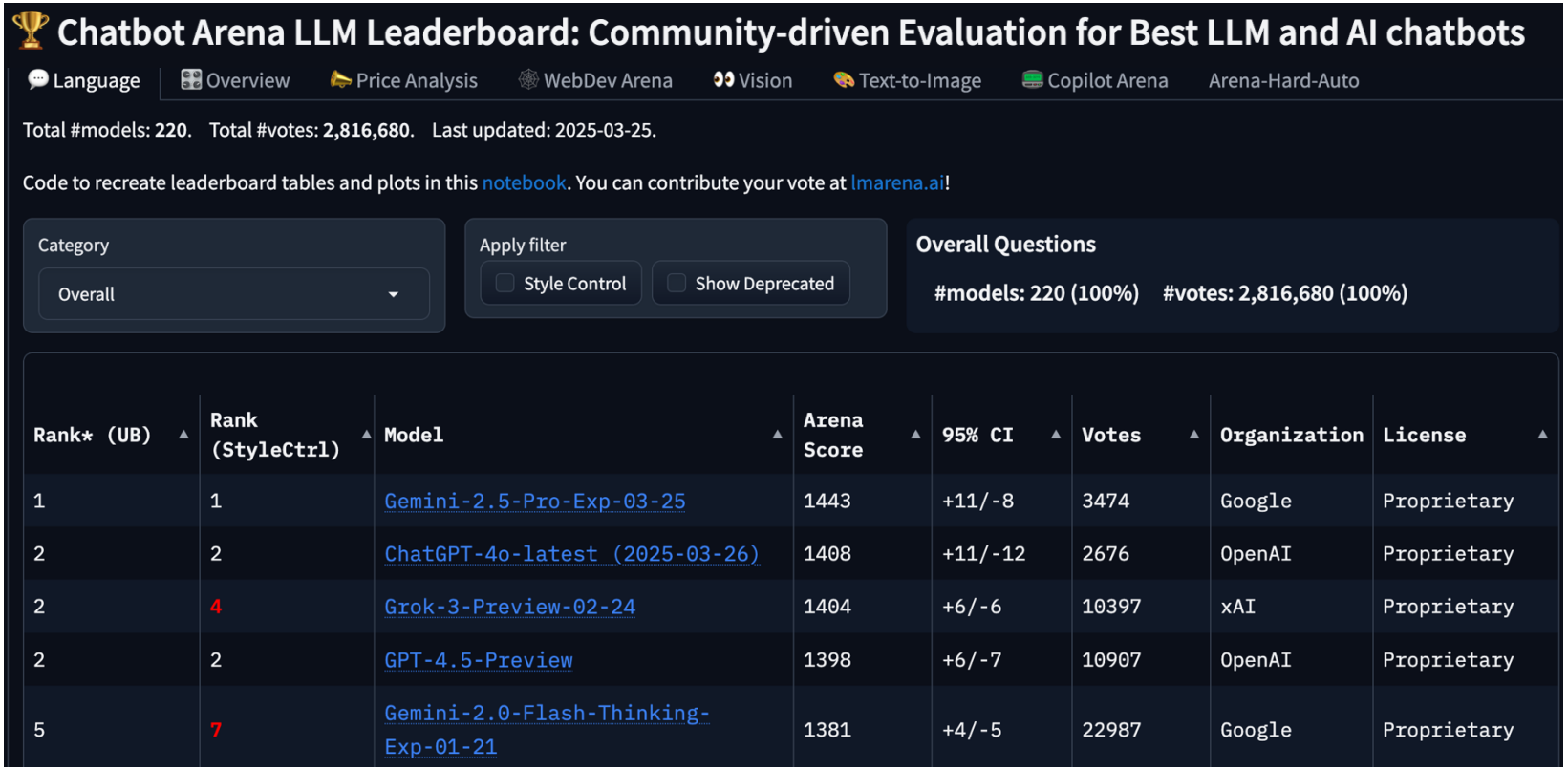
Merlin Evaluation Criteria

- **Zero-Shot Classification**: F1, AUROC, etc
- **Phenotype Prediction**: AUROC, AUPRC, etc
- **Retrieval**: Recall @k
- **Disease Prediction**: AUROC, AUPRC, etc
- **Report Generation**: ROUGE, GREEN, LLM-as-a-judge. etc
- **Segmentation**: Dice, ASSD, etc

Foundation Models

- Do we evaluate/regulate the model or the tasks?

Contemporary LLM Evaluation



Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots

Language | Overview | Price Analysis | WebDev Arena | Vision | Text-to-Image | Copilot Arena | Arena-Hard-Auto

Total #models: **220**. Total #votes: **2,816,680**. Last updated: 2025-03-25.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at lmarena.ai!

