🏥 MedArena.ai

A Clinician-Centric Platform for Expert Preferences on Medical Queries

Eric Wu, Kevin Wu, James Zou
Stanford University

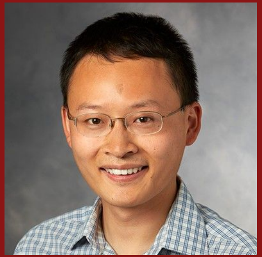
Stanford University

Stanford University

Stanford University

Eric Wu

Kevin Wu

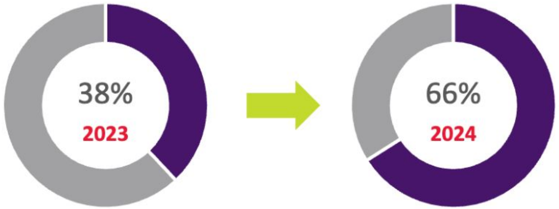


James Zou

GXL

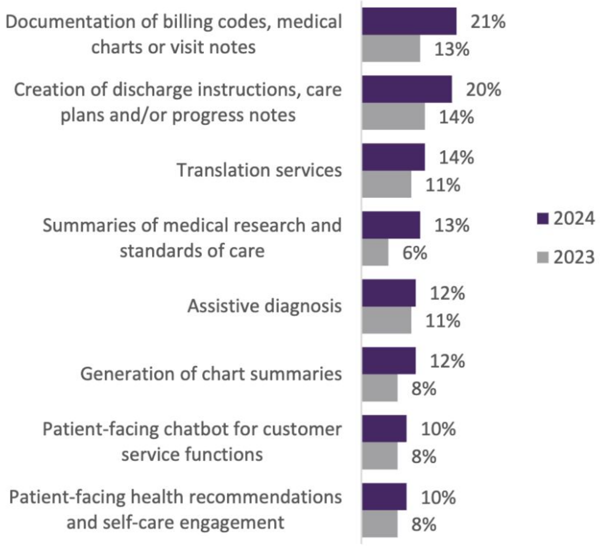
Two-thirds of clinicians report using AI tools

Rapid increase in AI users



Percent of respondents stating they currently use at least one of the 15 AI use cases presented

Top use cases gain users



American Medical Association, 2025

But LLM evaluation in medicine lags behind

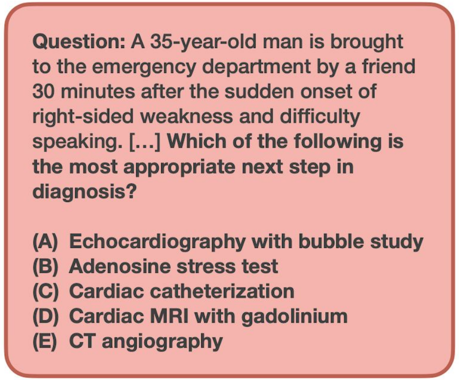
LLMs are evaluated on fixed multiple-choice datasets:

- MedQA
- MedMCQA
- PubMedQA
- MMLU
- etc.

Problem: Real-world clinical questions are not *fixed* and do not present with *multiple choices*

QA datasets are *static* and don't evolve over time.

MedQA



Question: A 35-year-old man is brought to the emergency department by a friend 30 minutes after the sudden onset of right-sided weakness and difficulty speaking. [...] **Which of the following is the most appropriate next step in diagnosis?**

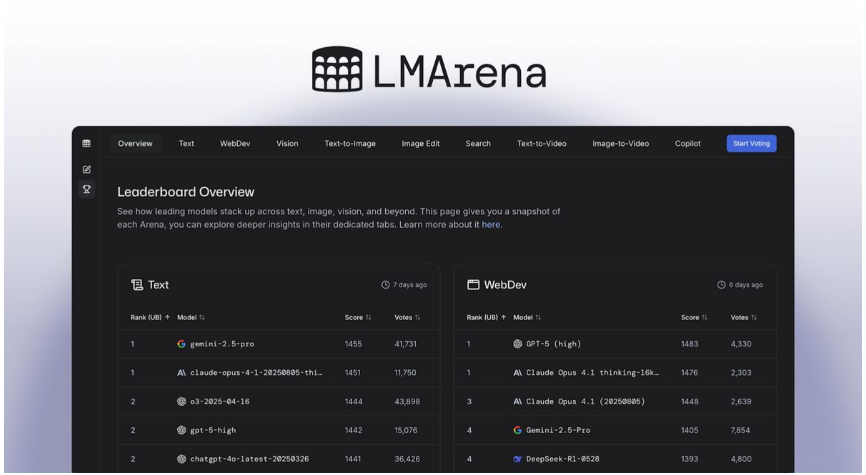
(A) Echocardiography with bubble study
(B) Adenosine stress test
(C) Cardiac catheterization
(D) Cardiac MRI with gadolinium
(E) CT angiography

USMLE℠
United States Medical Licensing Exam

What questions do clinicians ask LLMs?

Which LLMs do clinicians prefer?

Existing benchmarks moving beyond static QA



Leaderboard Overview

See how leading models stack up across text, image, vision, and beyond. This page gives you a snapshot of each Arena, you can explore deeper insights in their dedicated tabs. Learn more about it here.

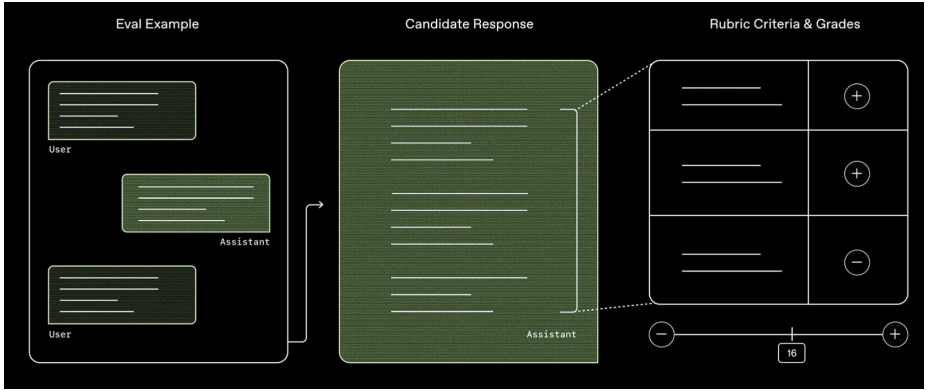
Text — 7 days ago

Rank (UB)	Model	Score	Votes
1	gemini-2.5-pro	1455	41,731
1	claude-opus-4-1-20250805-thi…	1451	11,750
2	o3-2025-04-16	1444	43,898
2	gpt-5-high	1442	15,076
2	chatgpt-4o-latest-20250326	1441	36,426

WebDev — 6 days ago

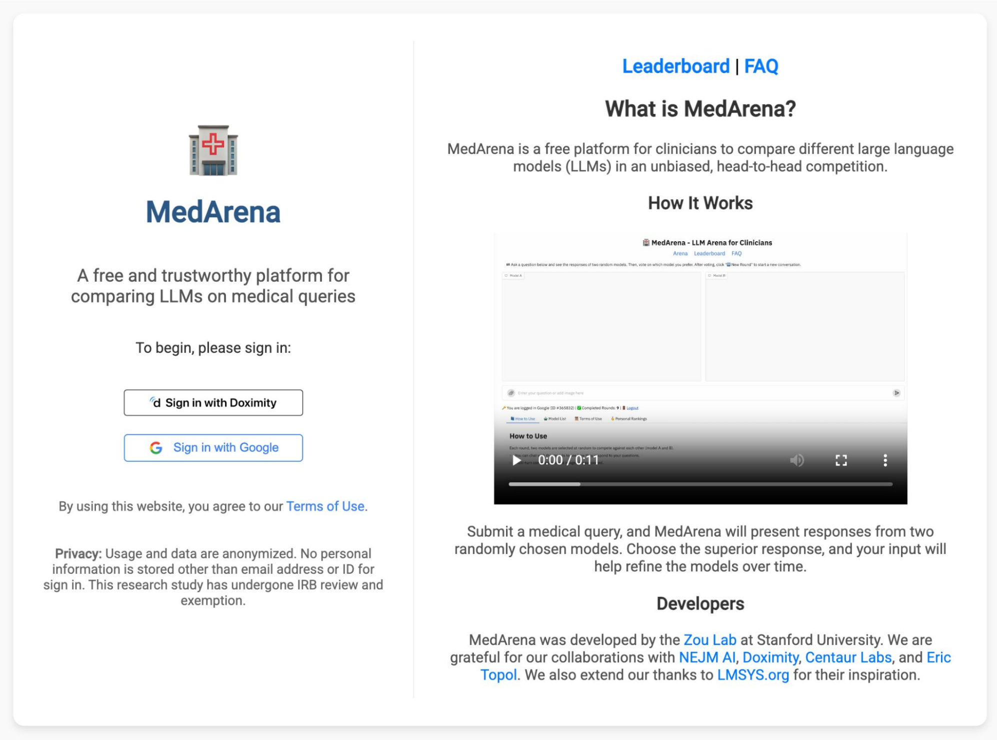
Rank (UB)	Model	Score	Votes
1	GPT-5 (high)	1483	4,330
1	Claude Opus 4.1 thinking-16k…	1476	2,303
3	Claude Opus 4.1 (20250805)	1448	2,639
4	Gemini-2.5-Pro	1405	7,854
4	DeepSeek-R1-0528	1393	4,800

OpenAI HealthBench



MedArena.ai

MedArena is a *free*, **clinician-only** platform for comparing LLM responses to medical queries



MedArena

A free and trustworthy platform for comparing LLMs on medical queries

To begin, please sign in:

By using this website, you agree to our Terms of Use.

Privacy: Usage and data are anonymized. No personal information is stored other than email address or ID for sign in. This research study has undergone IRB review and exemption.

What is MedArena?

MedArena is a free platform for clinicians to compare different large language models (LLMs) in an unbiased, head-to-head competition.

How It Works

Submit a medical query, and MedArena will present responses from two randomly chosen models. Choose the superior response, and your input will help refine the models over time.

Developers

MedArena was developed by the Zou Lab at Stanford University. We are grateful for our collaborations with NEJM AI, Doximity, Centaur Labs, and Eric Topol. We also extend our thanks to LMSYS.org for their inspiration.

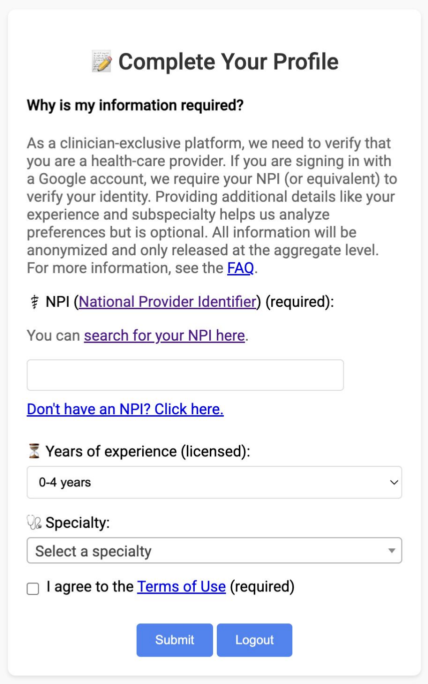
Clinician-only platform

- Clinicians are authenticated via National Provider Identifier (**NPI**) or **Doximity**
- International clinicians provide relevant credentials
- Reach out via physician networks

NPPES
National Plan & Provider Enumeration System

doximity

User data is anonymized



📝 Complete Your Profile

Why is my information required?

