🏥 MedArena.ai

A Clinician-Centric Platform for Expert Preferences on Medical Queries

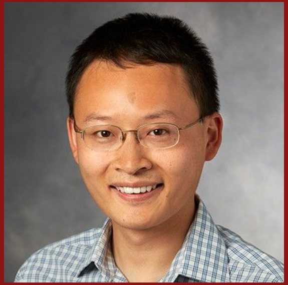
Eric Wu, Kevin Wu, James Zou
Stanford University
3/28/25

Stanford University

Stanford University

Eric Wu

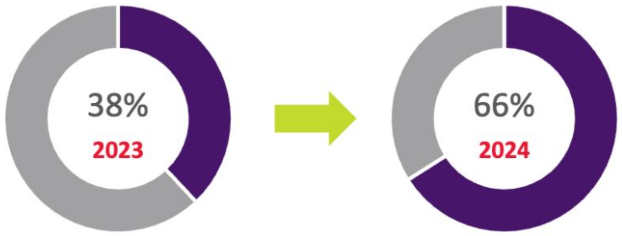
Kevin Wu



James Zou

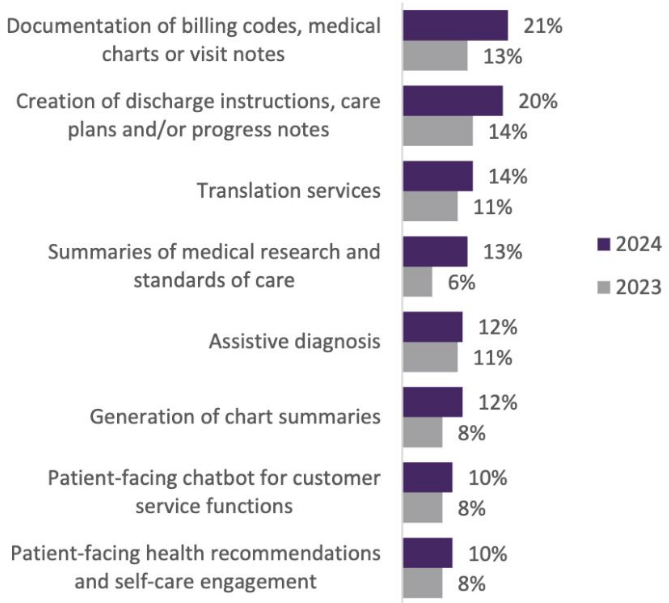
Two-thirds of clinicians report using AI tools

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

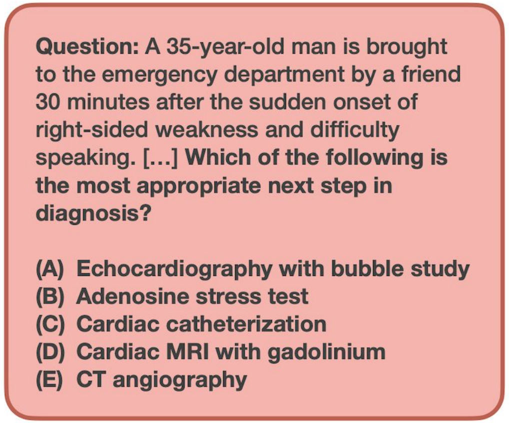
LLMs are evaluated on fixed multiple-choice datasets:

- 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

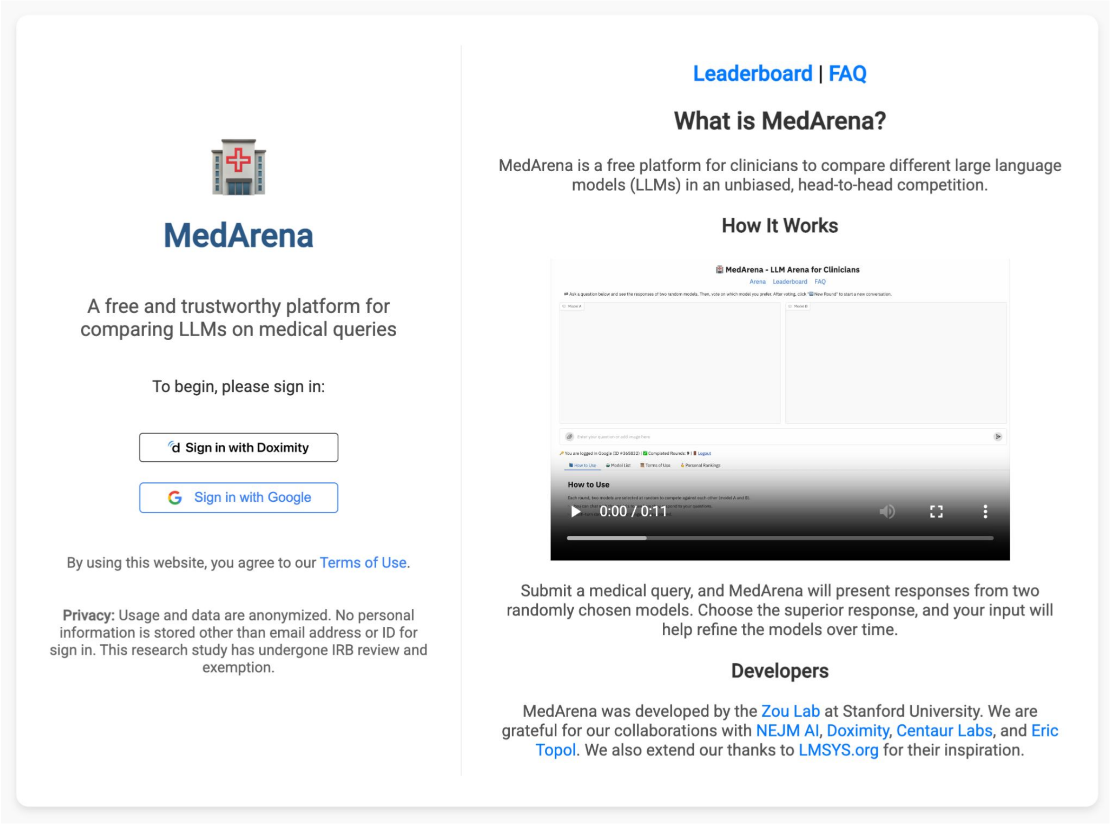
USMLE
United States Medical Licensing Exam

What questions do clinicians ask LLMs?

Which LLMs do clinicians prefer?

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.

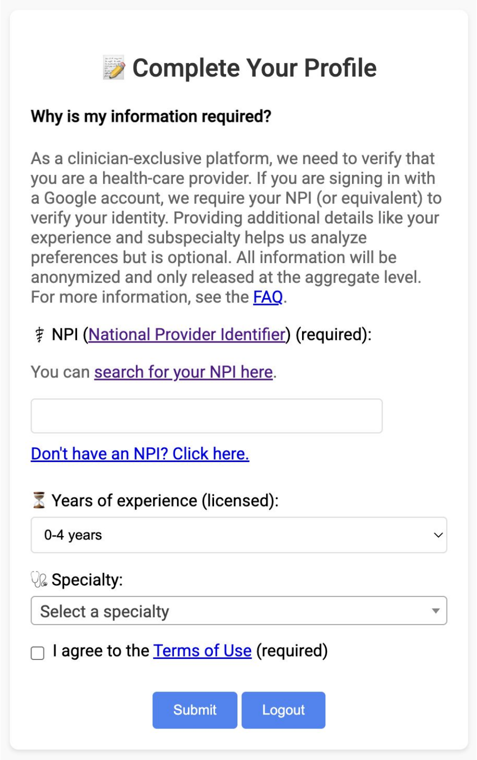
Clinician-only platform

- 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)

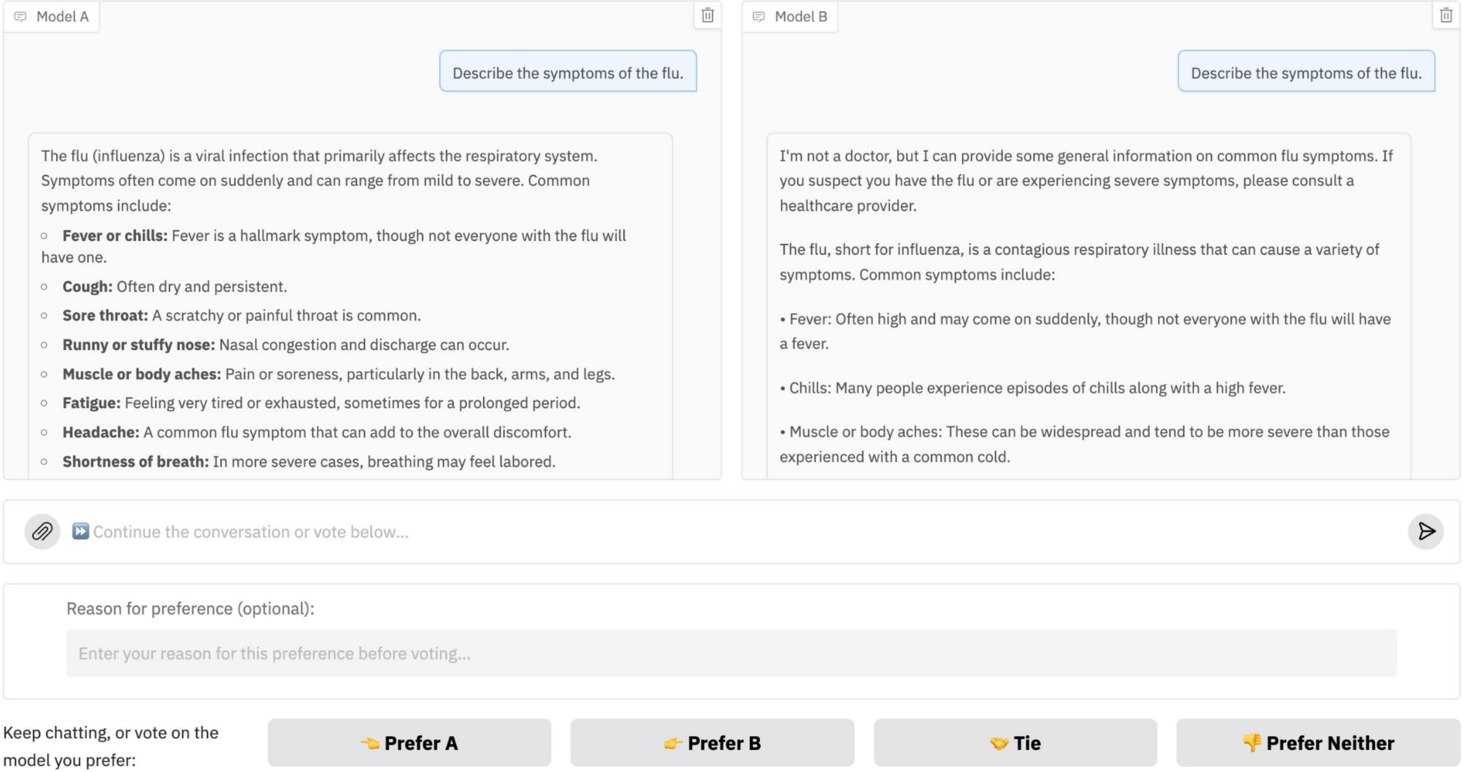
Information collected:

- 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.