Category: Overall

Apply filter: Style Control | Show Deprecated

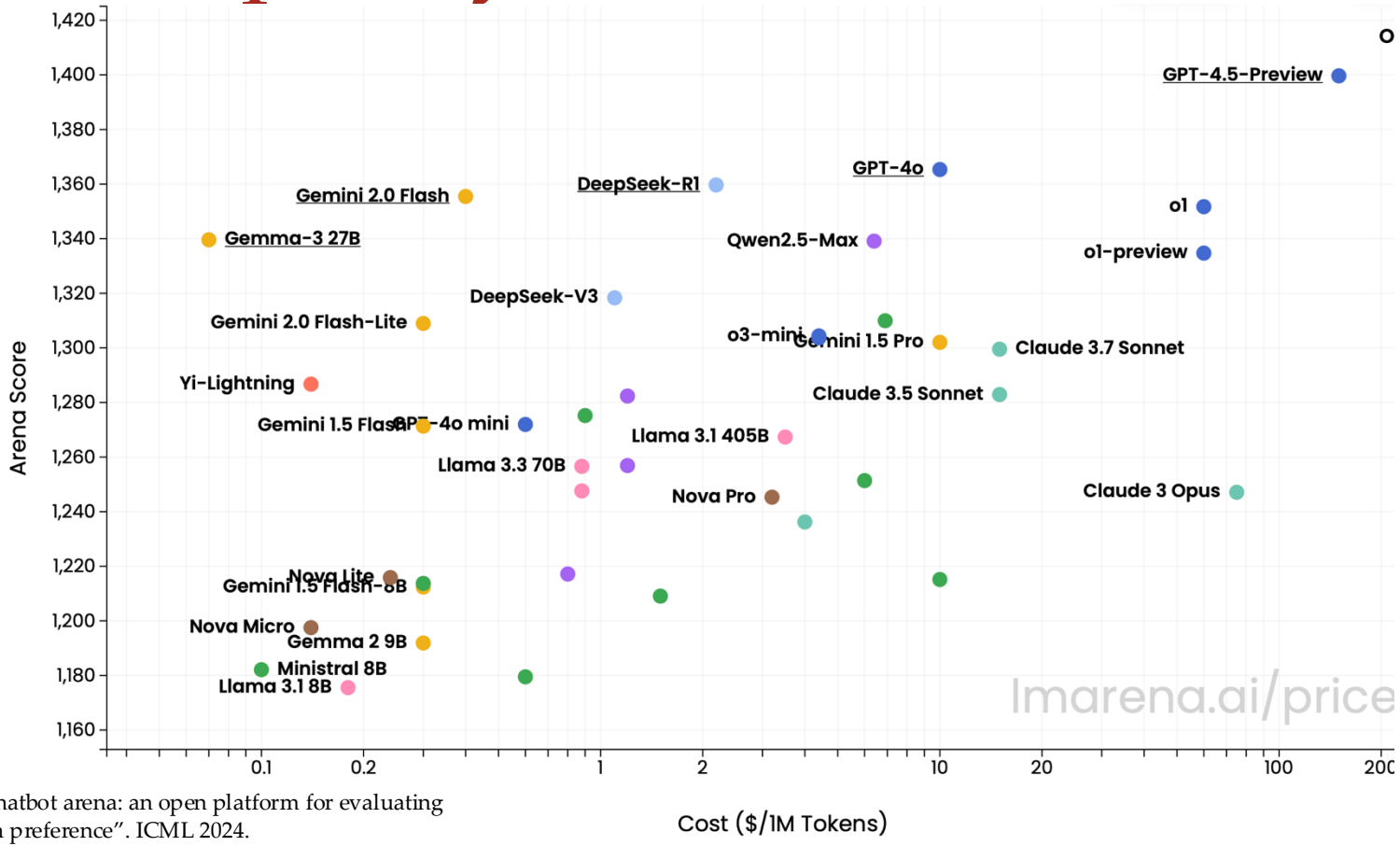
Overall Questions

#models: 220 (100%) #votes: 2,816,680 (100%)

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1443	+11/-8	3474	Google	Proprietary
2	2	ChatGPT-4o-latest (2025-03-26)	1408	+11/-12	2676	OpenAI	Proprietary
2	4	Grok-3-Preview-02-24	1404	+6/-6	10397	xAI	Proprietary
2	2	GPT-4.5-Preview	1398	+6/-7	10907	OpenAI	Proprietary
5	7	Gemini-2.0-Flash-Thinking-Exp-01-21	1381	+4/-5	22987	Google	Proprietary

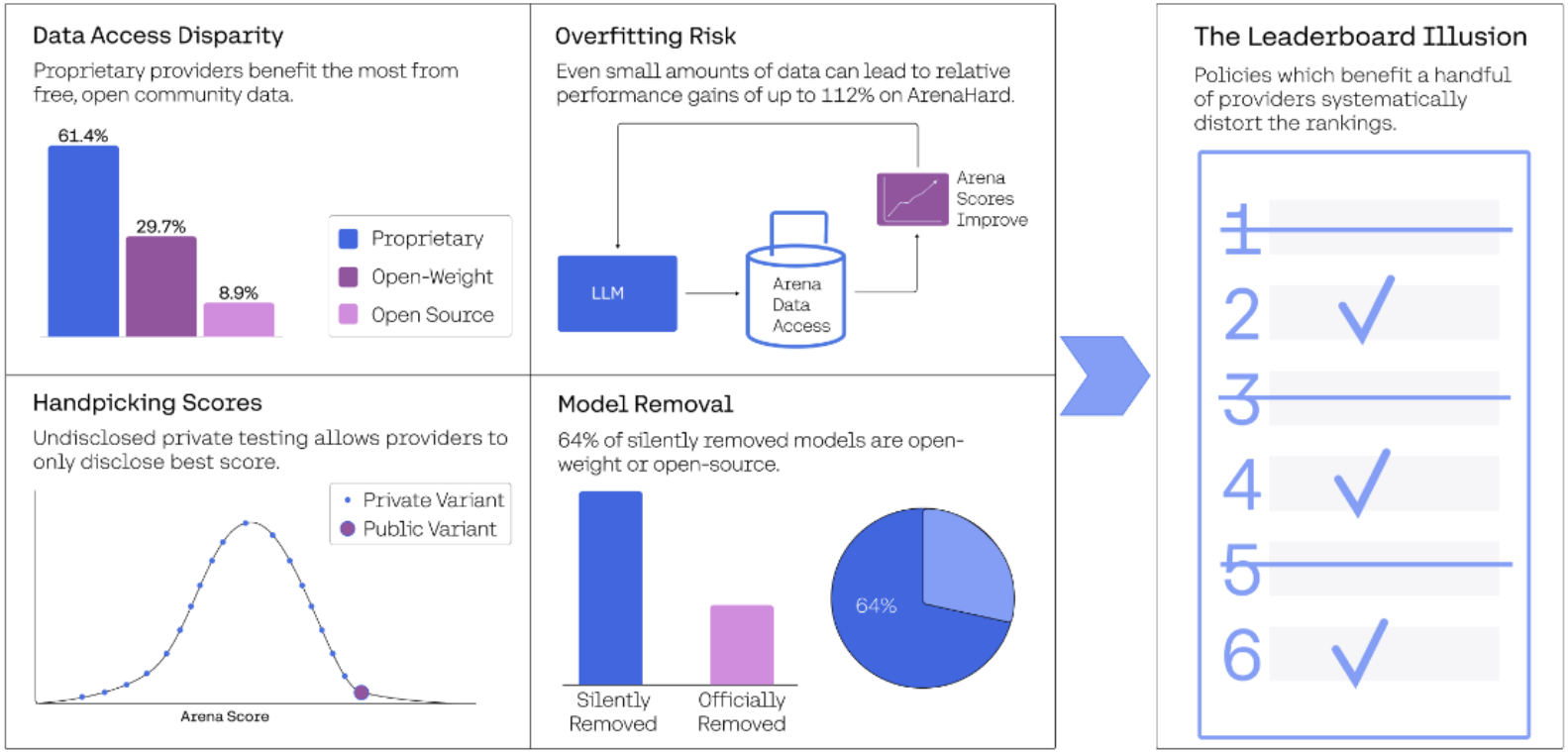
Chiang et al. "Chatbot arena: an open platform for evaluating LLMs by human preference". ICML 2024.