As a clinician-exclusive platform, we need to verify that you are a health-care provider. If you are signing in with a Google account, we require your NPI (or equivalent) to verify your identity. Providing additional details like your experience and subspecialty helps us analyze preferences but is optional. All information will be anonymized and only released at the aggregate level. For more information, see the FAQ.

⚕ NPI (National Provider Identifier) (required):

You can search for your NPI here.

Don't have an NPI? Click here.

⏳ Years of experience (licensed): 0-4 years

🩺 Specialty: Select a specialty

☐ I agree to the Terms of Use (required)

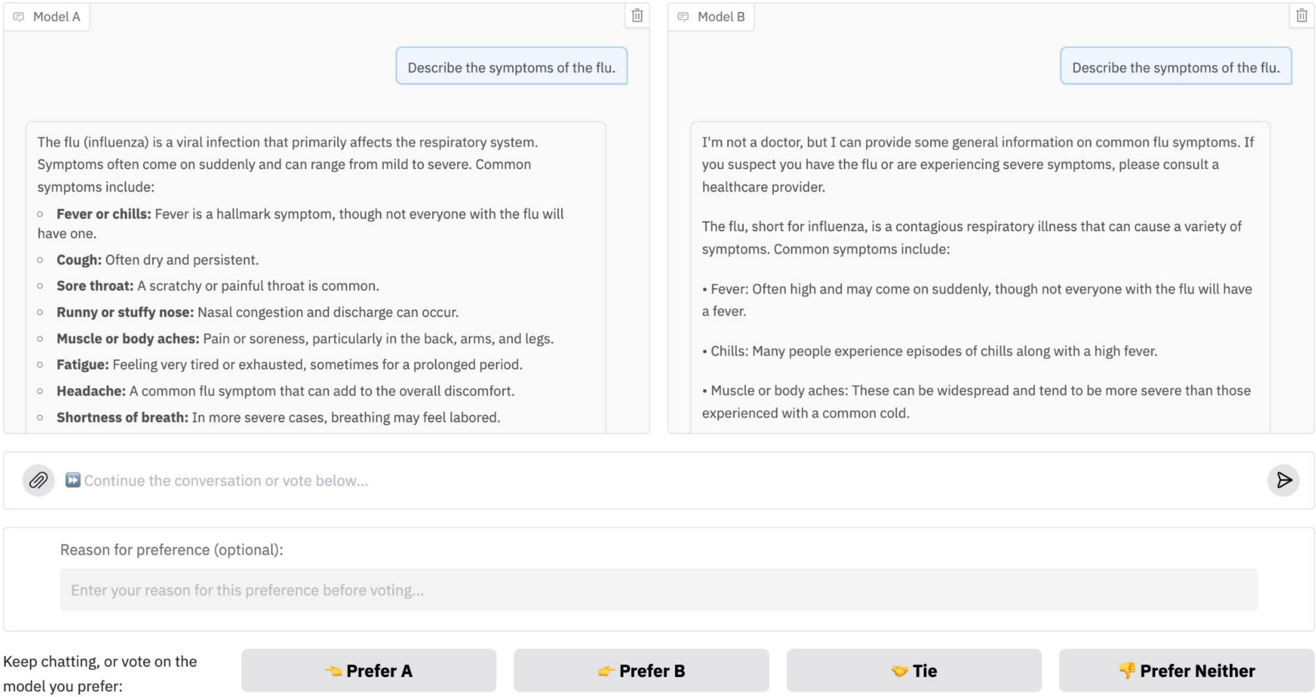
Information collected:

- Credentials
- Years of experience (optional)
- Specialty (optional)

Privacy:

- No personal information is stored
- Usage data is anonymized
- PHI is not allowed
- IRB exempt

Query receives two random model responses



Describe the symptoms of the flu.

The flu (influenza) is a viral infection that primarily affects the respiratory system. Symptoms often come on suddenly and can range from mild to severe. Common symptoms include:

- **Fever or chills:** Fever is a hallmark symptom, though not everyone with the flu will have one.
- **Cough:** Often dry and persistent.
- **Sore throat:** A scratchy or painful throat is common.
- **Runny or stuffy nose:** Nasal congestion and discharge can occur.
- **Muscle or body aches:** Pain or soreness, particularly in the back, arms, and legs.
- **Fatigue:** Feeling very tired or exhausted, sometimes for a prolonged period.
- **Headache:** A common flu symptom that can add to the overall discomfort.
- **Shortness of breath:** In more severe cases, breathing may feel labored.

Describe the symptoms of the flu.

I'm not a doctor, but I can provide some general information on common flu symptoms. If you suspect you have the flu or are experiencing severe symptoms, please consult a healthcare provider.

The flu, short for influenza, is a contagious respiratory illness that can cause a variety of symptoms. Common symptoms include:

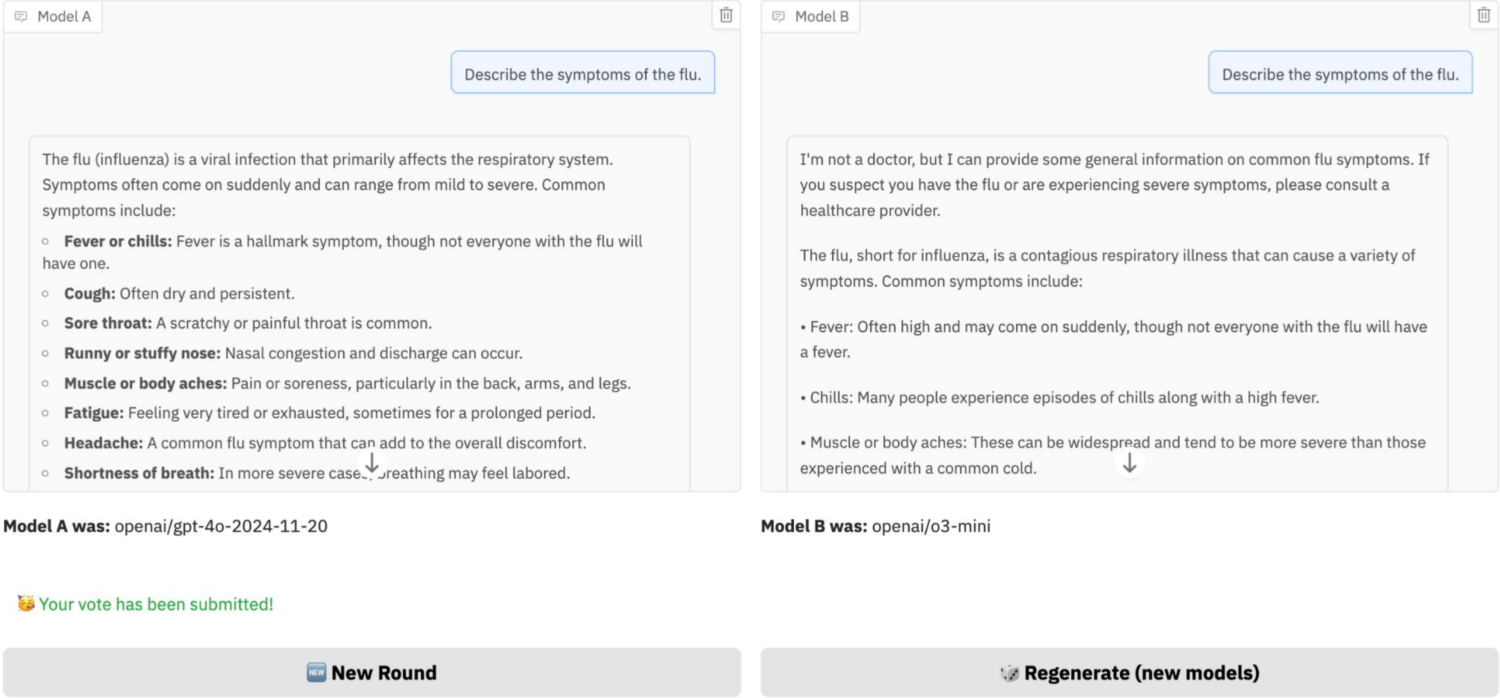
- Fever: Often high and may come on suddenly, though not everyone with the flu will have a fever.

- Chills: Many people experience episodes of chills along with a high fever.

- Muscle or body aches: These can be widespread and tend to be more severe than those experienced with a common cold.

⏩ Continue the conversation or vote below...
Reason for preference (optional):
Enter your reason for this preference before voting...
Keep chatting, or vote on the model you prefer:
👈 Prefer A
👉 Prefer B
🤝 Tie
👎 Prefer Neither

User submits a preference with optional reason



Describe the symptoms of the flu.

The flu (influenza) is a viral infection that primarily affects the respiratory system. Symptoms often come on suddenly and can range from mild to severe. Common symptoms include:

- **Fever or chills:** Fever is a hallmark symptom, though not everyone with the flu will have one.
- **Cough:** Often dry and persistent.
- **Sore throat:** A scratchy or painful throat is common.
- **Runny or stuffy nose:** Nasal congestion and discharge can occur.
- **Muscle or body aches:** Pain or soreness, particularly in the back, arms, and legs.
- **Fatigue:** Feeling very tired or exhausted, sometimes for a prolonged period.
- **Headache:** A common flu symptom that can add to the overall discomfort.
- **Shortness of breath:** In more severe cases, breathing may feel labored.

Describe the symptoms of the flu.

I'm not a doctor, but I can provide some general information on common flu symptoms. If you suspect you have the flu or are experiencing severe symptoms, please consult a healthcare provider.

The flu, short for influenza, is a contagious respiratory illness that can cause a variety of symptoms. Common symptoms include:

- Fever: Often high and may come on suddenly, though not everyone with the flu will have a fever.
- Chills: Many people experience episodes of chills along with a high fever.
- Muscle or body aches: These can be widespread and tend to be more severe than those experienced with a common cold.

Model A was: openai/gpt-4o-2024-11-20

Model B was: openai/o3-mini

🥳 Your vote has been submitted!

What questions do clinicians ask LLMs?