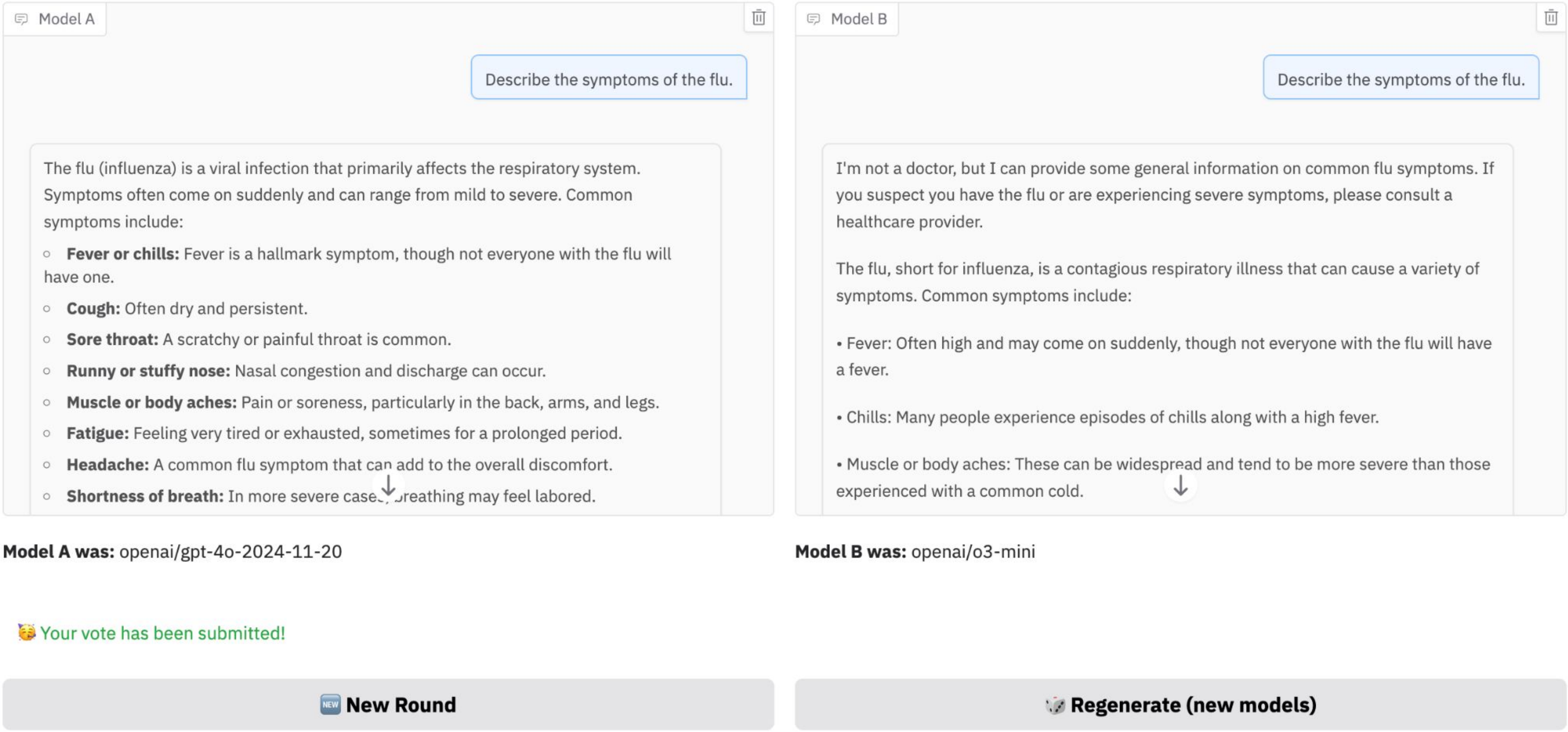
- Chills: Many people experience episodes of chills along with a high fever.

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

Reason for preference (optional):

Keep chatting, or vote on the model you prefer:

User submits a preference with optional reason



Model A

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.

Model B

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!

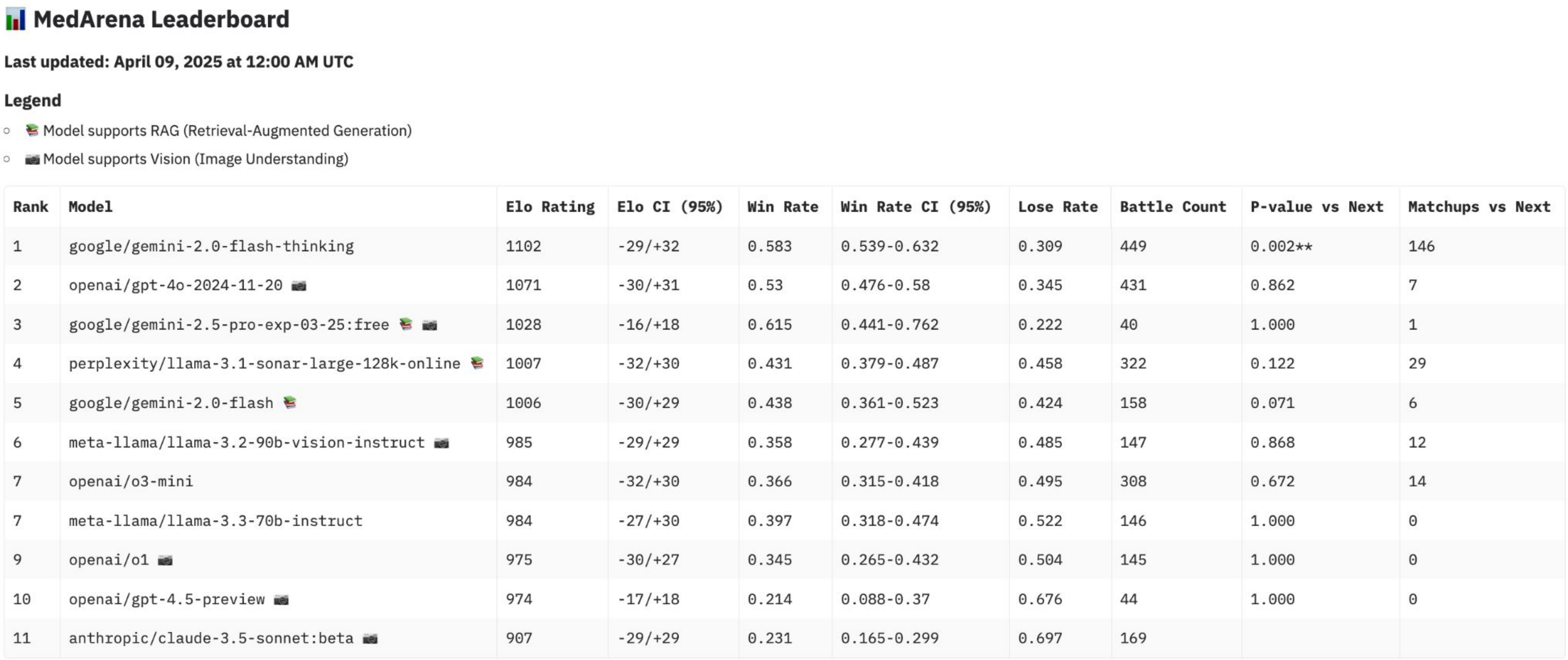
What questions do clinicians ask LLMs?

Fixed QA → Free-text

Which LLMs do clinicians prefer?

Preferences: rankings, reasons

Leaderboard based on clinician preferences



📊 MedArena Leaderboard

Last updated: April 09, 2025 at 12:00 AM UTC

Legend

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

Rank	Model	Elo Rating	Elo CI (95%)	Win Rate	Win Rate CI (95%)	Lose Rate	Battle Count	P-value vs Next	Matchups vs Next
1	google/gemini-2.0-flash-thinking	1102	-29/+32	0.583	0.539-0.632	0.309	449	0.002**	146
2	openai/gpt-4o-2024-11-20 📷	1071	-30/+31	0.53	0.476-0.58	0.345	431	0.862	7
3	google/gemini-2.5-pro-exp-03-25:free 📚 📷	1028	-16/+18	0.615	0.441-0.762	0.222	40	1.000	1
4	perplexity/llama-3.1-sonar-large-128k-online 📚	1007	-32/+30	0.431	0.379-0.487	0.458	322	0.122	29
5	google/gemini-2.0-flash 📚	1006	-30/+29	0.438	0.361-0.523	0.424	158	0.071	6
6	meta-llama/llama-3.2-90b-vision-instruct 📷	985	-29/+29	0.358	0.277-0.439	0.485	147	0.868	12
7	openai/o3-mini	984	-32/+30	0.366	0.315-0.418	0.495	308	0.672	14
7	meta-llama/llama-3.3-70b-instruct	984	-27/+30	0.397	0.318-0.474	0.522	146	1.000	0
9	openai/o1 📷	975	-30/+27	0.345	0.265-0.432	0.504	145	1.000	0
10	openai/gpt-4.5-preview 📷	974	-17/+18	0.214	0.088-0.37	0.676	44	1.000	0
11	anthropic/claude-3.5-sonnet:beta 📷	907	-29/+29	0.231	0.165-0.299	0.697	169		

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.

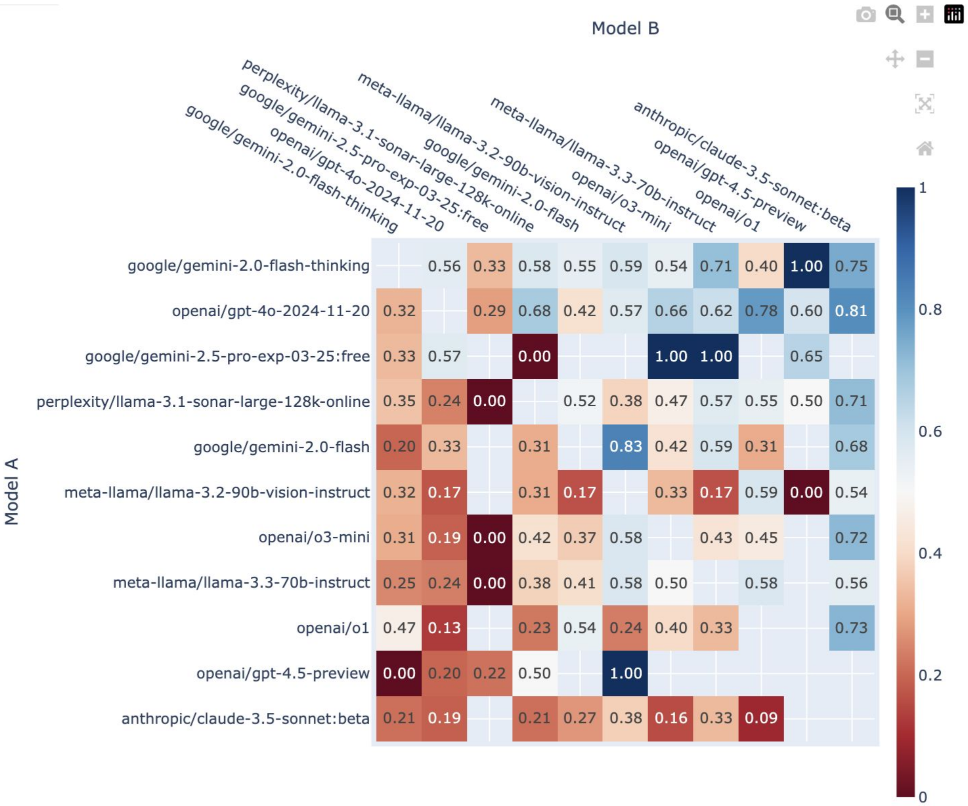
- Larger rating difference leads to larger rating change
- K=4 for stable ratings
- Base rating = 1000