Contemporary LLM Evaluation



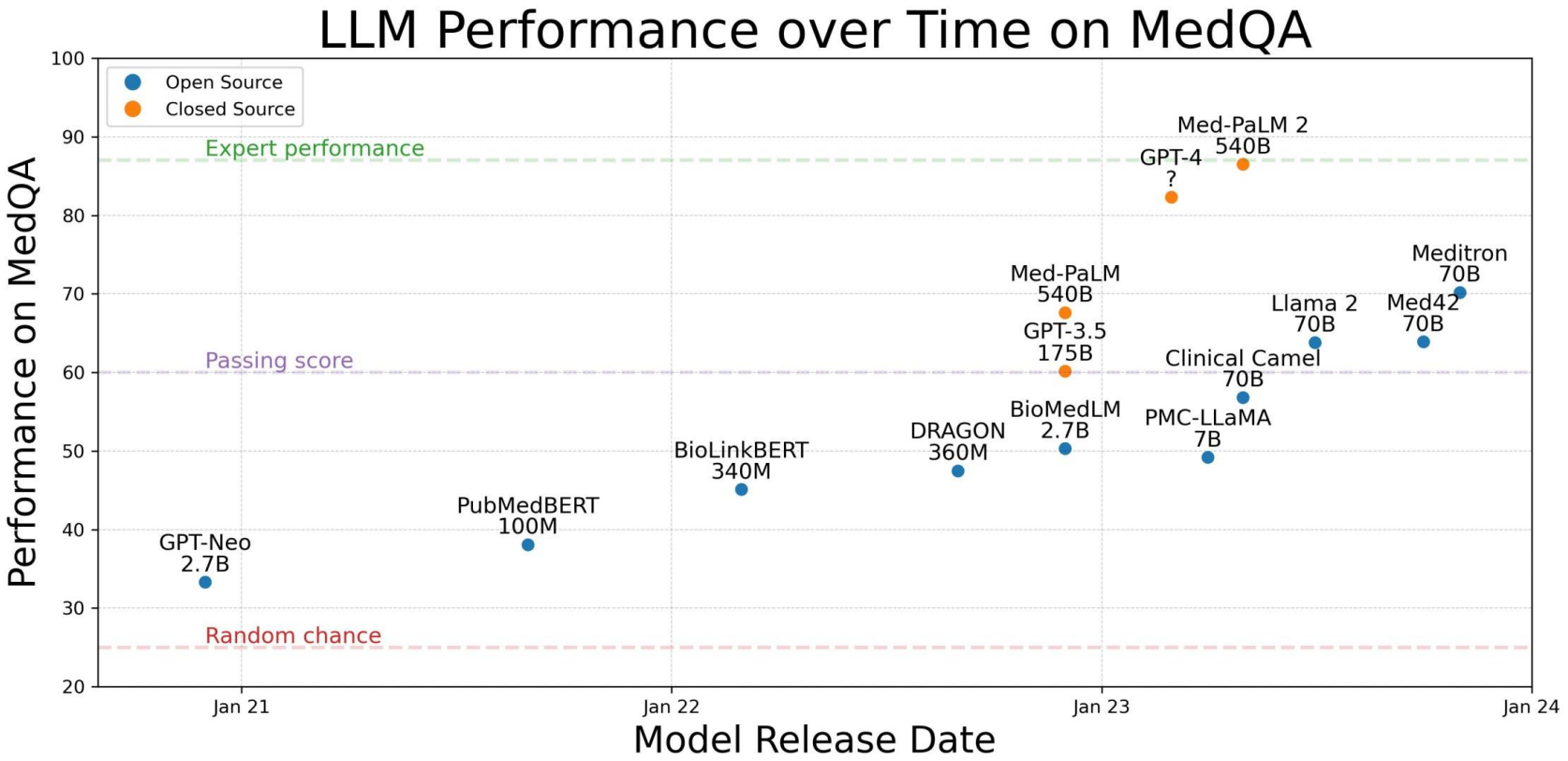
Chiang et al. "Chatbot arena: an open platform for evaluating LLMs by human preference". ICML 2024.