Fixed QA → Free-text

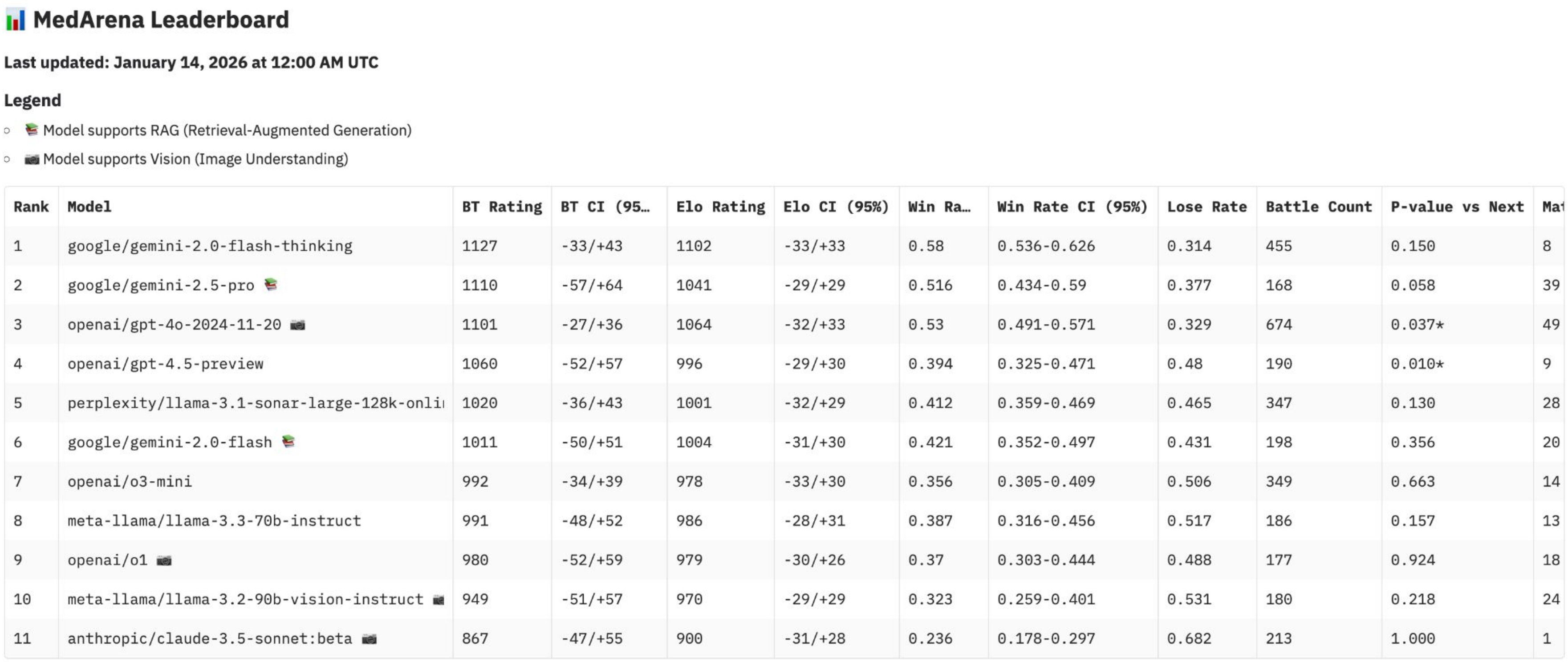
Static →Live

Which LLMs do clinicians prefer?

Preferences (rankings + reasons)

Updated with latest models

Leaderboard based on clinician preferences



📊 MedArena Leaderboard

Last updated: January 14, 2026 at 12:00 AM UTC

Legend

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

Rank	Model	BT Rating	BT CI (95…	Elo Rating	Elo CI (95%)	Win Ra…	Win Rate CI (95%)	Lose Rate	Battle Count	P-value vs Next	Ma
1	google/gemini-2.0-flash-thinking	1127	-33/+43	1102	-33/+33	0.58	0.536-0.626	0.314	455	0.150	8
2	google/gemini-2.5-pro 📚	1110	-57/+64	1041	-29/+29	0.516	0.434-0.59	0.377	168	0.058	39
3	openai/gpt-4o-2024-11-20 📷	1101	-27/+36	1064	-32/+33	0.53	0.491-0.571	0.329	674	0.037*	49
4	openai/gpt-4.5-preview	1060	-52/+57	996	-29/+30	0.394	0.325-0.471	0.48	190	0.010*	9
5	perplexity/llama-3.1-sonar-large-128k-onli	1020	-36/+43	1001	-32/+29	0.412	0.359-0.469	0.465	347	0.130	28
6	google/gemini-2.0-flash 📚	1011	-50/+51	1004	-31/+30	0.421	0.352-0.497	0.431	198	0.356	20
7	openai/o3-mini	992	-34/+39	978	-33/+30	0.356	0.305-0.409	0.506	349	0.663	14
8	meta-llama/llama-3.3-70b-instruct	991	-48/+52	986	-28/+31	0.387	0.316-0.456	0.517	186	0.157	13
9	openai/o1 📷	980	-52/+59	979	-30/+26	0.37	0.303-0.444	0.488	177	0.924	18
10	meta-llama/llama-3.2-90b-vision-instruct 📷	949	-51/+57	970	-29/+29	0.323	0.259-0.401	0.531	180	0.218	24
11	anthropic/claude-3.5-sonnet:beta 📷	867	-47/+55	900	-31/+28	0.236	0.178-0.297	0.682	213	1.000	1

Elo Rating

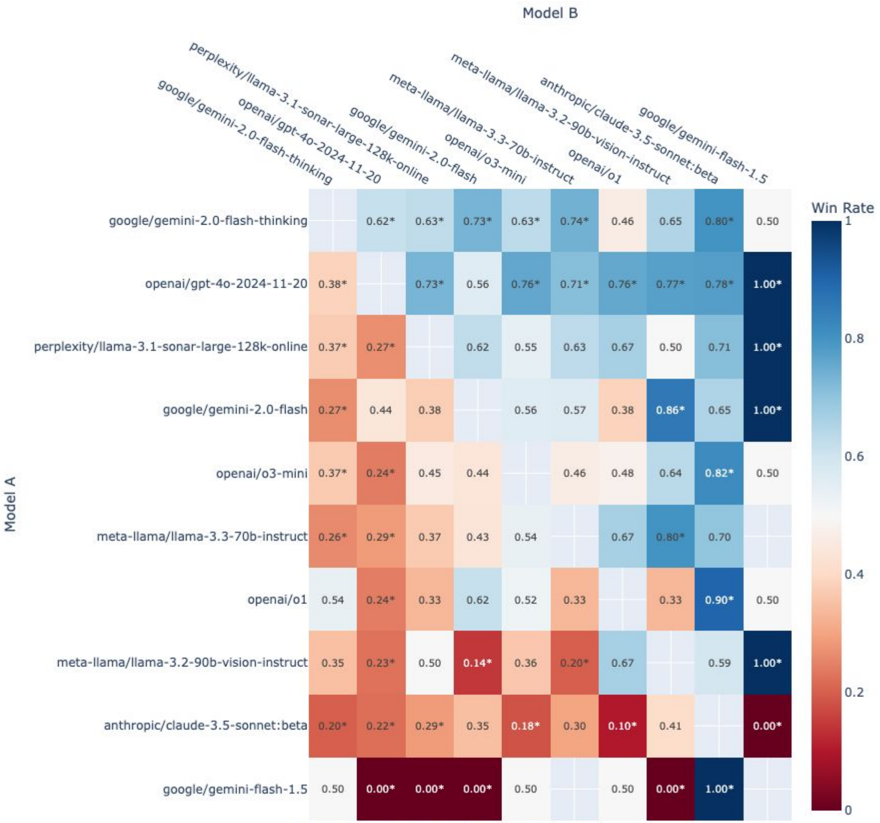
For each matchup, both models’ ratings are **updated** based on the **difference** between the **actual outcome** and their **expected scores**, using the Elo rating formula.

- Larger rating difference leads to larger rating change
- K=4 for stable ratings
- Base rating = 1000

$$E_{\mathsf{A}} = \frac{1}{1 + 10^{(R_{\mathsf{B}} - R_{\mathsf{A}})/400}}$$

$$R'_{\mathsf{A}} = R_{\mathsf{A}} + K \cdot (S_{\mathsf{A}} - E_{\mathsf{A}})$$

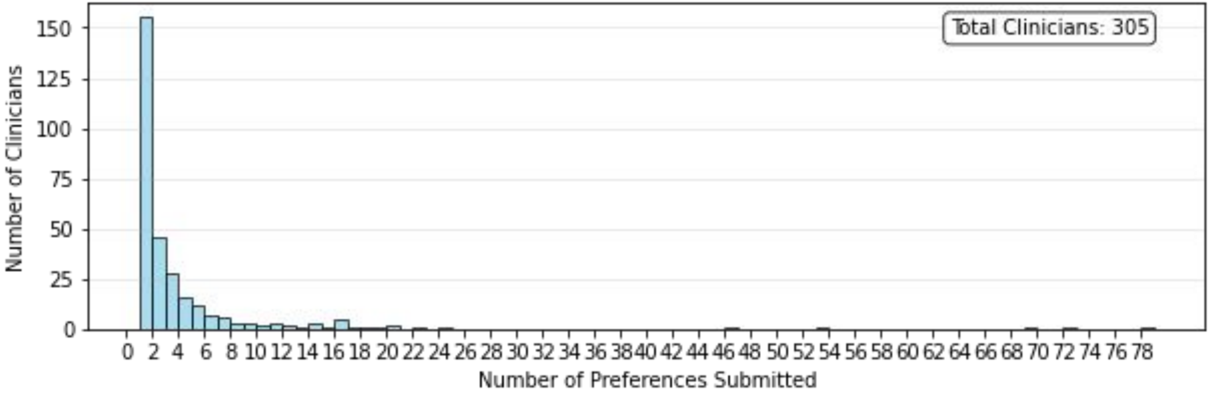
Pairwise head-to-head matchups

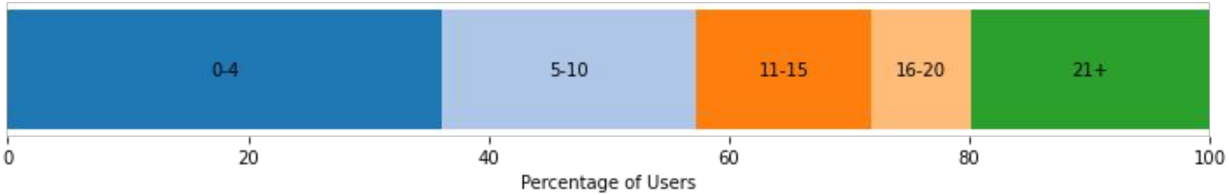


	google/gemini-2.0-flash-thinking	openai/gpt-4o-2024-11-20	perplexity/llama-3.1-sonar-large-128k-online	google/gemini-2.0-flash	openai/o3-mini	meta-llama/llama-3.3-70b-instruct	openai/o1	meta-llama/llama-3.2-90b-vision-instruct	anthropic/claude-3.5-sonnet:beta	google/gemini-flash-1.5
google/gemini-2.0-flash-thinking		0.62*	0.63*	0.73*	0.63*	0.74*	0.46	0.65	0.80*	0.50
openai/gpt-4o-2024-11-20	0.38*		0.73*	0.56	0.76*	0.71*	0.76*	0.77*	0.78*	1.00*
perplexity/llama-3.1-sonar-large-128k-online	0.37*	0.27*		0.62	0.55	0.63	0.67	0.50	0.71	1.00*
google/gemini-2.0-flash	0.27*	0.44	0.38		0.56	0.57	0.38	0.86*	0.65	1.00*
openai/o3-mini	0.37*	0.24*	0.45	0.44		0.46	0.48	0.64	0.82*	0.50
meta-llama/llama-3.3-70b-instruct	0.26*	0.29*	0.37	0.43	0.54		0.67	0.80*	0.70	
openai/o1	0.54	0.24*	0.33	0.62	0.52	0.33		0.33	0.90*	0.50
meta-llama/llama-3.2-90b-vision-instruct	0.35	0.23*	0.50	0.14*	0.36	0.20*	0.67		0.59	1.00*
anthropic/claude-3.5-sonnet:beta	0.20*	0.22*	0.29*	0.35	0.18*	0.30	0.10*	0.41		0.00*
google/gemini-flash-1.5	0.50	0.00*	0.00*	0.00*	0.50		0.50	0.00*	1.00*	