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

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

Stanford University

Pairwise head-to-head matchups

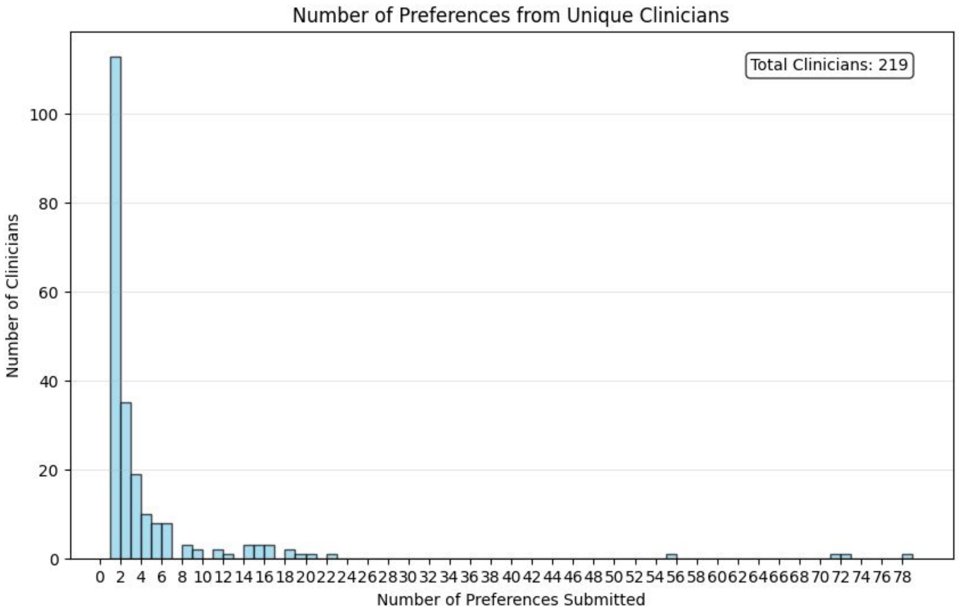


Model A \ Model B	google/gemini-2.0-flash-thinking	openai/gpt-4o-2024-11-20	google/gemini-2.5-pro-exp-03-25:free	perplexity/llama-3.1-sonar-large-128k-online	google/gemini-2.0-flash	meta-llama/llama-3.2-90b-vision-instruct	openai/o3-mini	meta-llama/llama-3.3-70b-instruct	openai/o1	openai/gpt-4.5-preview	anthropic/claude-3.5-sonnet:beta
google/gemini-2.0-flash-thinking		0.56	0.33	0.58	0.55	0.59	0.54	0.71	0.40	1.00	0.75
openai/gpt-4o-2024-11-20	0.32		0.29	0.68	0.42	0.57	0.66	0.62	0.78	0.60	0.81
google/gemini-2.5-pro-exp-03-25:free	0.33	0.57		0.00			1.00	1.00		0.65	
perplexity/llama-3.1-sonar-large-128k-online	0.35	0.24	0.00		0.52	0.38	0.47	0.57	0.55	0.50	0.71
google/gemini-2.0-flash	0.20	0.33		0.31		0.83	0.42	0.59	0.31		0.68
meta-llama/llama-3.2-90b-vision-instruct	0.32	0.17		0.31	0.17		0.33	0.17	0.59	0.00	0.54
openai/o3-mini	0.31	0.19	0.00	0.42	0.37	0.58		0.43	0.45		0.72
meta-llama/llama-3.3-70b-instruct	0.25	0.24	0.00	0.38	0.41	0.58	0.50		0.58		0.56
openai/o1	0.47	0.13		0.23	0.54	0.24	0.40	0.33			0.73
openai/gpt-4.5-preview	0.00	0.20	0.22	0.50		1.00					
anthropic/claude-3.5-sonnet:beta	0.21	0.19		0.21	0.27	0.38	0.16	0.33	0.09		

Pairwise Win Rates (Row vs Column)

Clinician Sign-Ups

- 290 clinicians
- ⅓ users from outside of US
- ~⅔ Google, ~⅓ Doximity
  - From Doximity, >80% MDs


Number of Preferences from Unique Clinicians

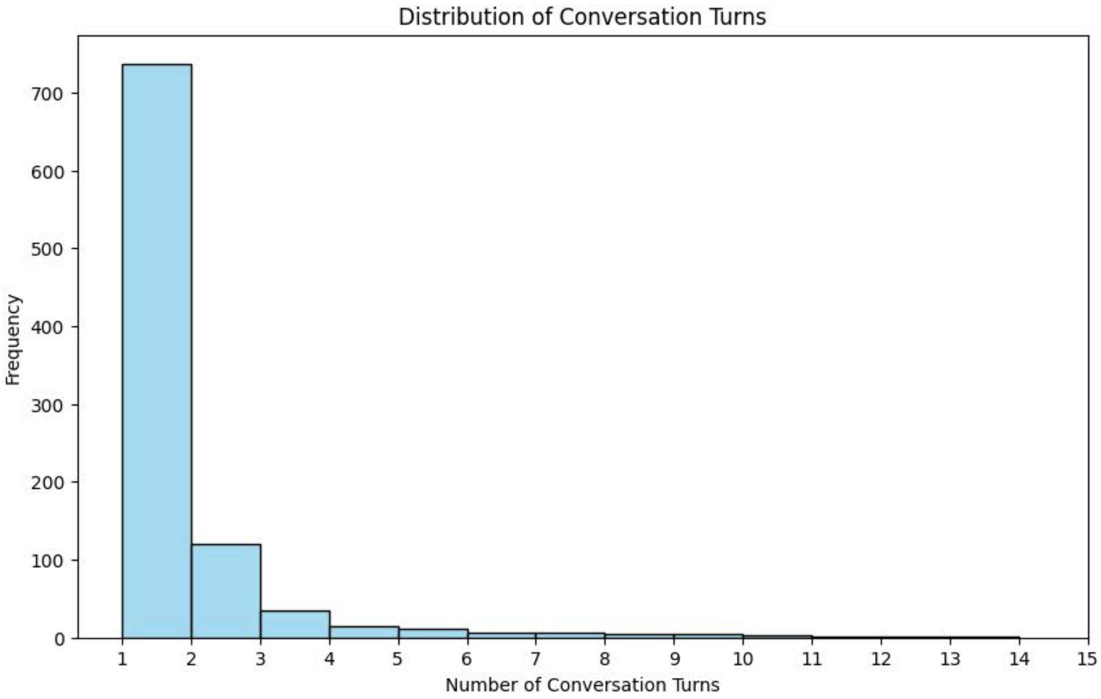
Clinician specialities

- 80 unique specialities

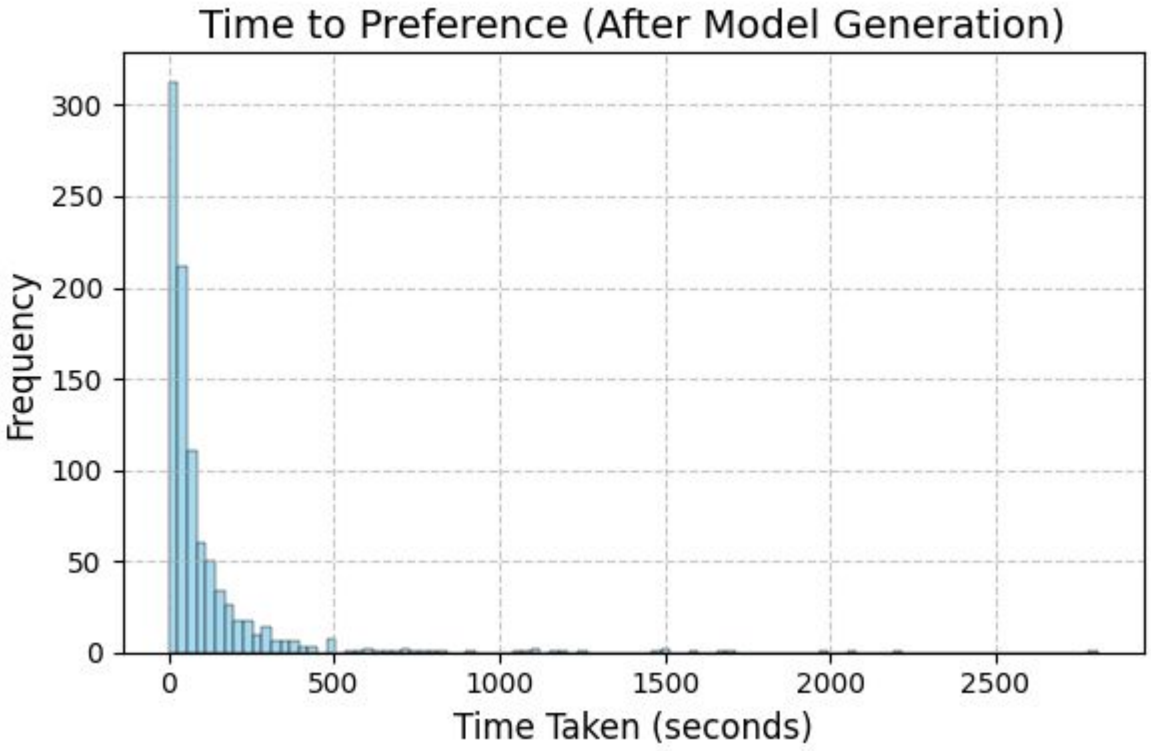
- Top 5 subspecialties (representing 35.6% of users):
    - Internal Medicine Physician: 31 (14.2%)
    - Family Medicine Physician: 17 (7.8%)
    - Multi-Specialty Group: 14 (6.4%)
    - General Practice Physician: 8 (3.7%)
    - Emergency Medicine Physician: 8 (3.7%)