Challenges with Chatbot Arena

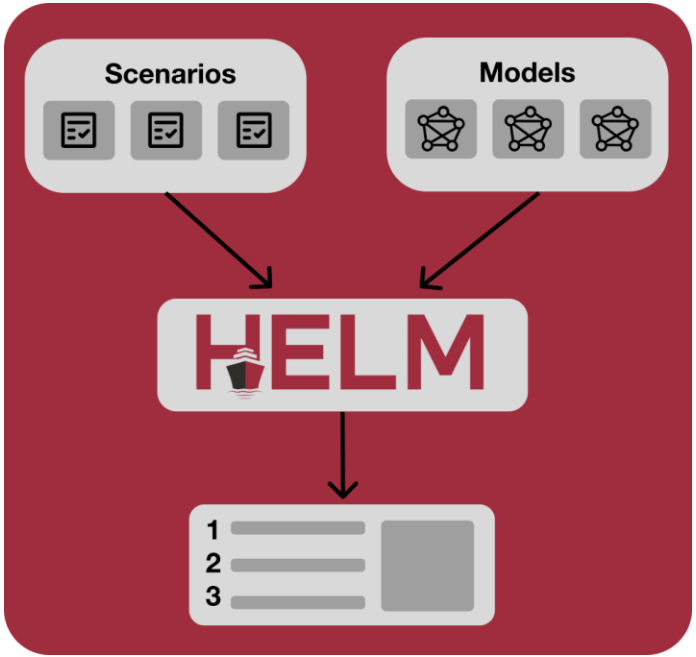


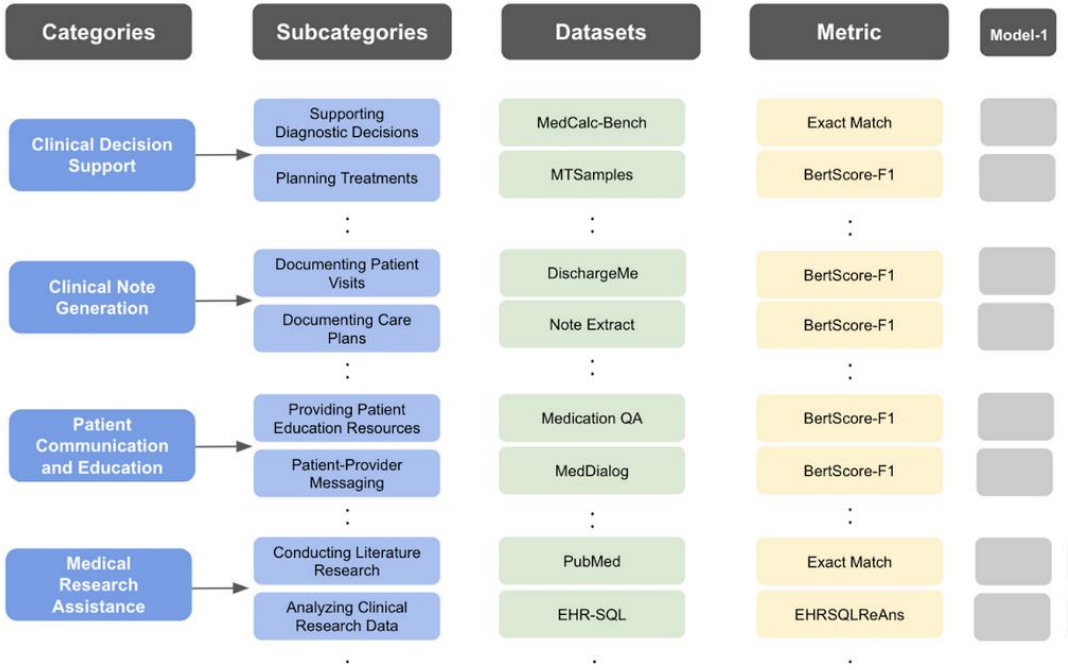
Singh et al. The Leaderboard Illusion. 2025.

Healthcare LLM Eval Any Better?



Scalable Evaluation Benchmarks





MedHELM

MedArena

🏥 MedArena - LLM Arena for Clinicians

Arena Leaderboard FAQ

📊 MedArena Leaderboard

Last updated: May 19, 2025 at 12:00 AM UTC

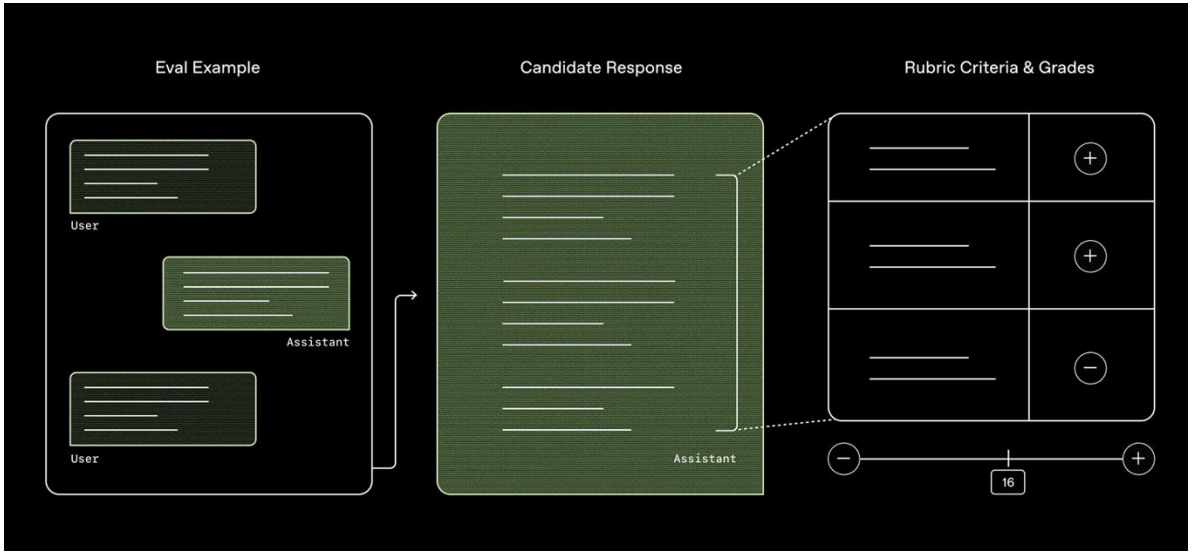
Legend

- 📚 Model supports RAG (Retrieval-Augmented Generation)
- 📷 Model supports Vision (Image Understanding)

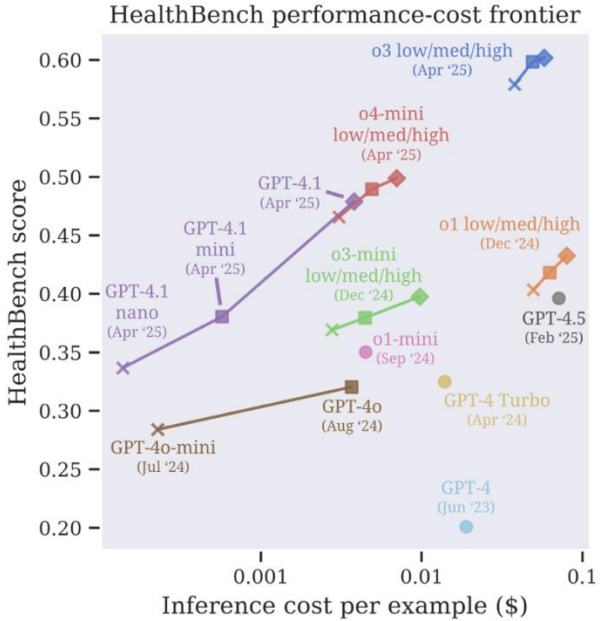
Model	BT Rating	BT CI (95…	Elo Rating	Elo CI (95%)	Win Ra…	Win Rate CI (95%)	Lose Rate	Battle
google/gemini-2.0-flash-thinking	1135	-37/+43	1103	-31/+32	0.58	0.535-0.626	0.313	460
openai/gpt-4o-2024-11-20 📷	1114	-34/+36	1066	-32/+30	0.53	0.48-0.576	0.344	477
google/gemini-2.5-pro	1112	-68/+74	1035	-30/+30	0.533	0.453-0.61	0.374	147
openai/gpt-4.5-preview 📷	1046	-65/+77	983	-27/+29	0.374	0.296-0.456	0.532	150
perplexity/llama-3.1-sonar-large-128k-online	1030	-39/+41	1004	-31/+34	0.425	0.373-0.478	0.458	326
google/gemini-2.0-flash 📚	1015	-57/+61	1008	-31/+29	0.446	0.369-0.525	0.419	164
openai/o3-mini	999	-38/+48	981	-31/+34	0.362	0.308-0.418	0.497	311

Health Bench

- New benchmark released by OpenAI in May 2025



"HealthBench is a rubric evaluation."