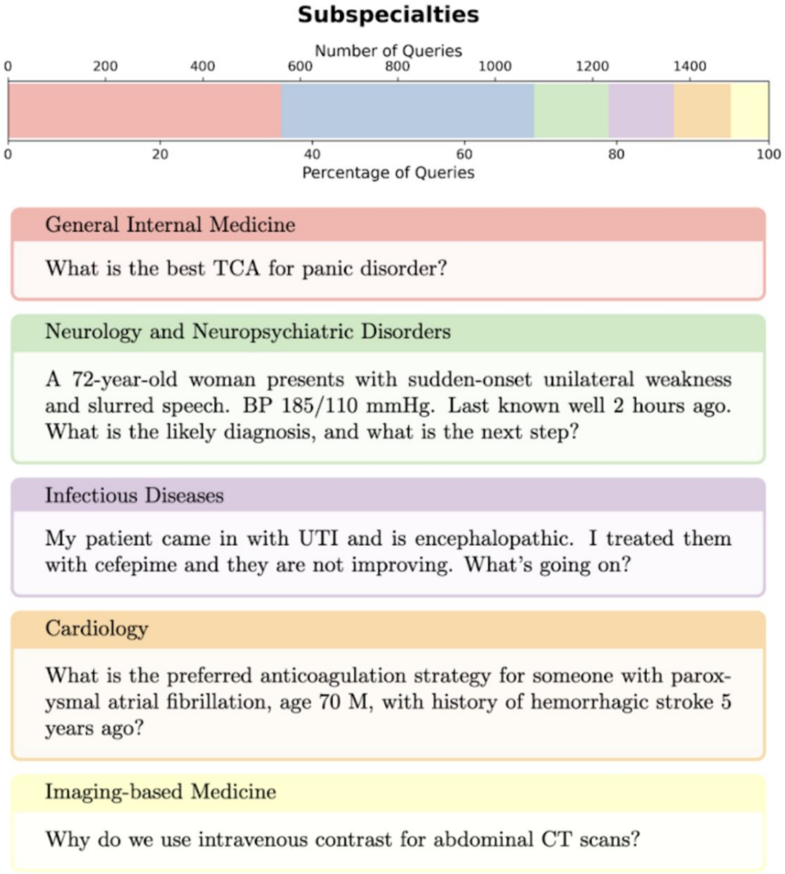
Pairwise Win Rates (Row vs Column)

Clinician Sign-Ups


Number of Preferences from Unique Clinicians


Years of Licensed Clinical Experience

Clinician specialities



General Internal Medicine

What is the best TCA for panic disorder?

Neurology and Neuropsychiatric Disorders

A 72-year-old woman presents with sudden-onset unilateral weakness and slurred speech. BP 185/110 mmHg. Last known well 2 hours ago. What is the likely diagnosis, and what is the next step?

Infectious Diseases

My patient came in with UTI and is encephalopathic. I treated them with cefepime and they are not improving. What's going on?

Cardiology

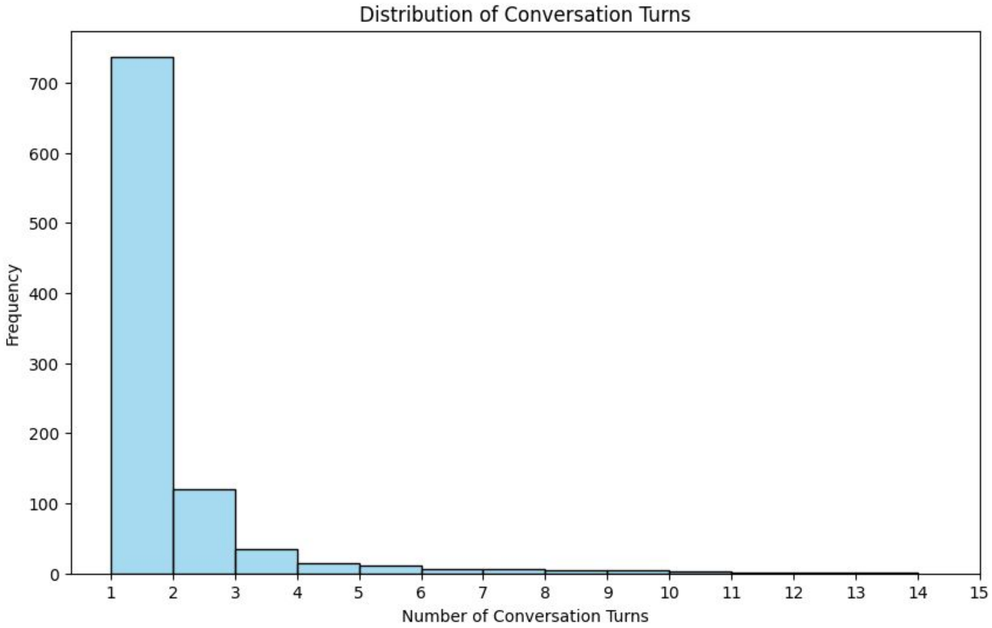
What is the preferred anticoagulation strategy for someone with paroxysmal atrial fibrillation, age 70 M, with history of hemorrhagic stroke 5 years ago?

Imaging-based Medicine

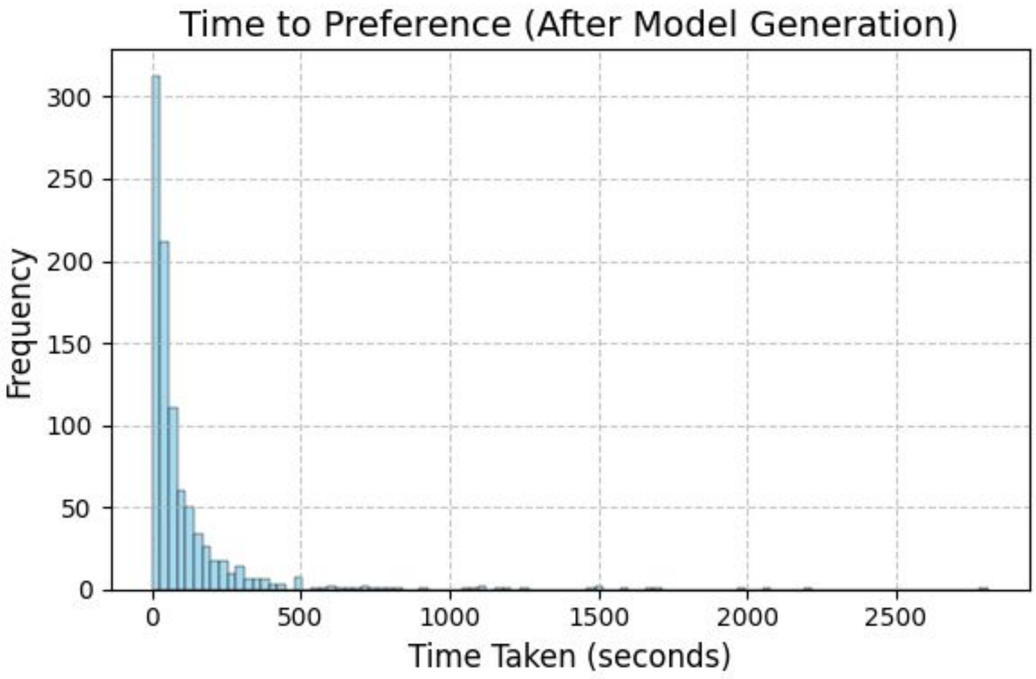
Why do we use intravenous contrast for abdominal CT scans?

Number of conversation turns

- 22% of conversations are multi-turn
- *Longer convos are over distinct questions