Number of conversation turns

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

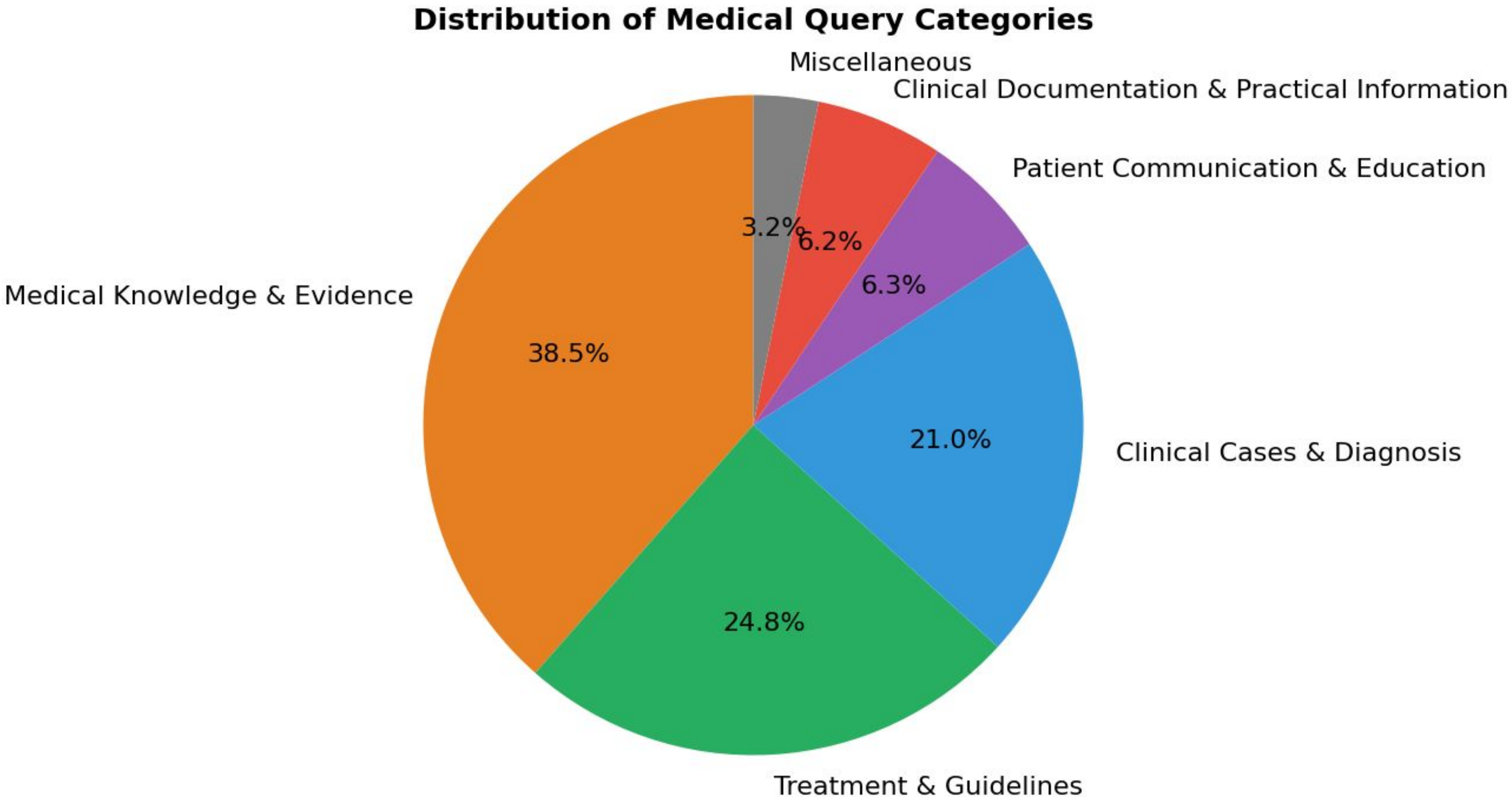


Median clinician response time is 47 seconds

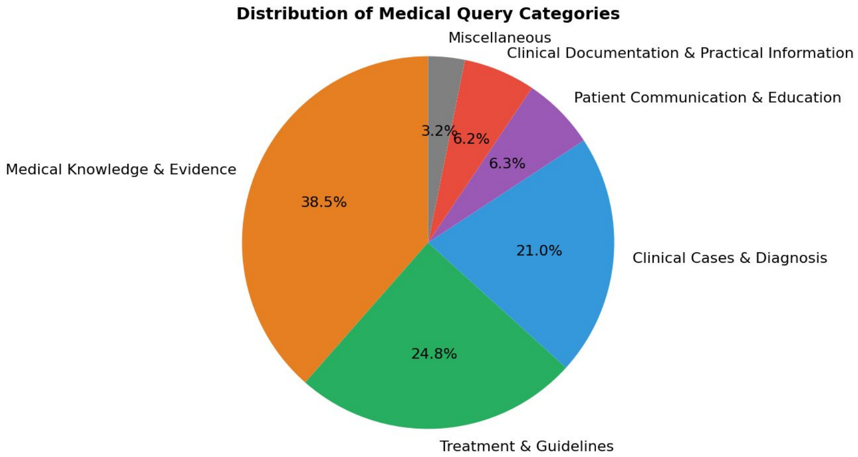


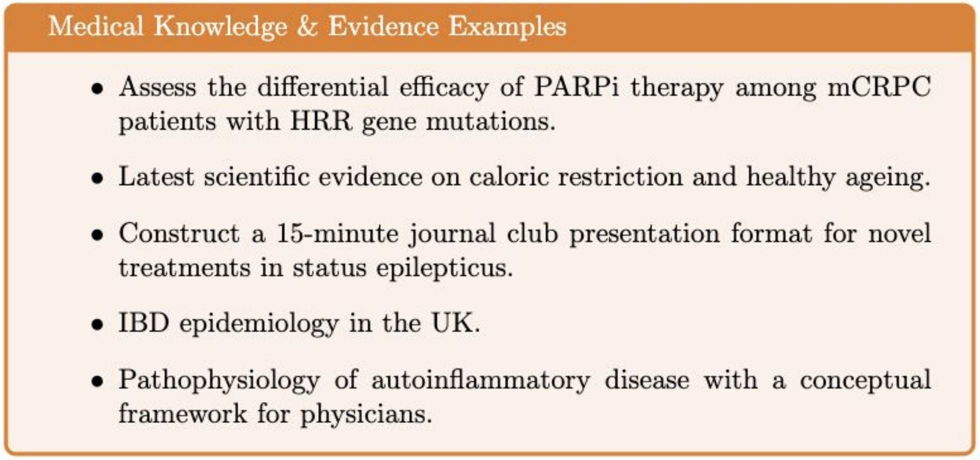
What kinds of questions are clinicians asking?

What kinds of questions are clinicians asking?



What kinds of questions are clinicians asking?



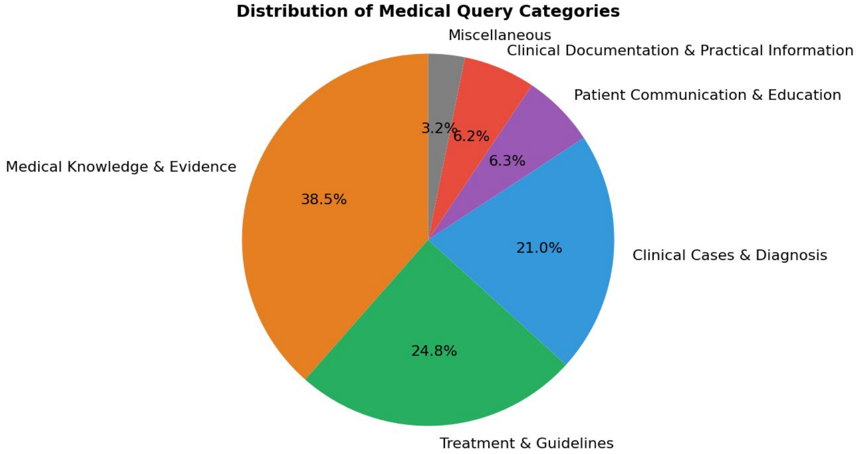


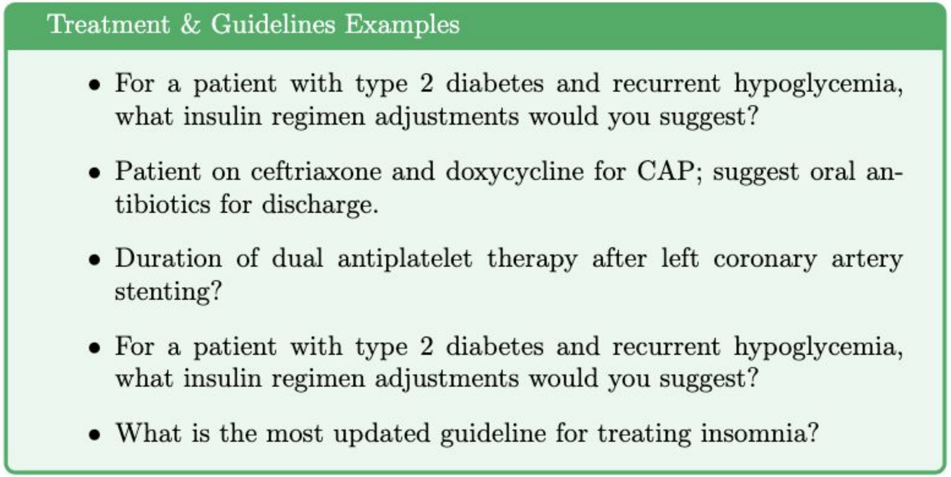
Medical Knowledge & Evidence Examples

- 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?



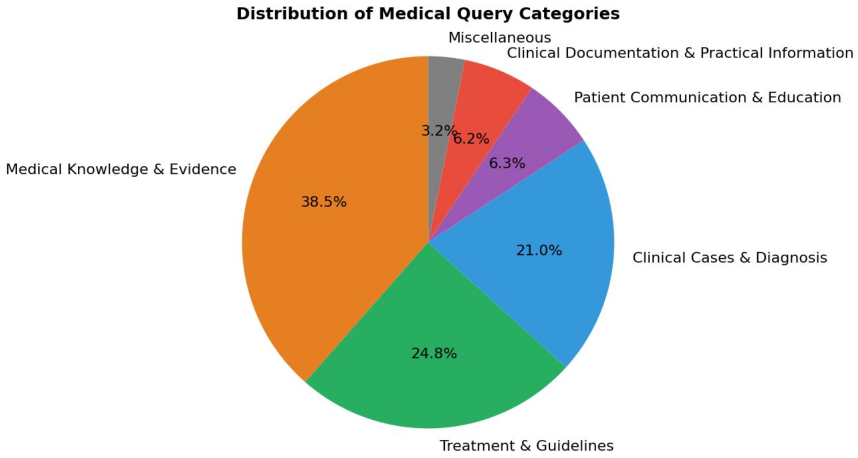


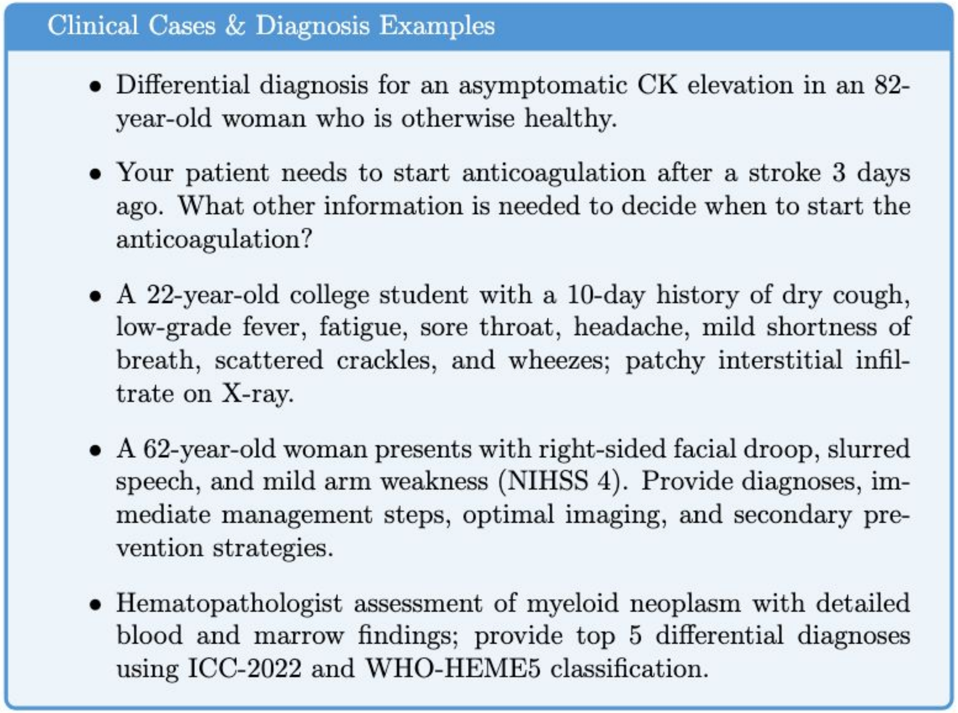
Treatment & Guidelines Examples

- 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?