Arora et al. "HealthBench: Evaluating Large Language Models Towards Improved Human Health". arxiv 2025.

Health Bench

- Themes and rubric criteria

Table 2: Distribution of themes in HealthBench.

Theme	Count	(%)
Total examples	**5,000**	(100.0%)
Global health	1,097	(21.9%)
Responding under uncertainty	1,071	(21.4%)
Expertise-tailored communication	919	(18.4%)
Context seeking	594	(11.9%)
Emergency referrals	482	(9.6%)
Health data tasks	477	(9.5%)
Response depth	360	(7.2%)

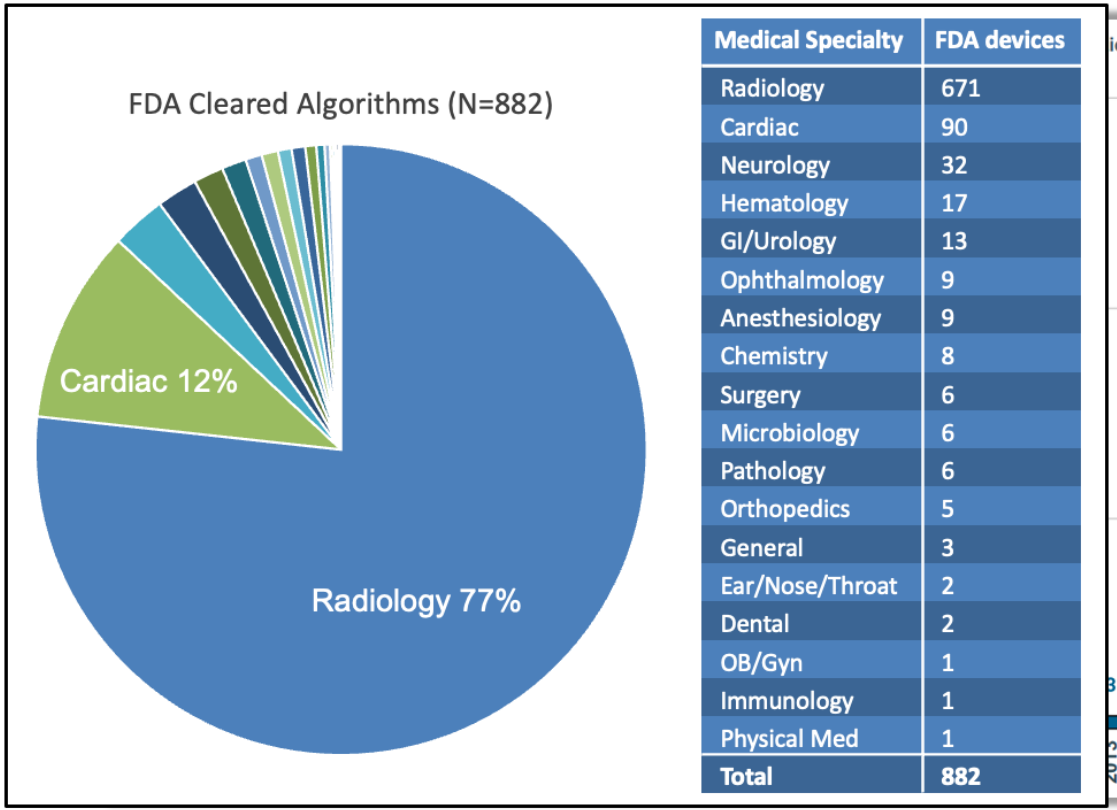
Table 3: Axes in HealthBench. Consensus rubric criteria are predefined and assigned by multiple physicians to an example, whereas example-specific criteria are written by physicians for each individual example.

Category	Count	(%)
All rubric criteria	**57,237**	**(100%)**
Consensus	8,053	(14%)
Example-specific	49,184	(86%)
Axis	**57,237**	**(100%)**
Completeness	22,285	(39%)
Accuracy	18,888	(33%)
Context awareness	8,991	(16%)
Communication quality	4,522	(8%)
Instruction following	2,551	(4%)

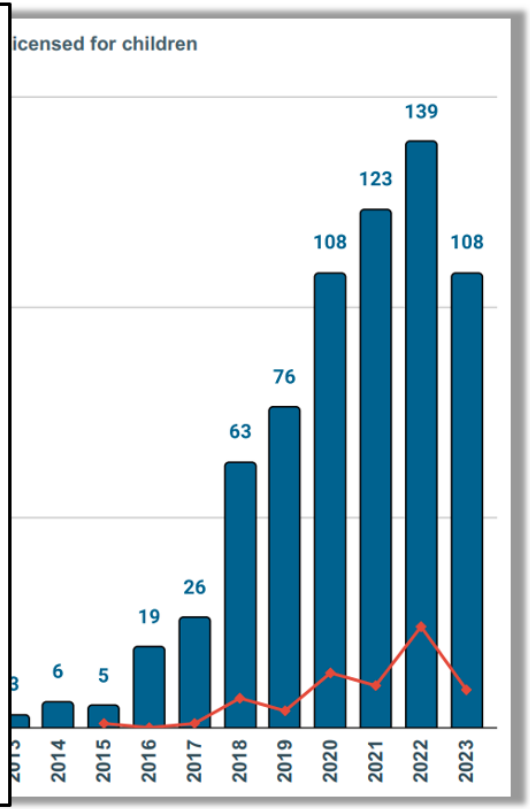
Arora et al. "HealthBench: Evaluating Large Language Models Towards Improved Human Health". arxiv 2025.

Evaluation Drives Translation

Regulation of Healthcare AI Models



Medical Specialty	FDA devices
Radiology	671
Cardiac	90
Neurology	32
Hematology	17
GI/Urology	13
Ophthalmology	9
Anesthesiology	9
Chemistry	8
Surgery	6
Microbiology	6
Pathology	6
Orthopedics	5
General	3
Ear/Nose/Throat	2
Dental	2
OB/Gyn	1
Immunology	1
Physical Med	1
Total	**882**



FDA Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. 2025.
Mulharidaran et al, A scoping review of reporting gaps in FDA-approved AI medical devices . npj Digital Medicine (2024)

Extensive Commercial Interest

AI vendors exhibiting at Radiological Society of North America 2023 Annual Meeting

Regulation of Healthcare AI Models

- Current FDA guidelines

Class 1

Low risk

- Includes: toothbrushes, tongue depressors, plasters
- 510(k) exempt
- FDA approval not needed
- Register device and company on FDA website