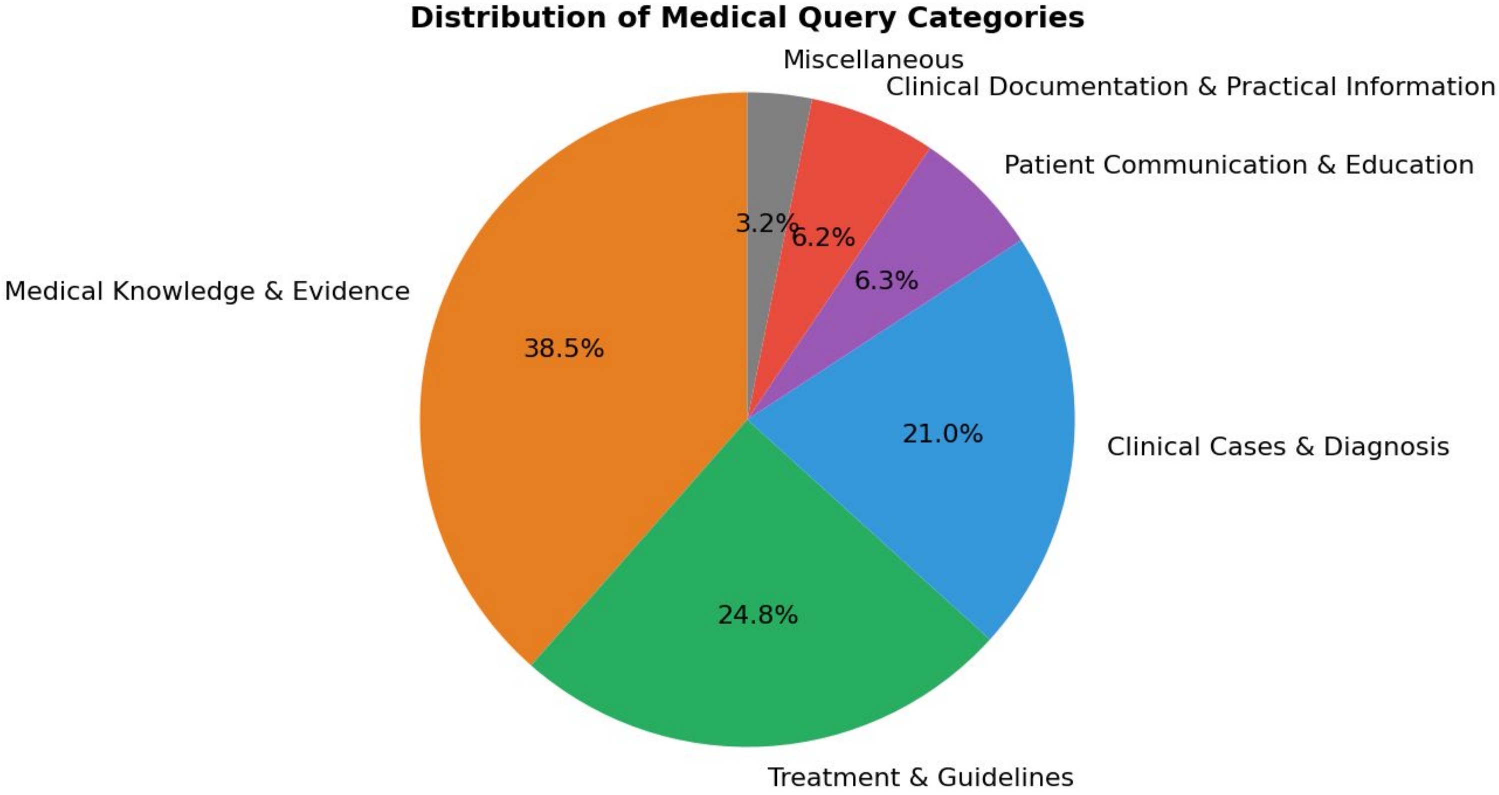


Median clinician response time is 47 seconds

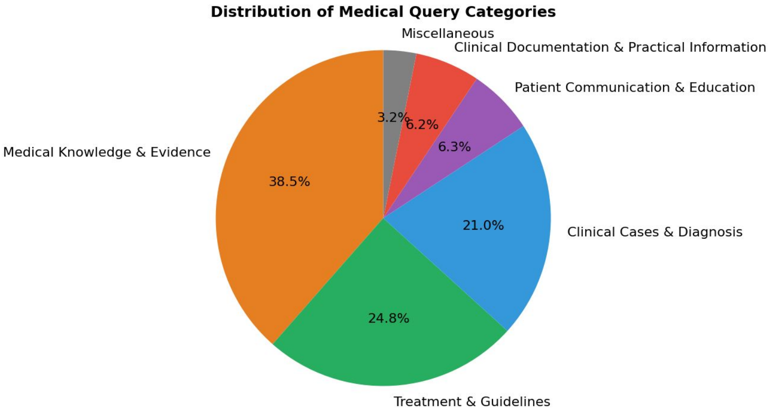


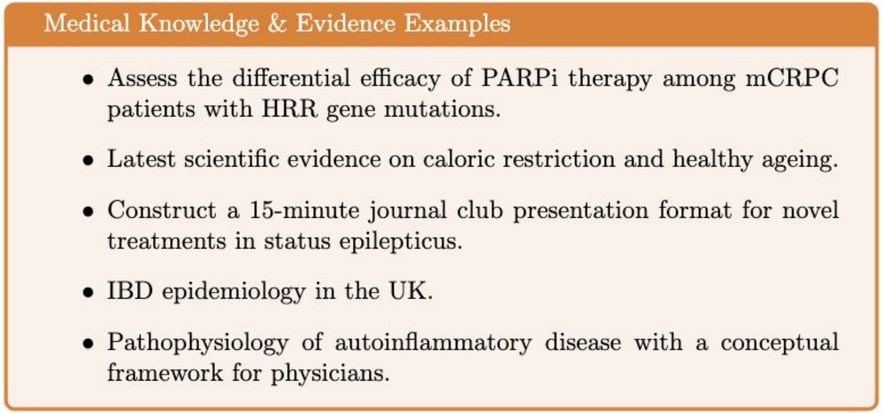
What kinds of questions are clinicians asking?

What kinds of questions are clinicians asking?



What kinds of questions are clinicians asking?



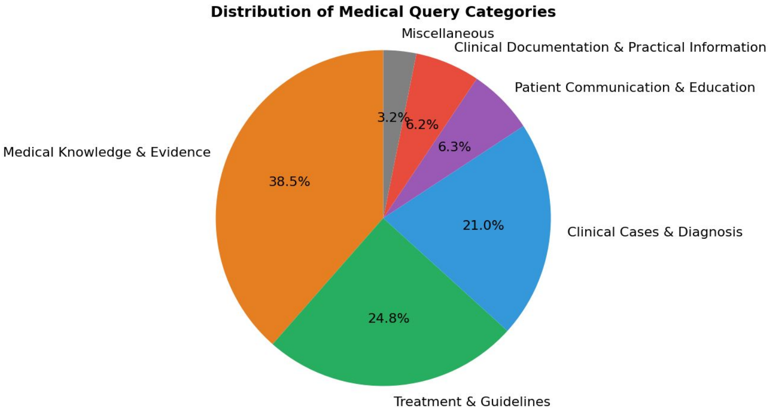


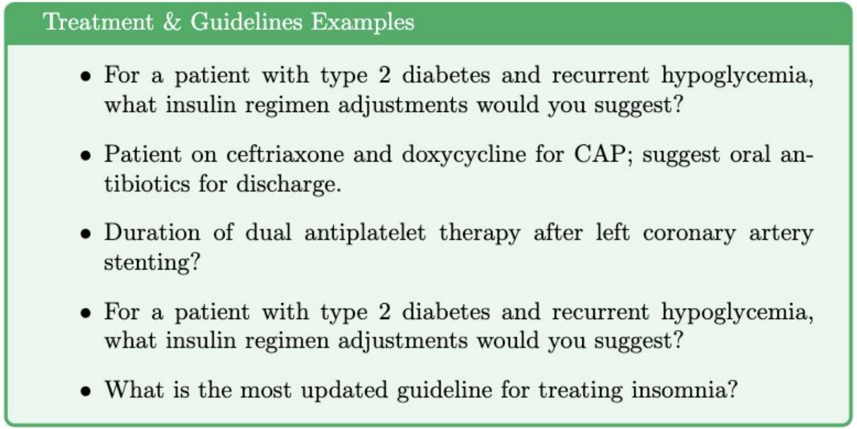
Medical Knowledge & Evidence Examples

- Assess the differential efficacy of PARPi therapy among mCRPC patients with HRR gene mutations.
- Latest scientific evidence on caloric restriction and healthy ageing.
- Construct a 15-minute journal club presentation format for novel treatments in status epilepticus.
- IBD epidemiology in the UK.
- Pathophysiology of autoinflammatory disease with a conceptual framework for physicians.

Asking the model to recall and organize medical knowledge and evidence

What kinds of questions are clinicians asking?



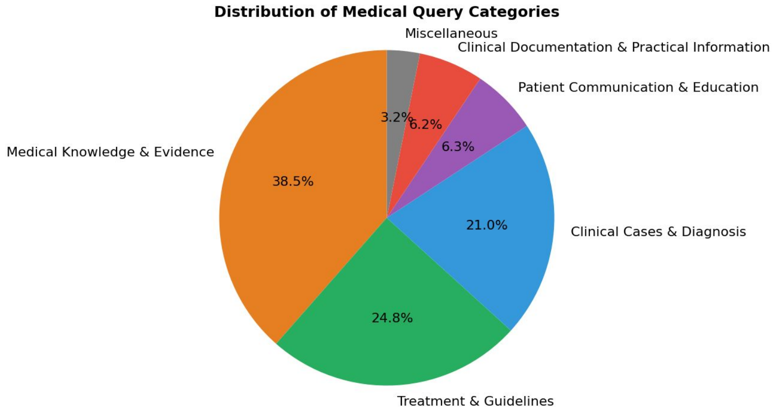


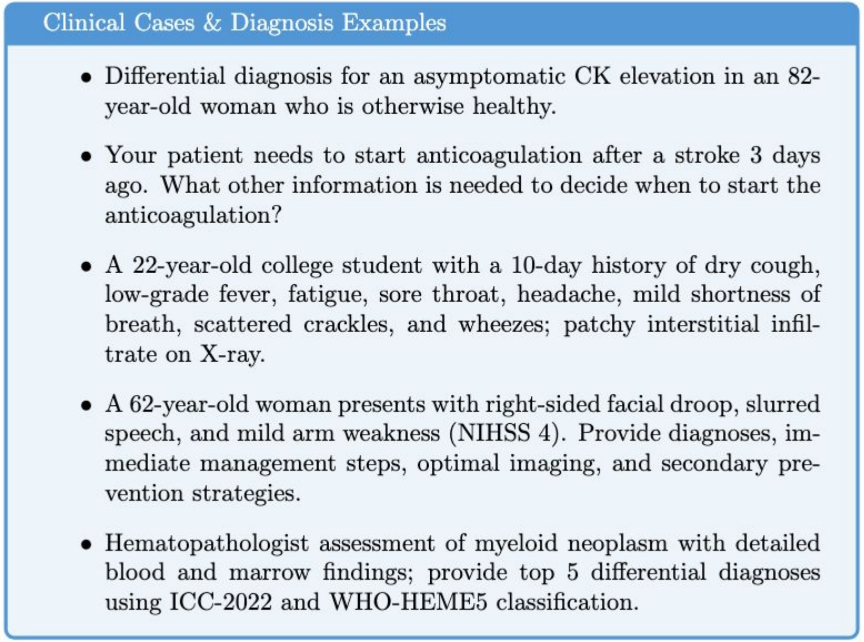
Treatment & Guidelines Examples

- For a patient with type 2 diabetes and recurrent hypoglycemia, what insulin regimen adjustments would you suggest?
- Patient on ceftriaxone and doxycycline for CAP; suggest oral antibiotics for discharge.
- Duration of dual antiplatelet therapy after left coronary artery stenting?
- For a patient with type 2 diabetes and recurrent hypoglycemia, what insulin regimen adjustments would you suggest?
- What is the most updated guideline for treating insomnia?

Asking the model to suggest treatments and provide clinical guidelines

What kinds of questions are clinicians asking?

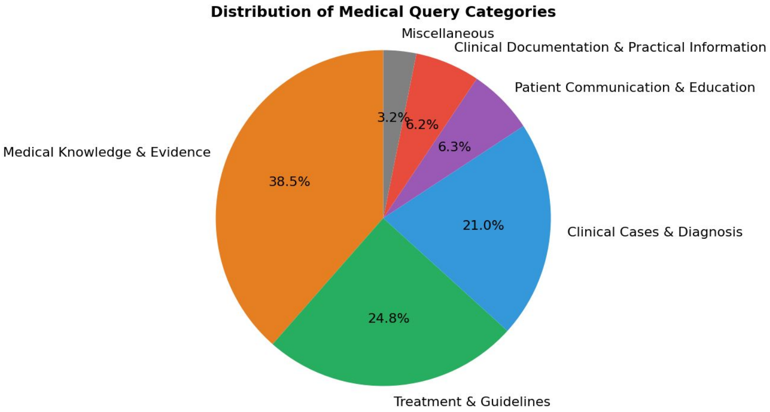


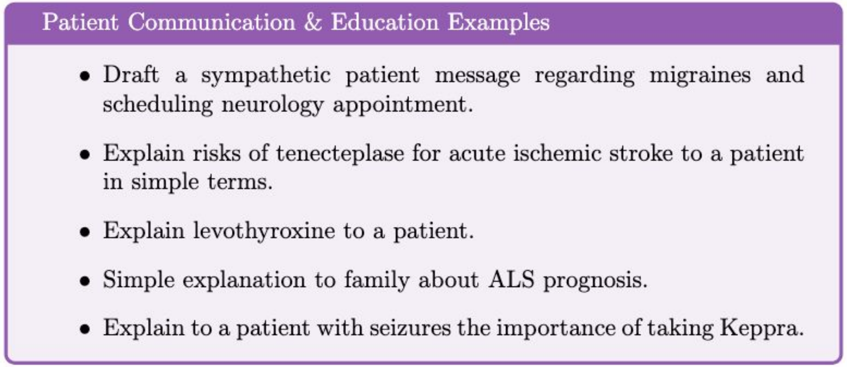


Clinical Cases & Diagnosis Examples

- Differential diagnosis for an asymptomatic CK elevation in an 82-year-old woman who is otherwise healthy.
- Your patient needs to start anticoagulation after a stroke 3 days ago. What other information is needed to decide when to start the anticoagulation?
- A 22-year-old college student with a 10-day history of dry cough, low-grade fever, fatigue, sore throat, headache, mild shortness of breath, scattered crackles, and wheezes; patchy interstitial infiltrate on X-ray.
- A 62-year-old woman presents with right-sided facial droop, slurred speech, and mild arm weakness (NIHSS 4). Provide diagnoses, immediate management steps, optimal imaging, and secondary prevention strategies.
- Hematopathologist assessment of myeloid neoplasm with detailed blood and marrow findings; provide top 5 differential diagnoses using ICC-2022 and WHO-HEME5 classification.

What kinds of questions are clinicians asking?