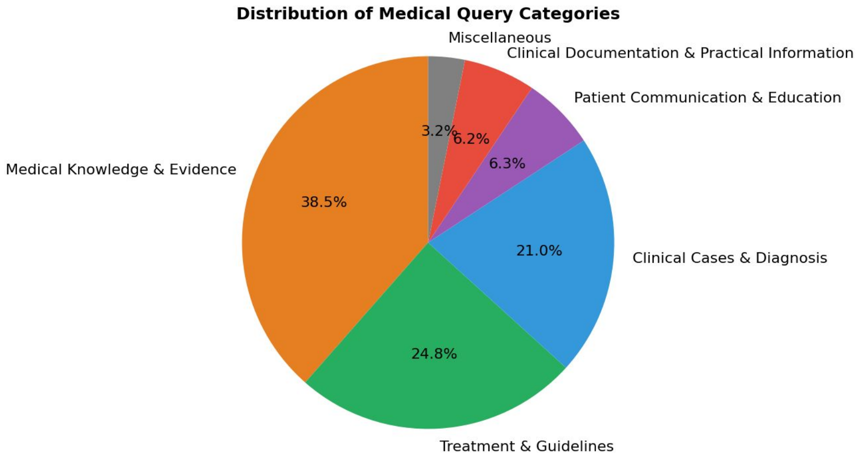


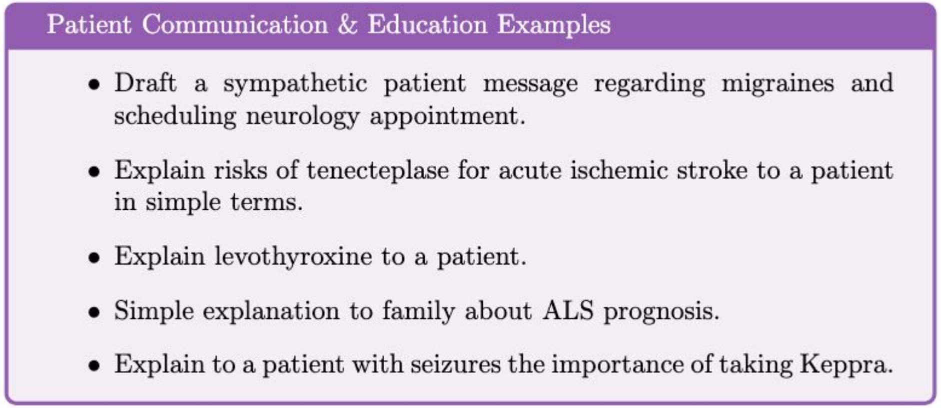


Clinical Cases & Diagnosis Examples

- 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?



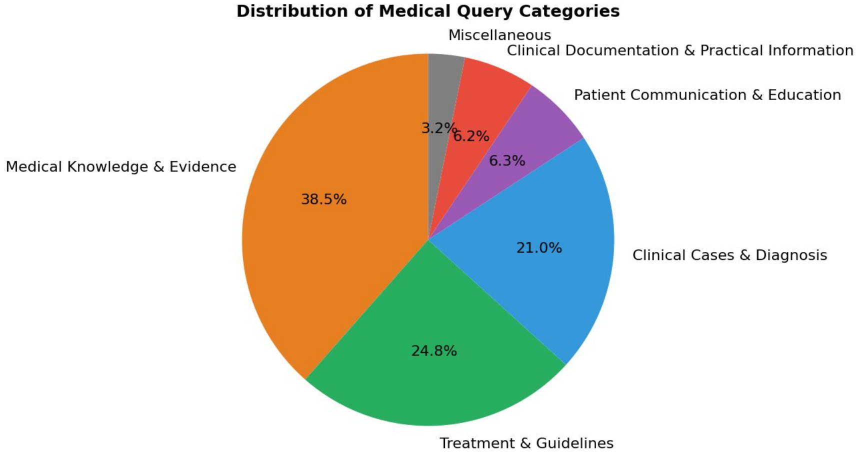


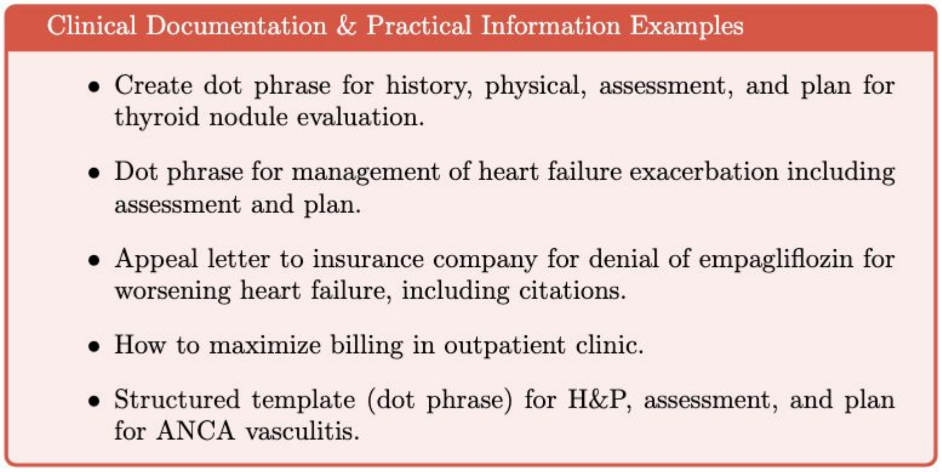
Patient Communication & Education Examples

- 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?



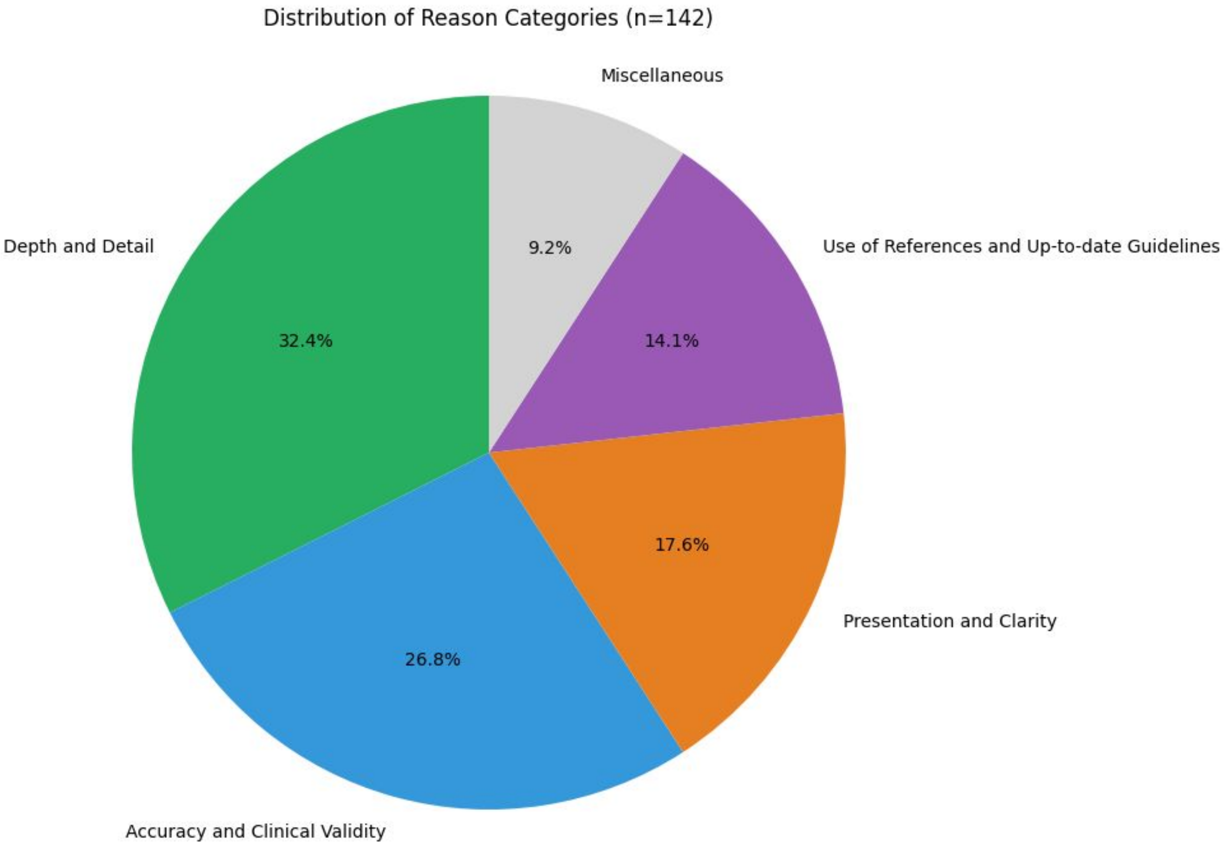


Clinical Documentation & Practical Information Examples

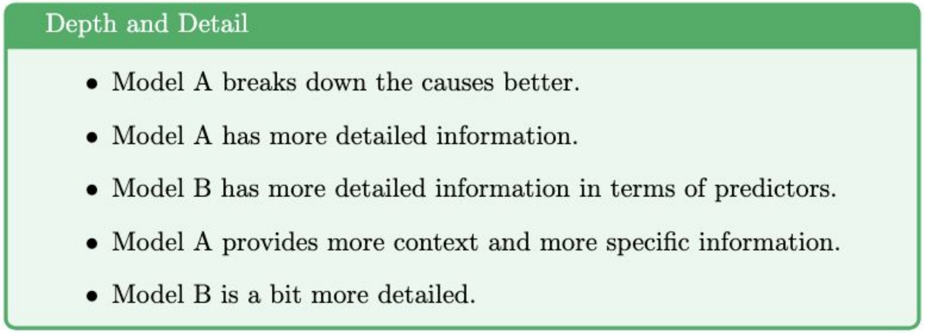
- Create dot phrase for history, physical, assessment, and plan for thyroid nodule evaluation.
- Dot phrase for management of heart failure exacerbation including assessment and plan.
- Appeal letter to insurance company for denial of empagliflozin for worsening heart failure, including citations.
- 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?

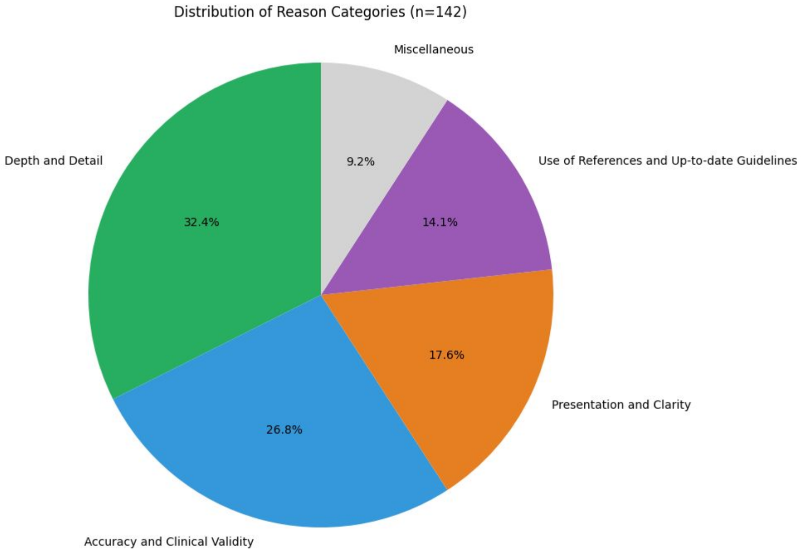


Why do clinicians prefer model A vs B?