Time frame:
30-90 days

Some exceptions to 510(k) requirements

Class 2

Moderate risk

- Includes: pregnancy tests, blood pressure cuffs, surgical gloves
- FDA clearance required
- Most require 510(k) premarket notification

510(k) time frame:
1-9 months

510(k) & clinical trial requirements may vary

Class 3

High risk

- Includes: pacemakers, defibrillators, implanted devices
- FDA appproval required
- Premarket Approval process (PMA)
- Clinical trial duration 1 - 3 years

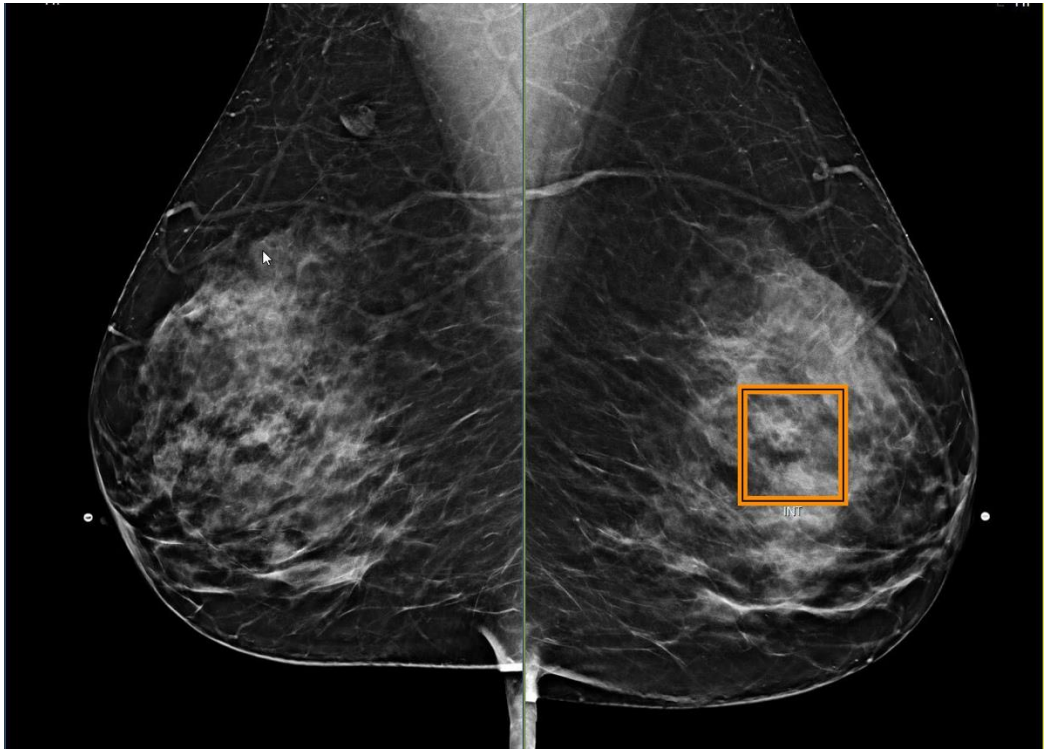
PMA approval time frame:
9-36 months

Byrne J. FDA 510(k) vs PMA: What's the Best Path for Your Device? 2025.

Computer Aided Diagnosis Triage (CADt)

O'Neil et al. Active Reprioritization of the Reading Worklist Using Artificial Intelligence Has a Beneficial Effect on the Turnaround Time for Interpretation of Head CT with Intracranial Hemorrhage. Radiology AI. 2024

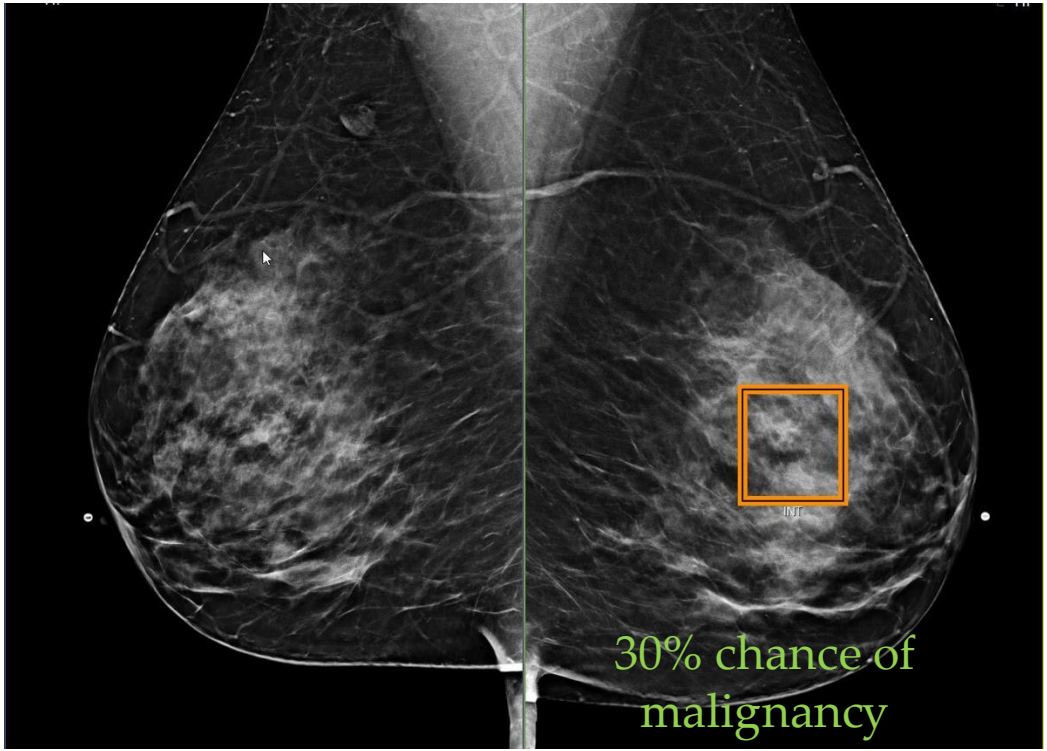
CAD e vs CAD x



- CAD e: Detection
  - Visual annotation overlay on image

Image Modified from Radiological Society of North America (Medical Xpress: Women pay for AI to boost mammogram findings)