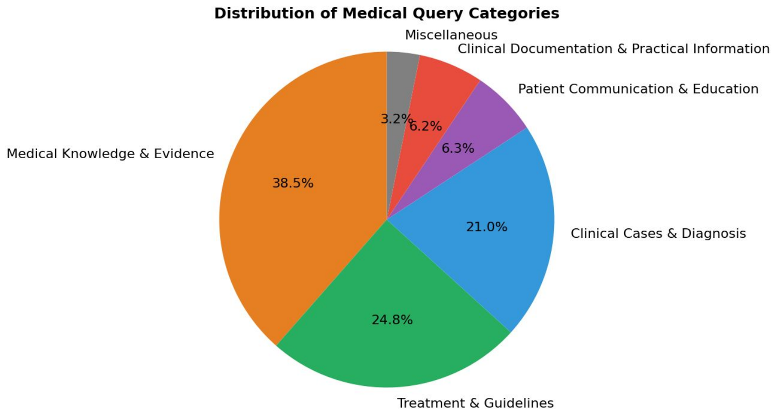


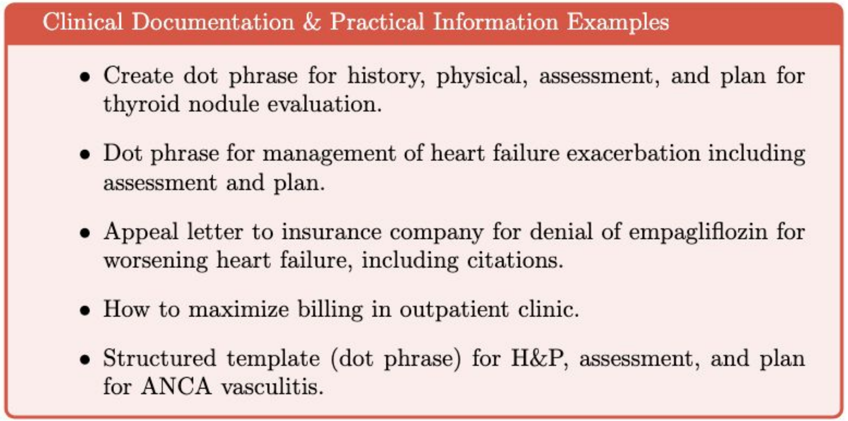
Patient Communication & Education Examples

- Draft a sympathetic patient message regarding migraines and scheduling neurology appointment.
- Explain risks of tenecteplase for acute ischemic stroke to a patient in simple terms.
- Explain levothyroxine to a patient.
- Simple explanation to family about ALS prognosis.
- Explain to a patient with seizures the importance of taking Keppra.

Asking the model to provide patient-friendly explanations

What kinds of questions are clinicians asking?



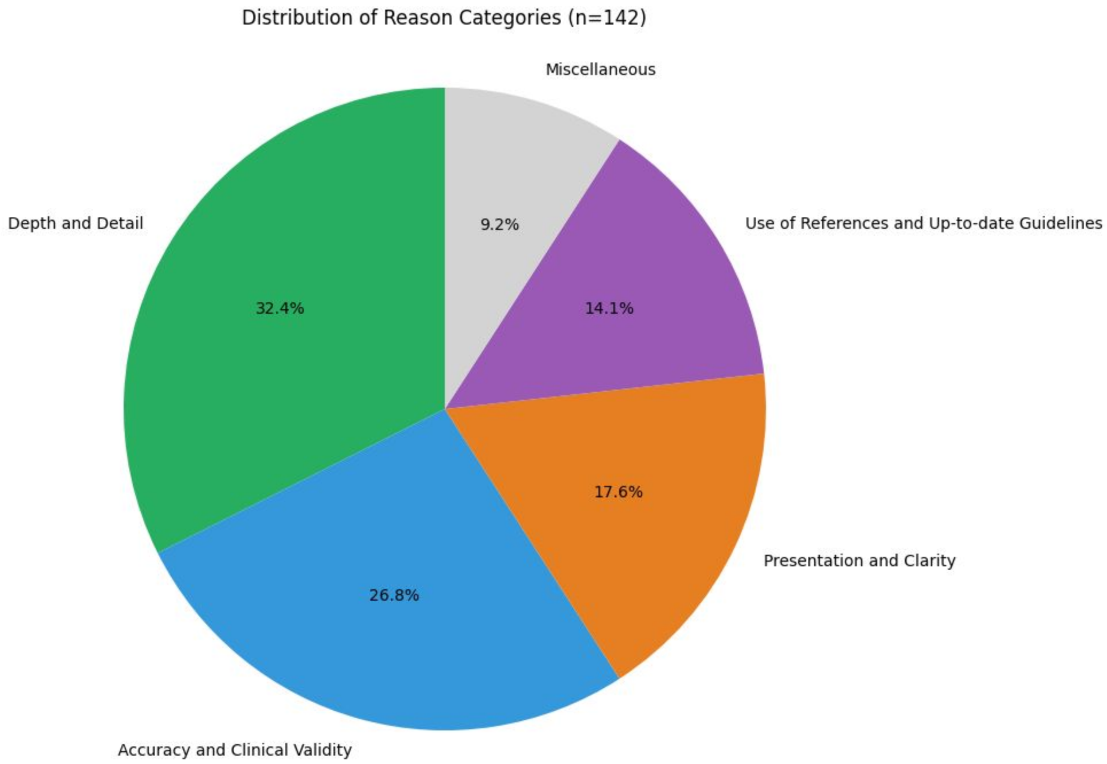


Clinical Documentation & Practical Information Examples

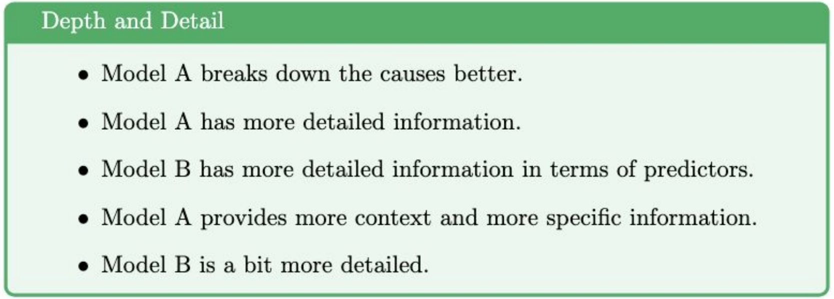
- Create dot phrase for history, physical, assessment, and plan for thyroid nodule evaluation.
- Dot phrase for management of heart failure exacerbation including assessment and plan.
- Appeal letter to insurance company for denial of empagliflozin for worsening heart failure, including citations.
- How to maximize billing in outpatient clinic.
- Structured template (dot phrase) for H&P, assessment, and plan for ANCA vasculitis.

Asking the model to help with note writing

Why do clinicians prefer model A vs B?

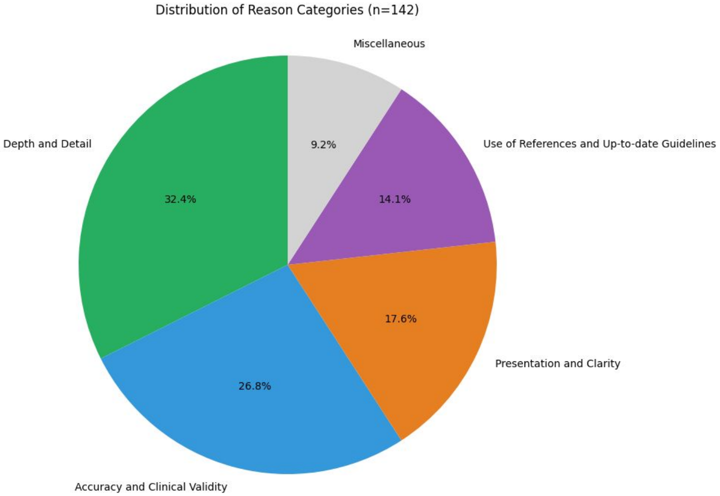


Why do clinicians prefer model A vs B?

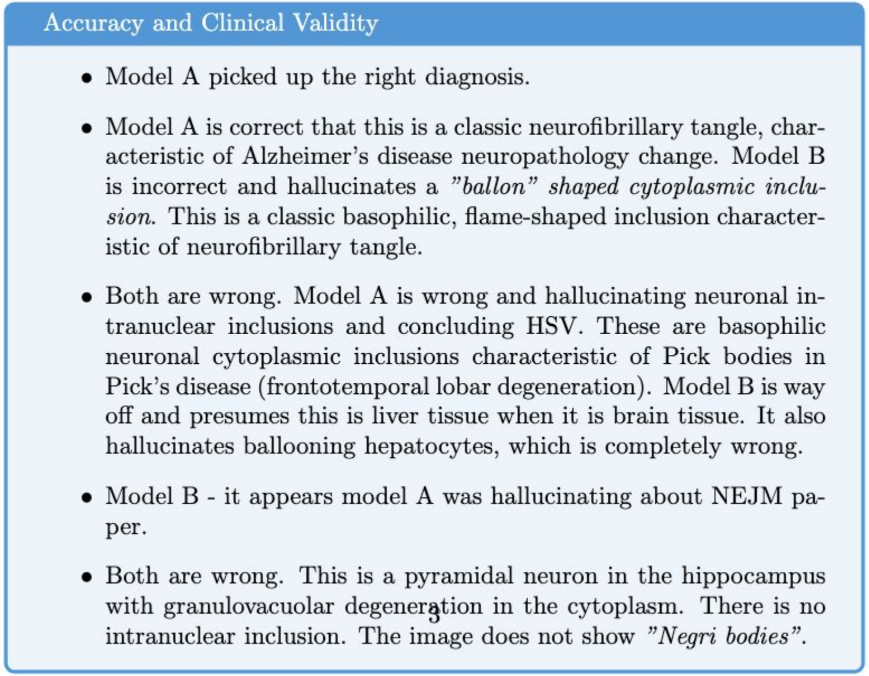


Depth and Detail

- Model A breaks down the causes better.
- Model A has more detailed information.
- Model B has more detailed information in terms of predictors.
- Model A provides more context and more specific information.
- Model B is a bit more detailed.

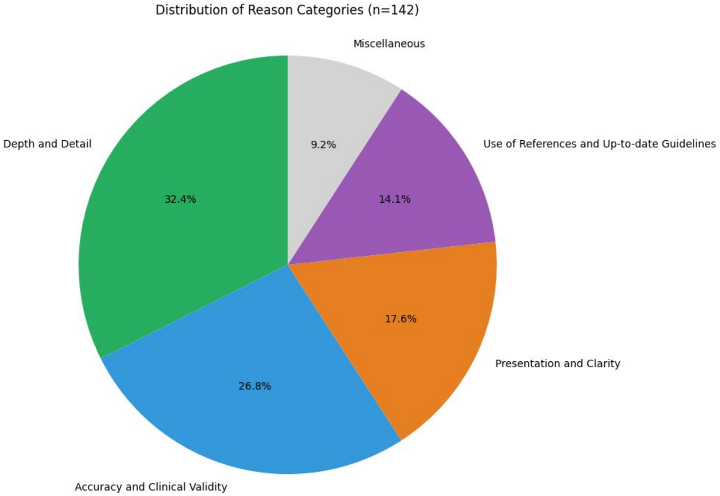


Why do clinicians prefer model A vs B?