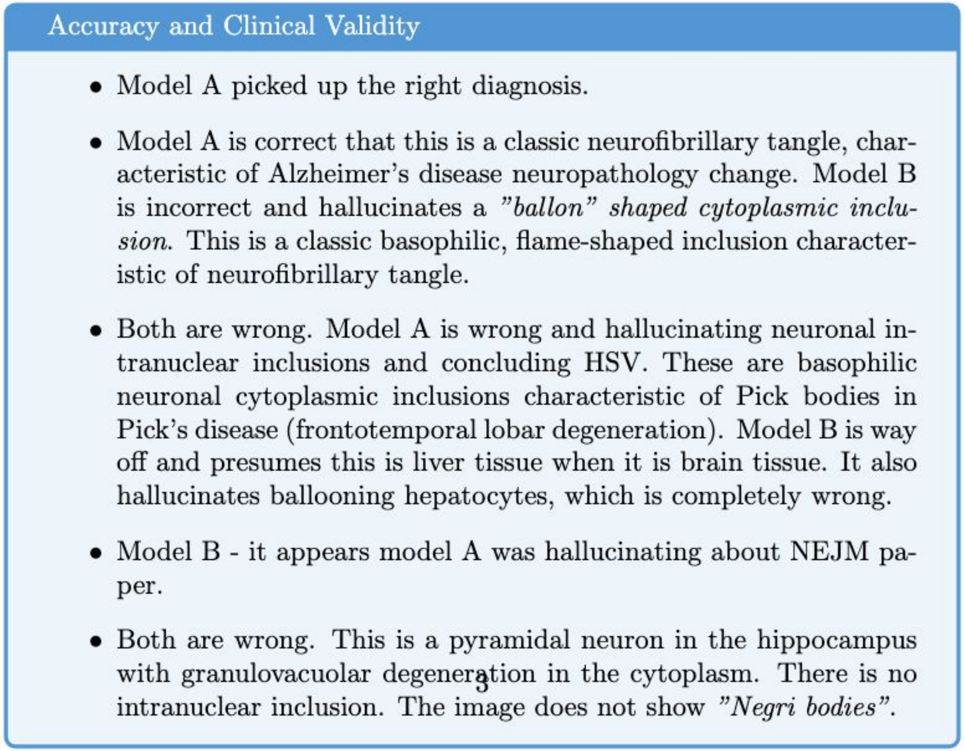


Depth and Detail

- 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.

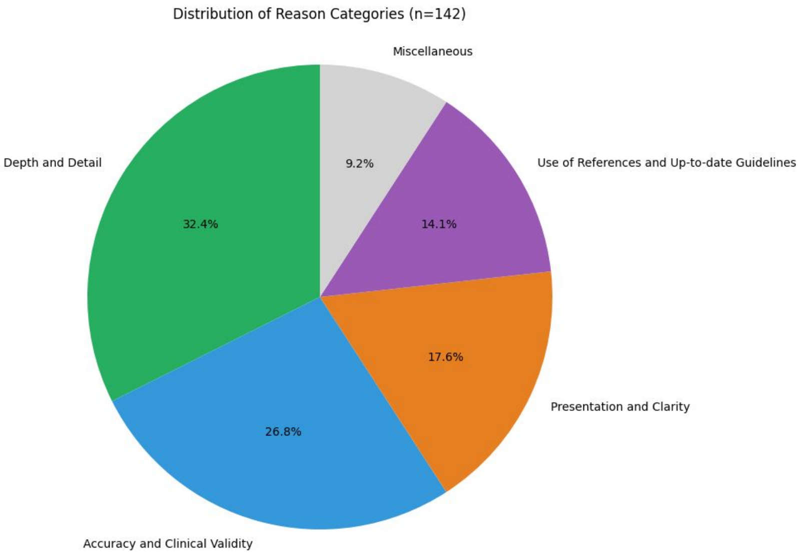


Why do clinicians prefer model A vs B?

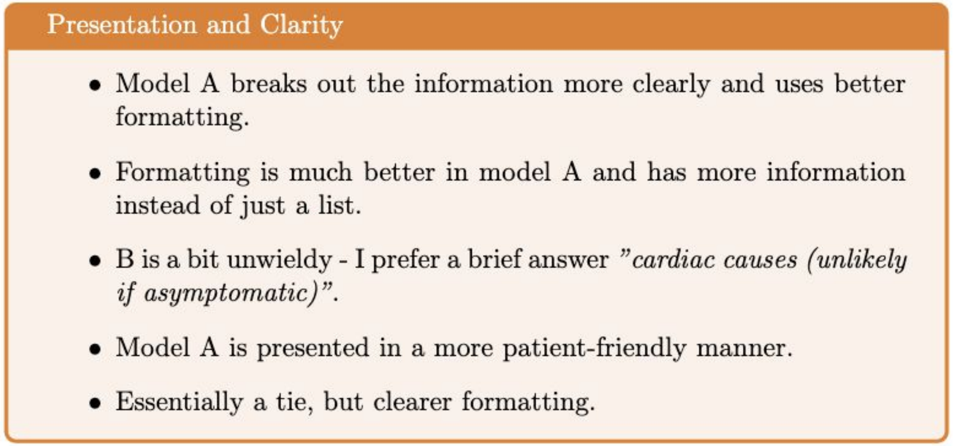


Accuracy and Clinical Validity

- 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"*.

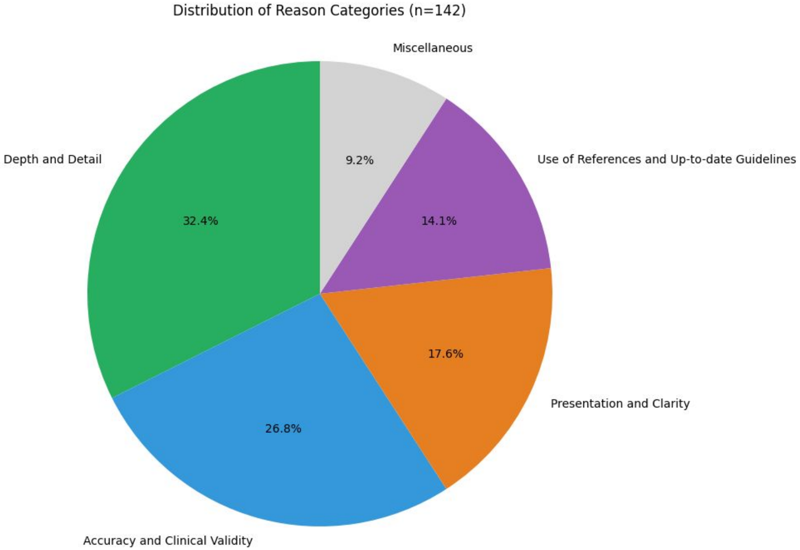


Why do clinicians prefer model A vs B?

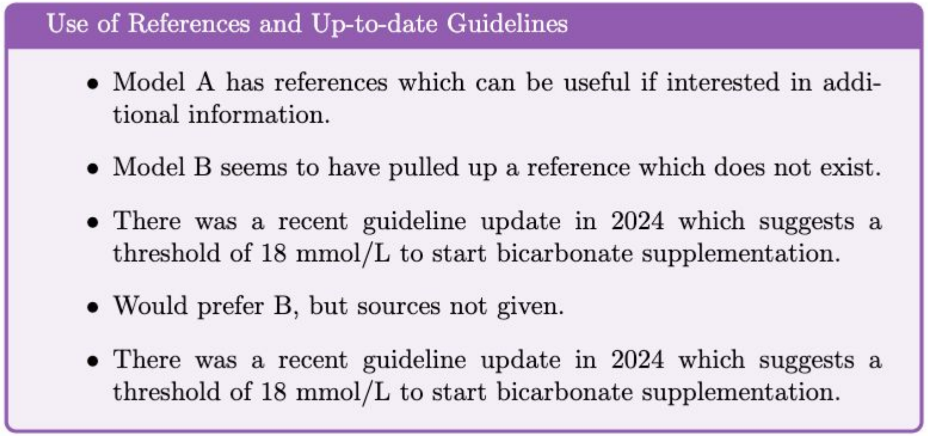


Presentation and Clarity

- 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.

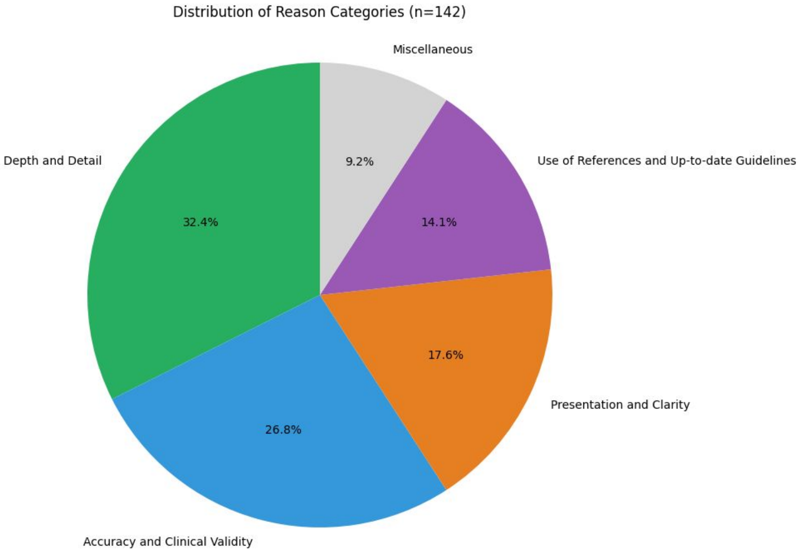


Why do clinicians prefer model A vs B?



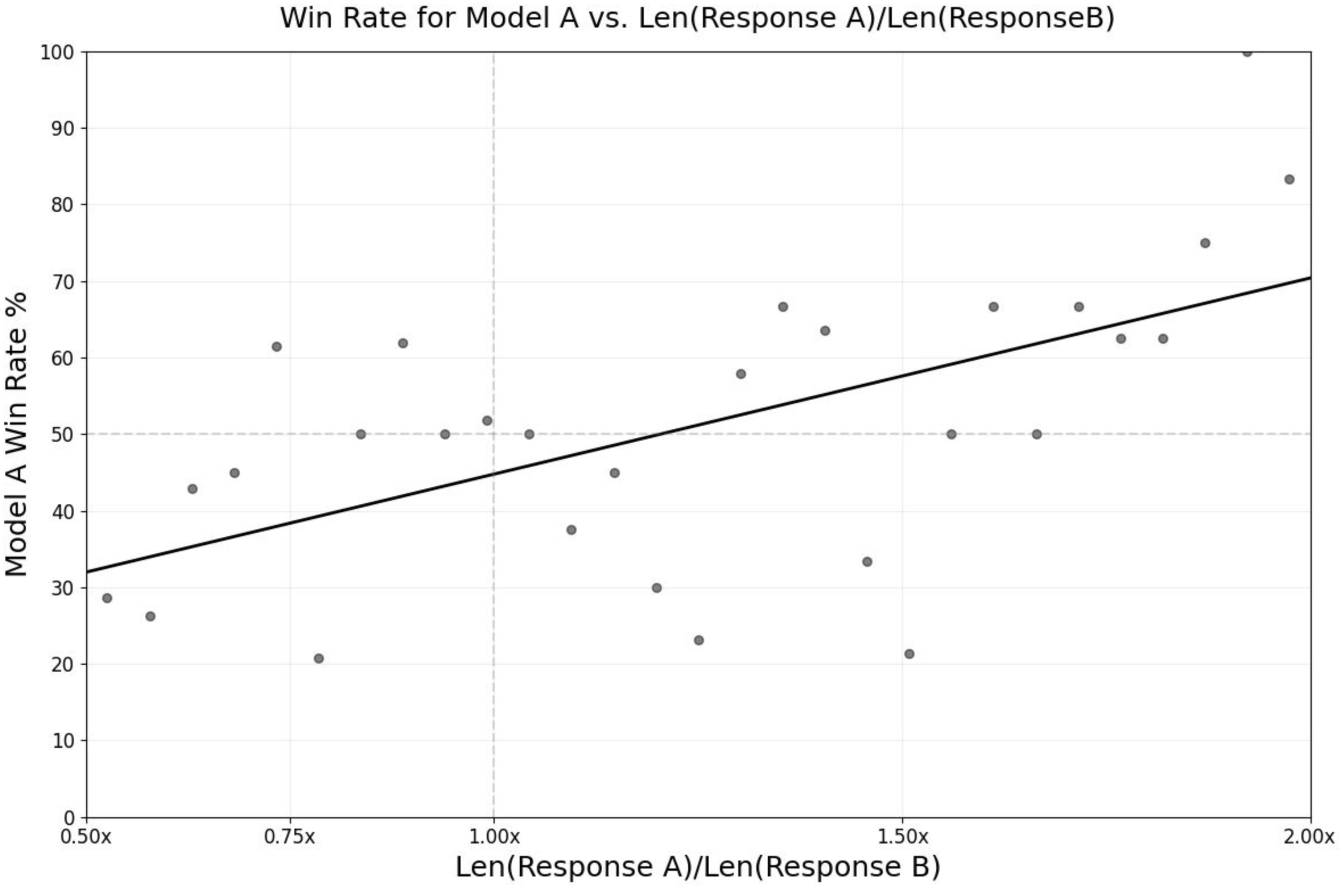
Use of References and Up-to-date Guidelines

- 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.