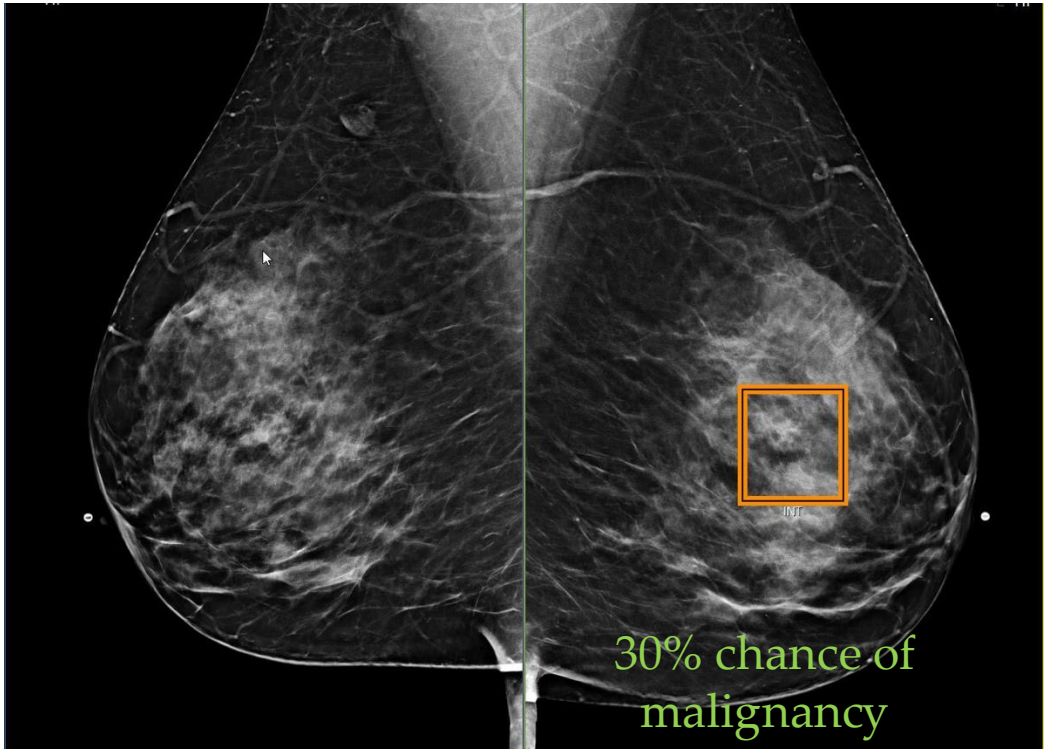
CAD e vs CAD x



- CAD e: Detection
  - Visual annotation overlay on image
- CAD x: Diagnosis
  - Prediction of some underlying health status/prognosis

Image Modified from Radiological Society of North America (Medical Xpress: Women pay for AI to boost mammogram findings)

CAD e vs CAD x



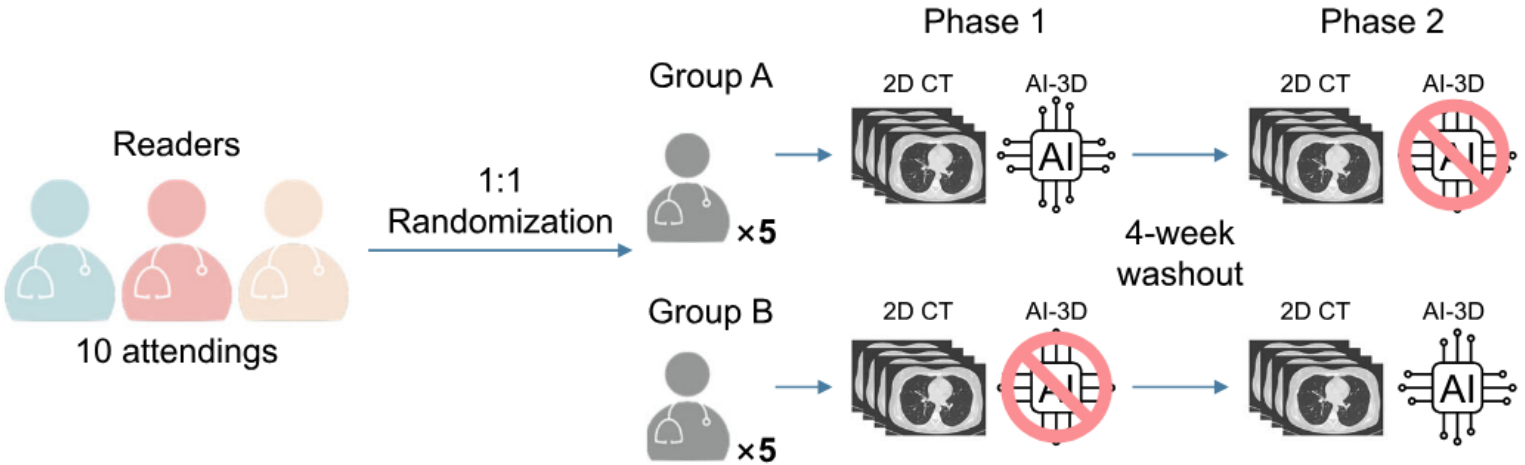
- CAD e: Detection
  - Visual annotation overlay on image
- CAD x: Diagnosis
  - Prediction of some underlying health status/prognosis
- CAD e/x: Both

Image Modified from Radiological Society of North America (Medical Xpress: Women pay for AI to boost mammogram findings)

Common Validation Studies

- **CAD t**: Standalone study comparing AI model to experts

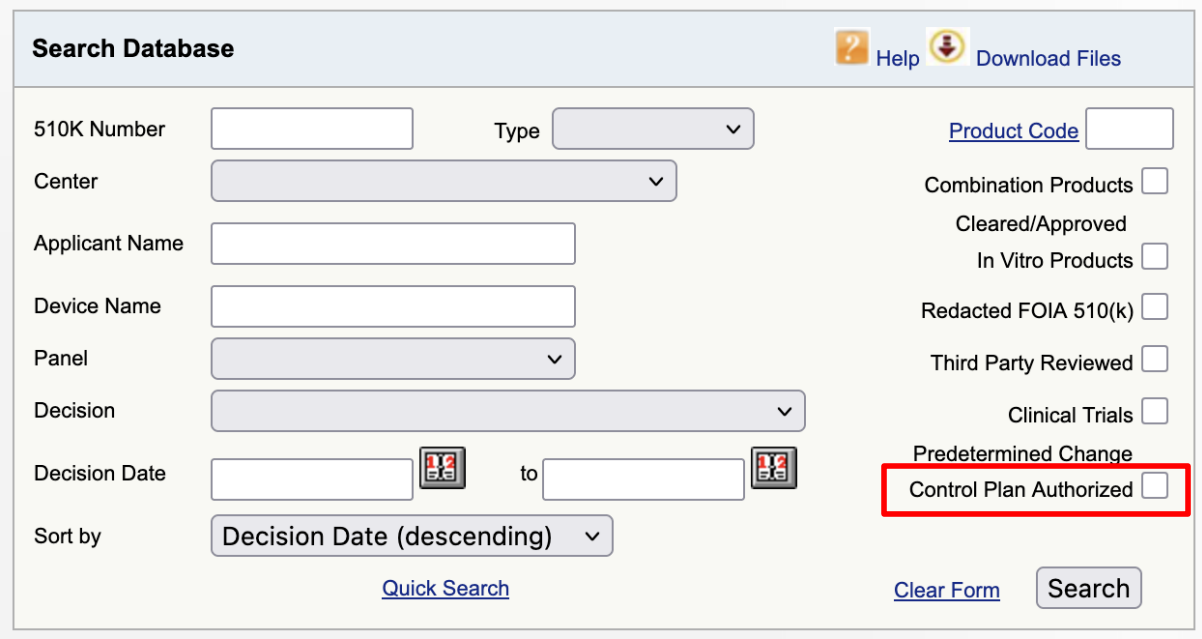
- **CAD e/x**: Standalone study + multi-reader multi-case study