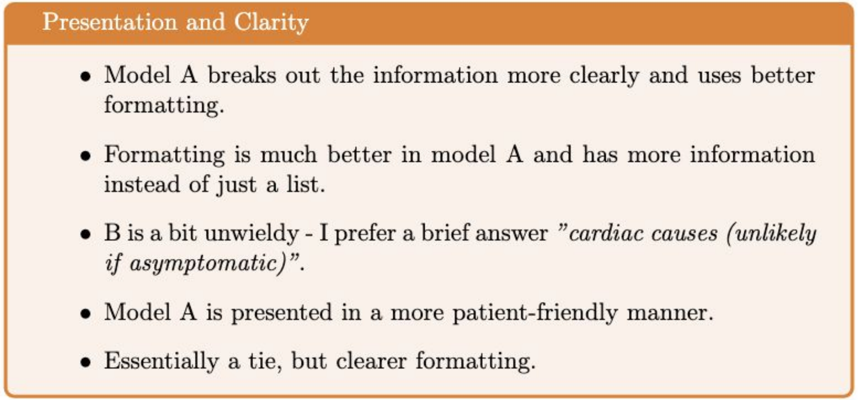


Accuracy and Clinical Validity

- Model A picked up the right diagnosis.
- Model A is correct that this is a classic neurofibrillary tangle, characteristic of Alzheimer’s disease neuropathology change. Model B is incorrect and hallucinates a *”ballon” shaped cytoplasmic inclusion*. This is a classic basophilic, flame-shaped inclusion characteristic of neurofibrillary tangle.
- Both are wrong. Model A is wrong and hallucinating neuronal intranuclear inclusions and concluding HSV. These are basophilic neuronal cytoplasmic inclusions characteristic of Pick bodies in Pick’s disease (frontotemporal lobar degeneration). Model B is way off and presumes this is liver tissue when it is brain tissue. It also hallucinates ballooning hepatocytes, which is completely wrong.
- Model B - it appears model A was hallucinating about NEJM paper.
- Both are wrong. This is a pyramidal neuron in the hippocampus with granulovacuolar degeneration in the cytoplasm. There is no intranuclear inclusion. The image does not show *”Negri bodies”*.

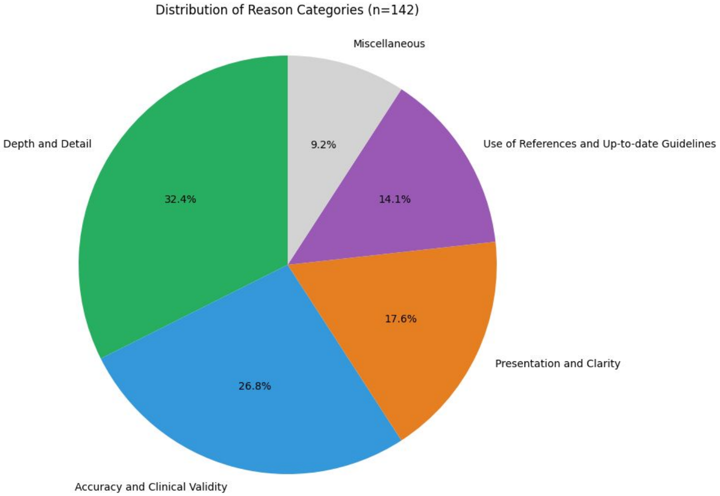


Why do clinicians prefer model A vs B?

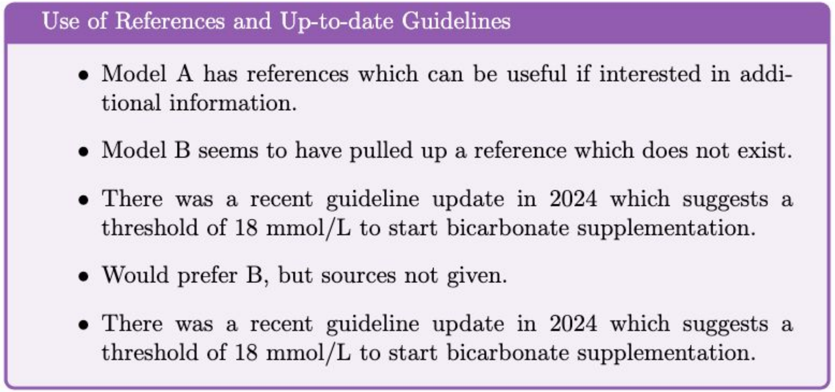


Presentation and Clarity

- Model A breaks out the information more clearly and uses better formatting.
- Formatting is much better in model A and has more information instead of just a list.
- B is a bit unwieldy - I prefer a brief answer *"cardiac causes (unlikely if asymptomatic)"*.
- Model A is presented in a more patient-friendly manner.
- Essentially a tie, but clearer formatting.

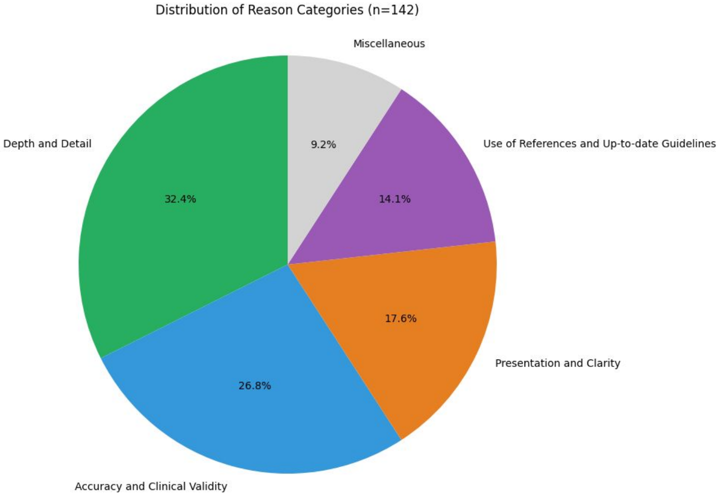


Why do clinicians prefer model A vs B?

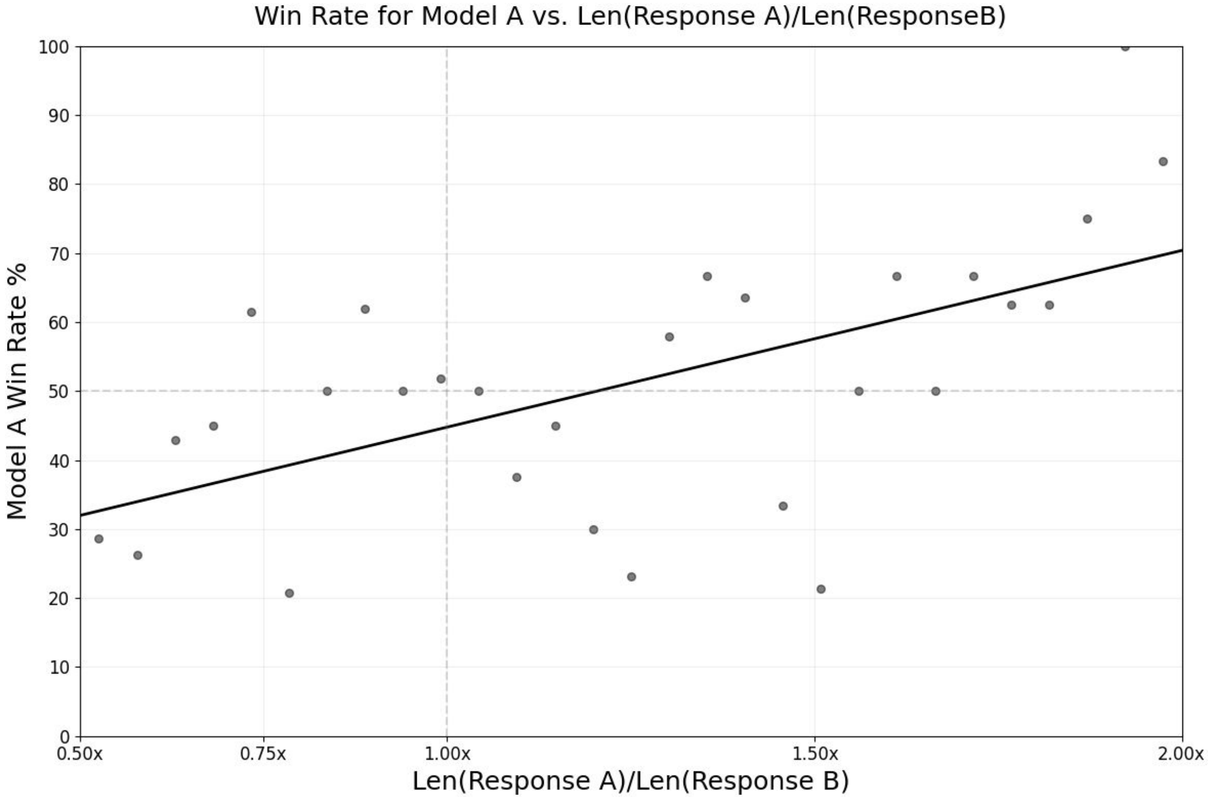


Use of References and Up-to-date Guidelines

- Model A has references which can be useful if interested in additional information.
- Model B seems to have pulled up a reference which does not exist.
- There was a recent guideline update in 2024 which suggests a threshold of 18 mmol/L to start bicarbonate supplementation.
- Would prefer B, but sources not given.
- There was a recent guideline update in 2024 which suggests a threshold of 18 mmol/L to start bicarbonate supplementation.



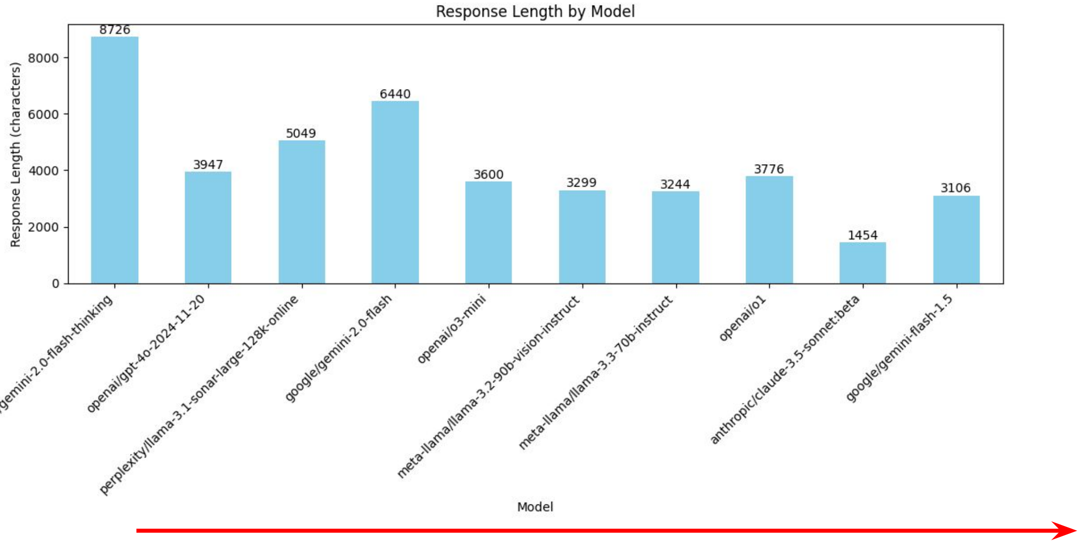
Models with longer responses win more



Longer responses are preferred

Median response length (in characters):