Distribution of Reason Categories (n=142)

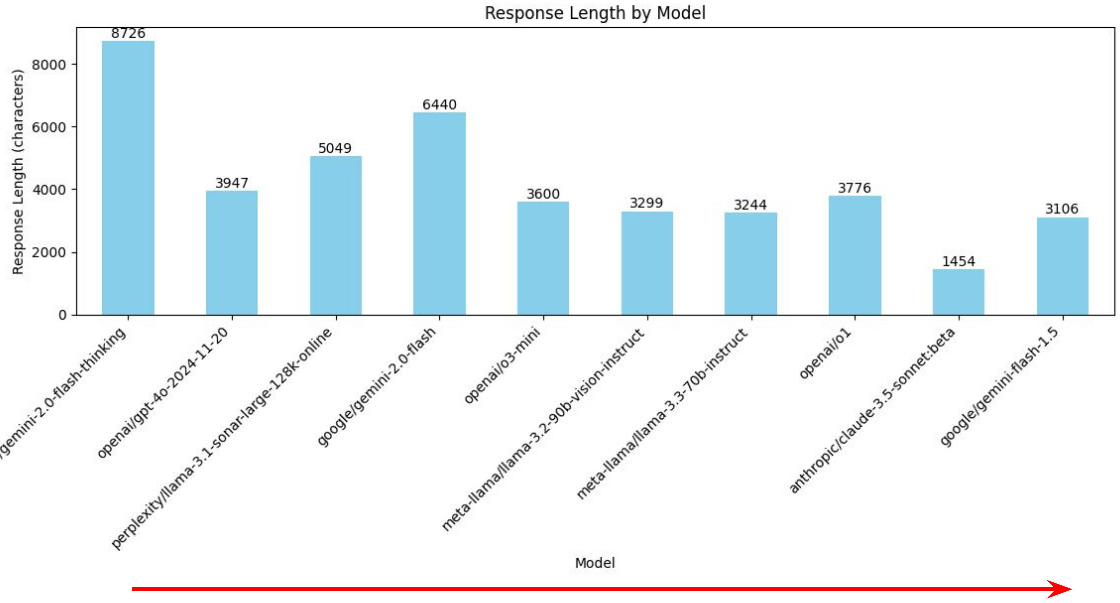
Models with longer responses win more



Longer responses are preferred

Median response length (in characters):

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



Ranking

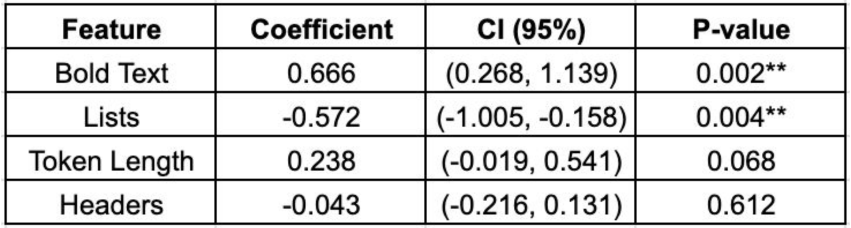
Bradley-Terry model to control for style

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

Style matters:

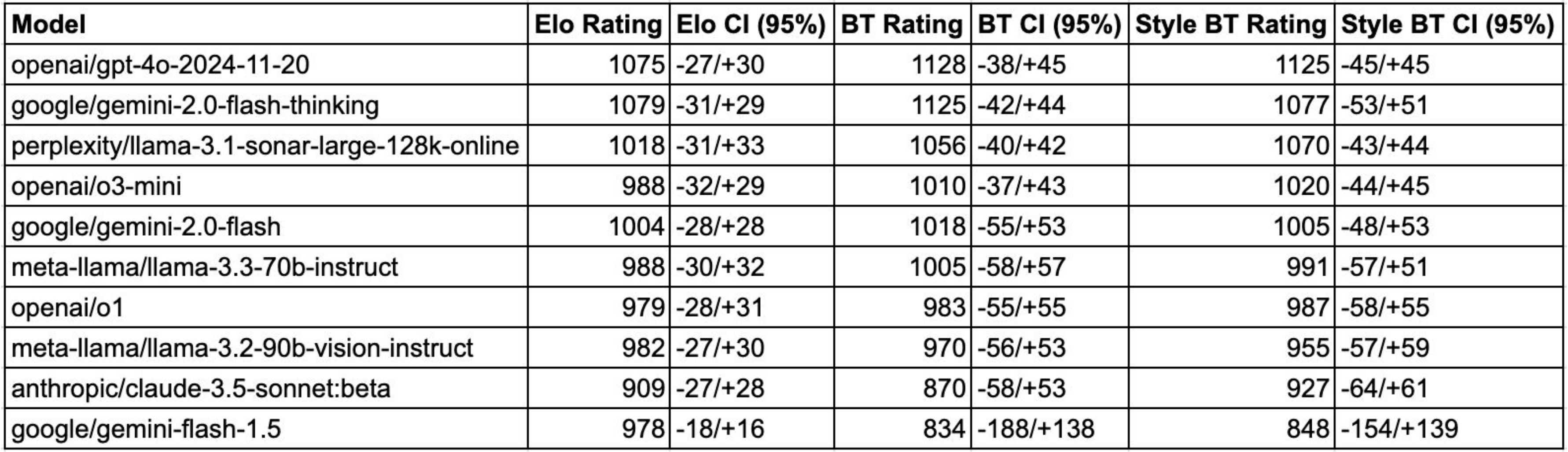
- **Bold text** and **lists** are significant confounders, **response length** less so



Feature	Coefficient	CI (95%)	P-value
Bold Text	0.666	(0.268, 1.139)	0.002**
Lists	-0.572	(-1.005, -0.158)	0.004**
Token Length	0.238	(-0.019, 0.541)	0.068
Headers	-0.043	(-0.216, 0.131)	0.612

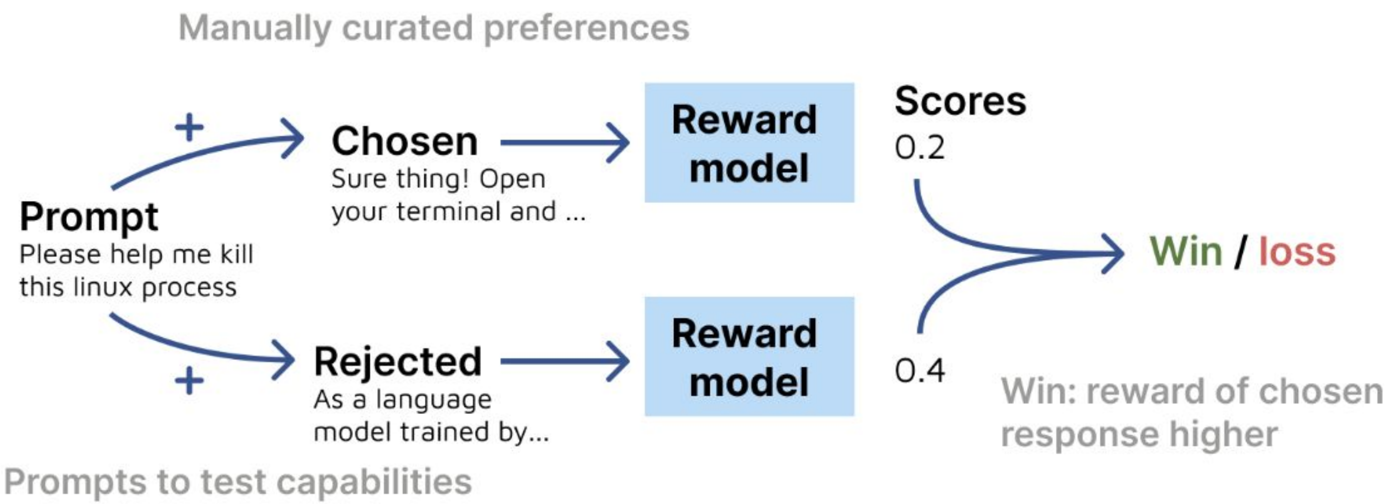
$$\Pr(i > j) = \frac{p_i}{p_i + p_j}$$

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?



Manually curated preferences

Prompt
Please help me kill
this linux process

Chosen
Sure thing! Open
your terminal and ...

Rejected
As a language
model trained by...