Chen et al. Artificial intelligence driven 3D reconstruction for enhanced lung surgery planning. Nature Communications 2025.

Transparency of Regulatory AI Eval

- Searchable database of regulated products



Transparency of Regulatory AI Eval

- Example product – Bunkerhill BMD

Proposed Device

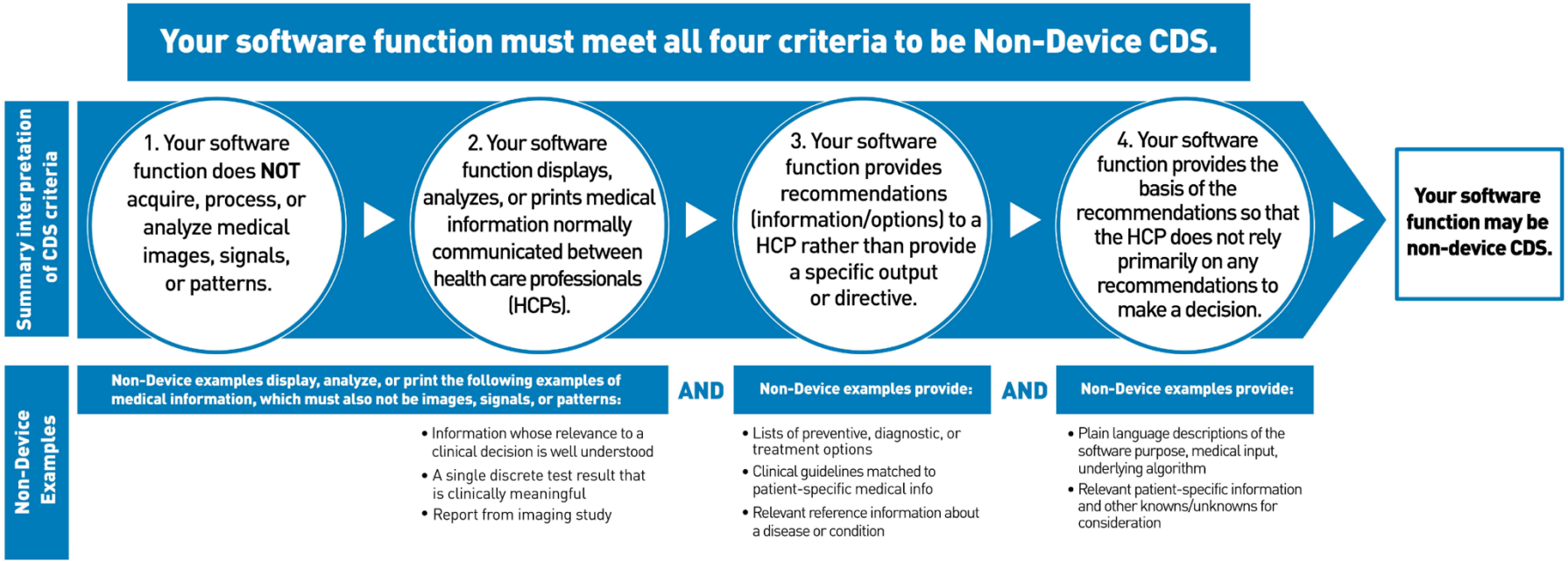
Proprietary Name	Bunkerhill BMD
Classification Name	Bone Densitometer
Regulation Number	21 CFR 892.1170
Product Code	KGI
Regulatory Class	II

Predicate Device

Proprietary Name	ABMD software
Premarket Notification	K213760
Classification Name	Bone Densitometer
Regulation Number	21 CFR 892.1170
Product Code	KGI
Regulatory Class	II

- *Bunkerhill BMD performance was validated in a stand-alone retrospective study for overall agreement of the device output compared to the established ground truth.*
- *The pivotal testing dataset consisted of 371 CT studies from four (4) geographically diverse sites.*
- *The Bunkerhill BMD algorithm achieved a sensitivity of 81.0 (74.0 - 86.8) and specificity of78.4 (72.3 - 83.7),…*

What is Not Regulated?



FDA. Clinical Decision Support Software Frequently Asked Questions (FAQs). 2024.

What is Not Regulated?

- Ambient scribes record the conversation between a patient and physician
- Transcription tools generate text and create an encounter note
- Encounter note can include diagnoses that would be used for billing

Reminder Slide from Past Lecture

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.

Not Regulated

Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine, 2024.

Future Approaches for Evaluation

- Shifting focusing from pre-deployment to total life cycle
- Emphasizing need for continuous monitoring of deployed models

Clinical Review & Education

JAMA | Special Communication | AI IN MEDICINE

FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine

Haider J. Warraich, MD; Troy Tazbaz, BS; Robert M. Califf, MD

Request from CA Attorney General

State of California
Office of the Attorney General
ROB BONTA
ATTORNEY GENERAL

August 31, 2022

- Key Requests sent to all Hospital CEOs in California:

- A list of all commercially available or purchased decision-making tools, products, software systems, or algorithmic methodologies currently in use that assist or contribute to the performance of any of the following functions:
- The purposes for which these tools are currently used, how these tools inform decisions, and any policies, procedures, training, or protocols that apply to use of these tools; and
- The name or contact information of the person(s) responsible for evaluating the purpose and use of these tools and ensuring that they do not have a disparate impact based on race or other protected characteristics.