- Preferred model (4385) vs Not preferred (3725)
  - ***very significant

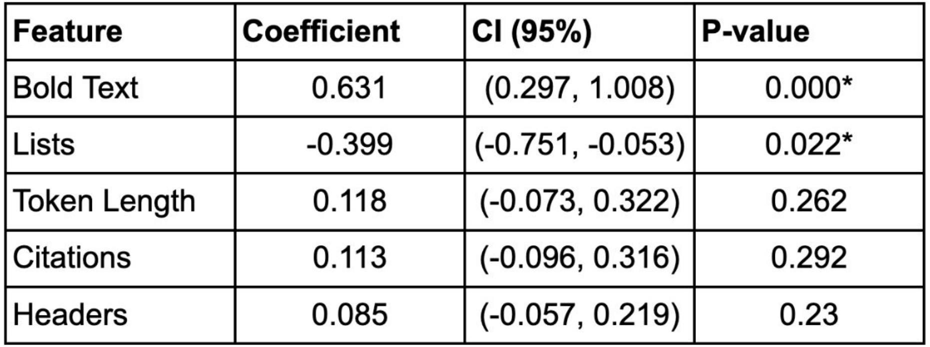


Ranking

Bradley-Terry model to control for style

Bradley-Terry model: *logistic regression* where input features are pairwise comparisons between models:

- +1 for winner, -1 for loser, 0 otherwise
- Coefficients are model scores
- Scaled to 1000 base rating



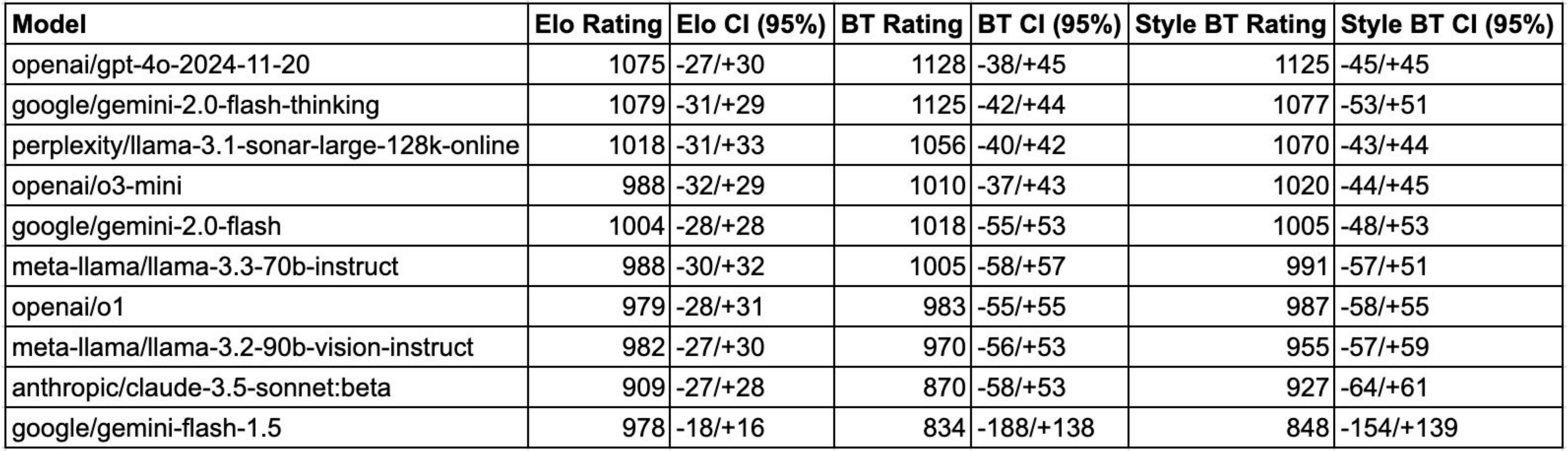
Feature	Coefficient	CI (95%)	P-value
Bold Text	0.631	(0.297, 1.008)	0.000*
Lists	-0.399	(-0.751, -0.053)	0.022*
Token Length	0.118	(-0.073, 0.322)	0.262
Citations	0.113	(-0.096, 0.316)	0.292
Headers	0.085	(-0.057, 0.219)	0.23

$$\Pr(i > j) = \frac{p_i}{p_i + p_j}$$

Style matters:

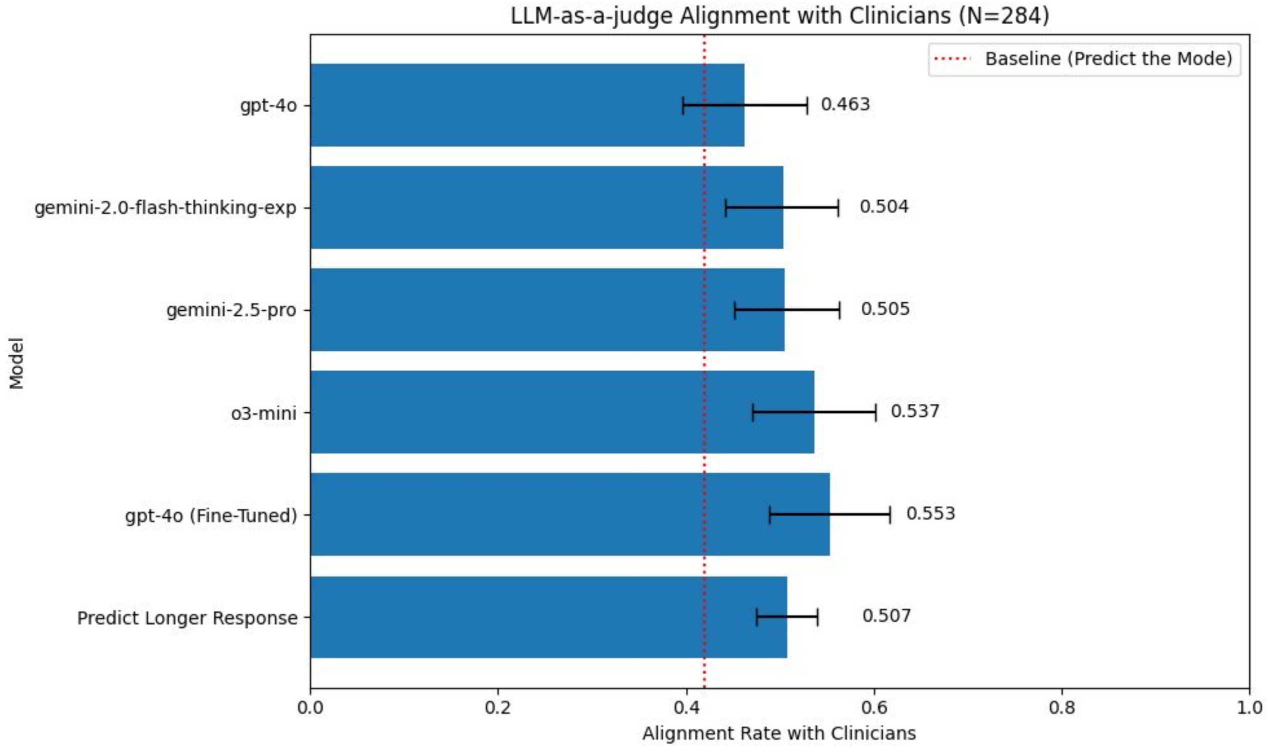
- **Bold text** and **lists** are significant confounders, **response length** less so

BT ratings (with and w/o style control)



Model	Elo Rating	Elo CI (95%)	BT Rating	BT CI (95%)	Style BT Rating	Style BT CI (95%)
openai/gpt-4o-2024-11-20	1075	-27/+30	1128	-38/+45	1125	-45/+45
google/gemini-2.0-flash-thinking	1079	-31/+29	1125	-42/+44	1077	-53/+51
perplexity/llama-3.1-sonar-large-128k-online	1018	-31/+33	1056	-40/+42	1070	-43/+44
openai/o3-mini	988	-32/+29	1010	-37/+43	1020	-44/+45
google/gemini-2.0-flash	1004	-28/+28	1018	-55/+53	1005	-48/+53
meta-llama/llama-3.3-70b-instruct	988	-30/+32	1005	-58/+57	991	-57/+51
openai/o1	979	-28/+31	983	-55/+55	987	-58/+55
meta-llama/llama-3.2-90b-vision-instruct	982	-27/+30	970	-56/+53	955	-57/+59
anthropic/claude-3.5-sonnet:beta	909	-27/+28	870	-58/+53	927	-64/+61
google/gemini-flash-1.5	978	-18/+16	834	-188/+138	848	-154/+139

Can LLMs predict clinician preferences?



Why do clinicians prefer Gemini over GPT-4o

What is the primary patency rate following stenting of the femoropopliteal artery?

Gemini Flash Thinking 2.0

- **Bare metal stents (BMS):** The 1-year primary patency rates range from **60% to 70%** in clinical studies, particularly in cases involving long and complex lesions.
- **Drug-eluting stents (DES):** These tend to show improved outcomes with 1-year primary patency rates of approximately **80% to 90%**, depending on the device and patient selection.
- **Drug-coated balloons (DCBs):** Although these are not technically stents, they have demonstrated 1-year patency rates of around **80%**, especially in shorter lesions.

<u>Gemini “summarizes like a clinician”</u>

- More concise and less vague
- More clinically actionable
- Exhibits greater domain knowledge

GPT-4o

The primary patency rate following stenting of the femoropopliteal artery is **variable** and depends on several factors, including:
* **Length and complexity of the treated lesion:** Longer and more complex lesions (e.g., TASC D lesions) generally have lower patency rates.
* **Severity of peripheral artery disease (PAD):** Patients with more severe PAD and poorer runoff vessels tend to have lower patency.
* **Patient characteristics:** Factors like diabetes, smoking, chronic kidney disease, and hyperlipidemia can negatively impact patency.
* **Type of stent used:** Drug-eluting stents (DES) generally demonstrate improved patency compared to bare metal stents (BMS), especially in longer and more complex lesions.
* **Adjunctive therapies:** Using techniques like atherectomy…

…… Truncated for Brevity ……

Why do clinicians prefer GPT-4o over Gemini

Where do watershed strokes typically occur in the brain?

Gemini

1. **Cortical Watershed Zones**: These occur between the territories of the anterior cerebral artery (ACA) and the middle cerebral artery (MCA), or between the MCA and the posterior cerebral artery (PCA). Cortical watershed strokes are commonly found at the edges of these territories in the cerebral cortex.

GPT-4o "Reasons like a medical student"

- More thorough
- More comprehensive
- Despite being vague and verbose, offers more rationale

GPT-4o

1. Cortical Watershed Zones: These are the most frequently discussed and often clinically relevant watershed zones. They are located in the cerebral cortex at the junctions between the territories of:
* **Anterior Watershed Zone (ACA-MCA Watershed):** This is the area between the territories of the **Anterior Cerebral Artery (ACA)** and the **Middle Cerebral Artery (MCA)**. It's typically located in the **superior frontal and parietal lobes**, often affecting:
    * **Lateral frontal lobe posteriorly:** This region is involved in motor function, especially for the **proximal upper extremity (shoulder and upper arm)** and sometimes the **leg**.
    * **Superior parietal lobe:** Impacts sensory function and spatial awareness.
  * **Posterior Watershed Zone (MCA-PCA Watershed):** This zone lies between the territories of the **Middle Cerebral Artery (MCA)** and the **Posterior Cerebral Artery (PCA)**. It's found in the **occipito-parieto-temporal junction region**, typically affecting:

…… Truncated for Brevity ……

Thanks to NEJM AI, Doximity, Centaur Labs, and Eric Topol!

Thank You