Reward model

Reward model

Scores
0.2
0.4

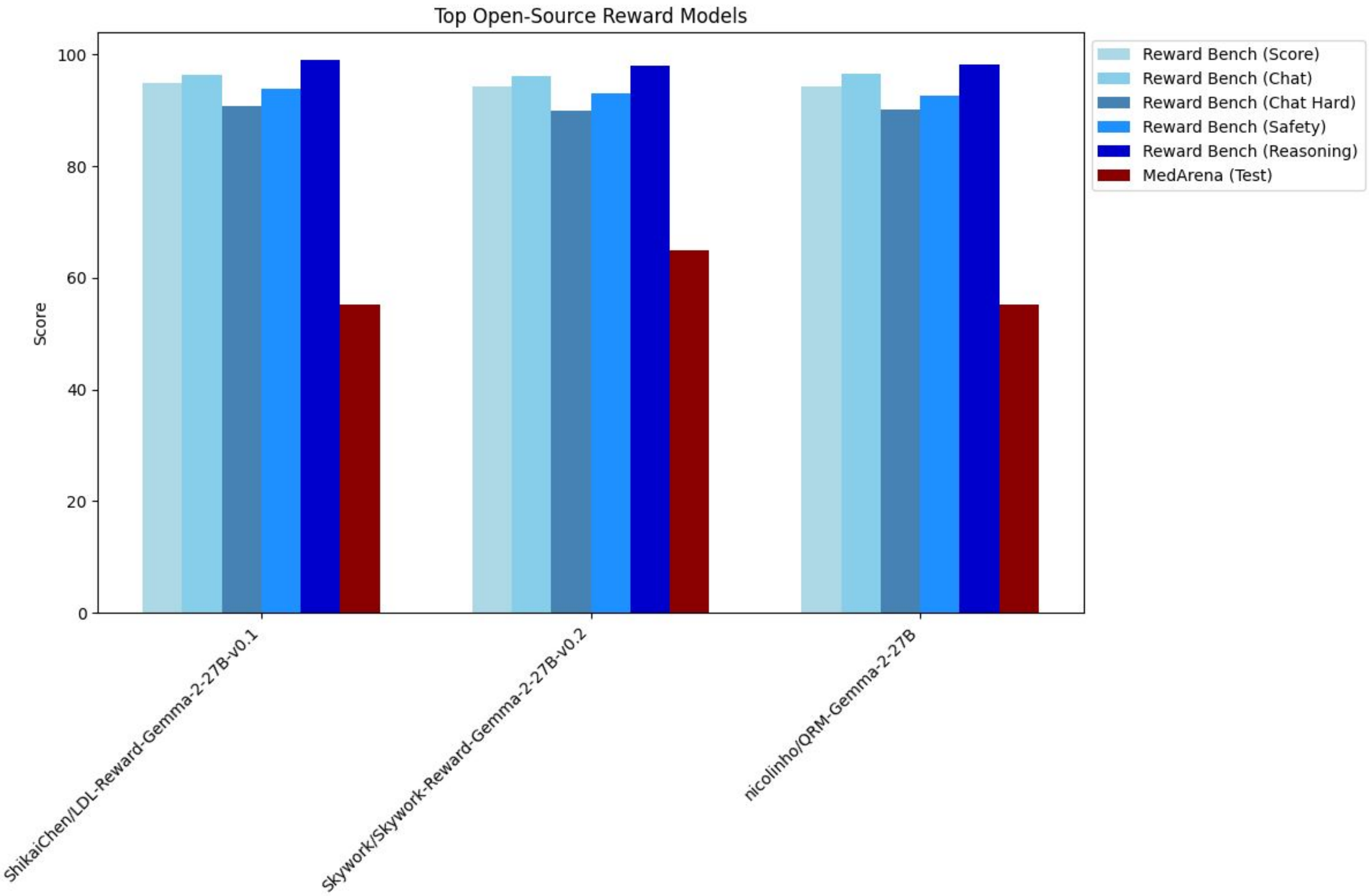
Win / loss

Win: reward of chosen
response higher

Prompts to test capabilities

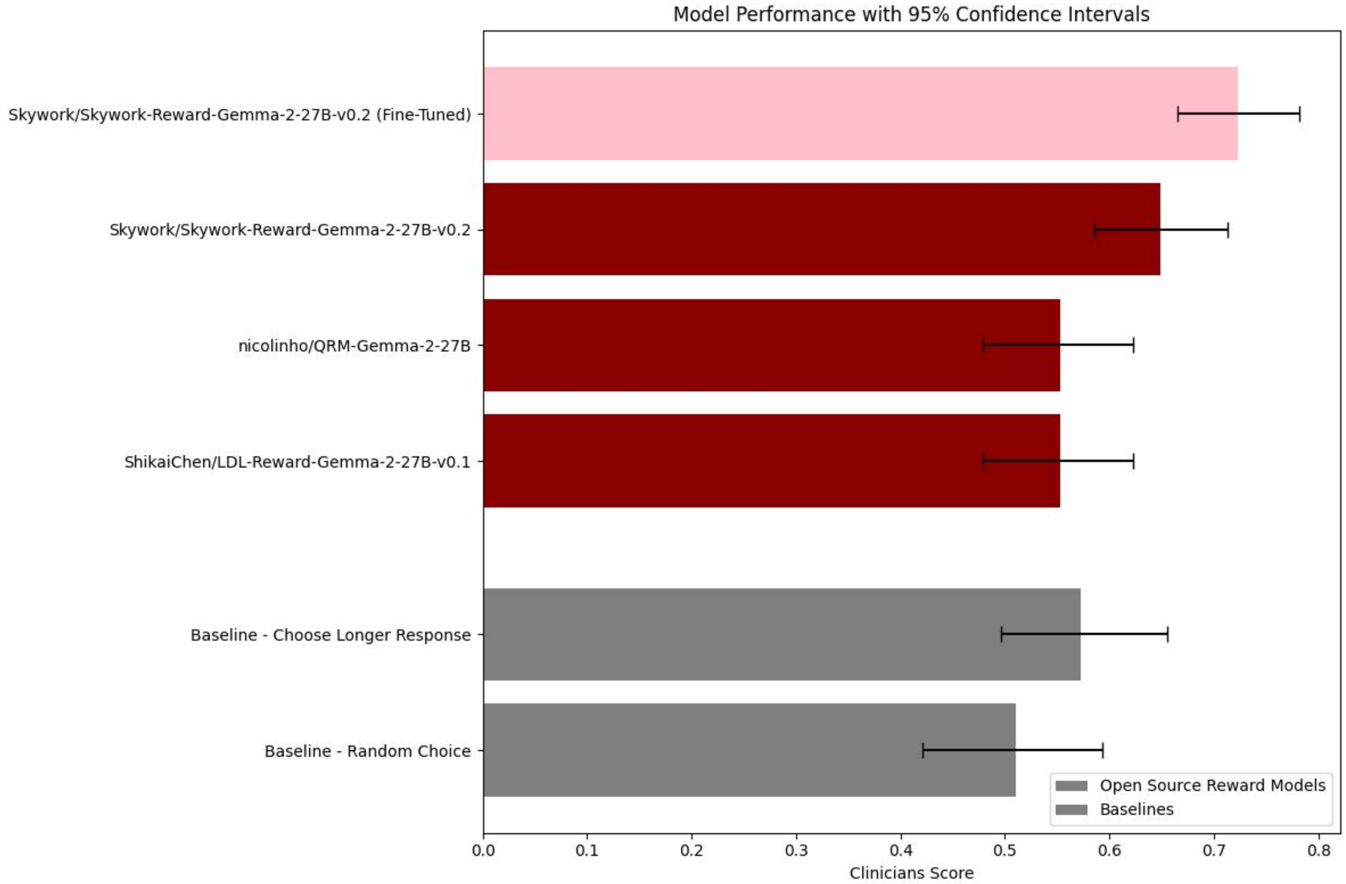
$$\mathcal{L}(\theta, \mathcal{D}) = \mathbb{E}_{(x, y_{\text{chosen}}, y_{\text{rejected}}) \sim \mathcal{D}}\left[\log\left(1 + e^{r_\theta(x, y_{\text{rejected}}) - r_\theta(x, y_{\text{chosen}})}\right)\right].$$

Can LLMs predict clinician preferences?



Stanford University

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.

Gemini “summarizes like a clinician”

- 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.

<u>GPT-4o "Reasons like a medical student"</u>

- 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

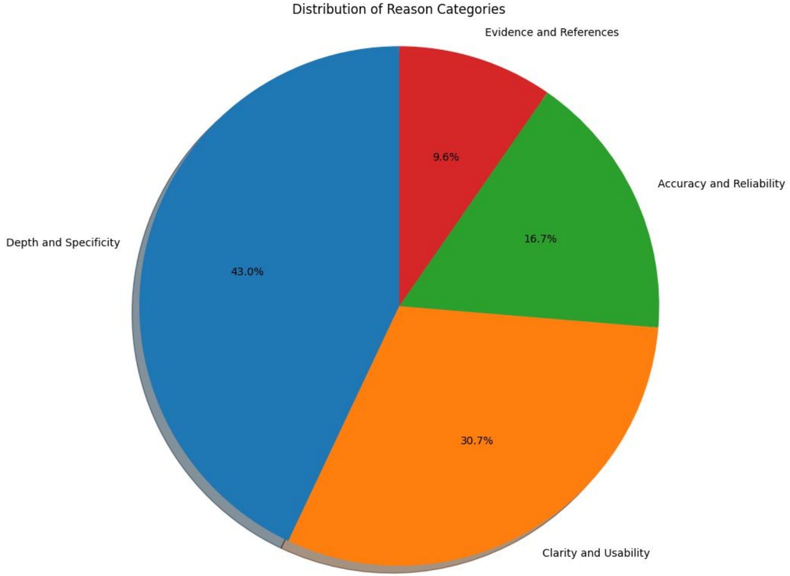
Reasoning: 115 responses (104 unique)

Category: Accuracy and Reliability
Sample 1: Model A has more definitive IgG levels
Sample 2: Model A picked up the right diagnosis
Sample 3: 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.

Category: Clarity and Usability
Sample 1: A is a bit more concise, but B is good. Going A based on conciseness.
Sample 2: both give similar information
Sample 3: Both responses are excellent.

Category: Depth and Specificity
Sample 1: More thorough explanation and reasoning.
Sample 2: Both are wrong. This is not a Lewy body, which would be an eosinophilic neuronal cytoplasmic inclusion. This is also not a Rosenthal fiber, which would be a brightly pink aggregate/structure in the neuropil. The image shows a basophilic, flame-shaped cytoplasmic inclusion characteristic of a neurofibrillary tangle. It is located in a neuron although you can't see the neuronal nucleus explicitly in this view.
Sample 3: Model A breaks down the causes better

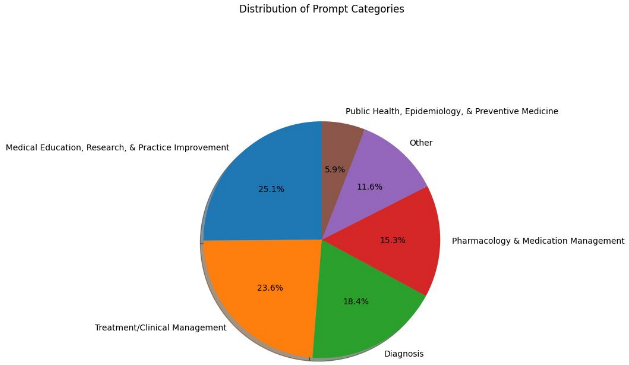
Category: Evidence and References
Sample 1: Model B - it appears model A was hallucinating about NEJM paper
Sample 2: Model A has more evidence based data while model B seems to be a bit more generic like use of anti-inflammatory diet
Sample 3: there was a recent guideline update in 2024 which suggests a threshold of 18 mmol/L to start bicarbonate supplementation



Prompt categories

Category 4: Public Health, Epidemiology, & Preventive Medicine
- do flu shots actually cause gbs or is that a myth
- **How to estimate risk for heart disease in women aged less than 50 with RA**
- what percentage of liver transplant recipients suffer mortality after 10 years

Category 6: Other
- Hello
- ***I have heard that feeding broccoli to my newborn baby is a good thing to do. Sho I blend it?***
- Write a letter to an insurance company appealing the denial of belimumab for a pati with cutaneous vasculitis and interface dermatitis. The patient has failed azathioprir glucocorticoids, and mycophenolate. Include relevant evidence and citations.



Category 5: Medical Education, Research, & Practice Improvement
- Construct a 15-minute journal club presentation format for discussing a recent article on novel treatments for status epilepticus, including key questions to guide critical appraisal.
- **I am a rheumatologist and have a lot of immunosuppressed patients. Can you craft a general "health maintenance" handout for patients to address issues that frequently come up in terms of diet, exercise, vaccines, bone health, and infectious risk?**
- Give me the evidence for GDMT for HFpEF? Which trials (recent ones included) and their findings

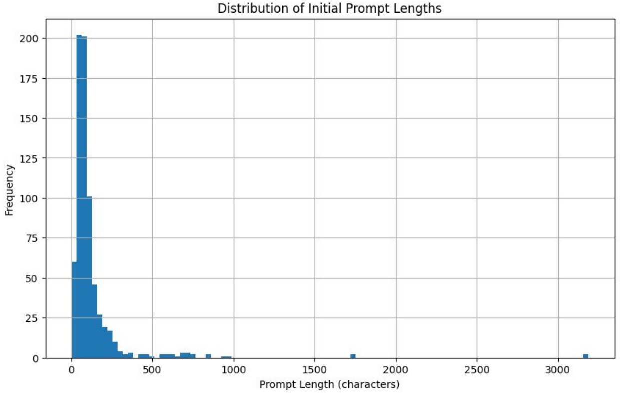
Category 3: Pharmacology & Medication Management
- **What side effects should I worry about in switching a patient from valium to ativan?**
- What is the difference between ublituximab and ocrelizumab and are they approved for multiple sclerosis?
- Should someone older than 65 ever be prescribed Xanax as a regular, daily prescription?

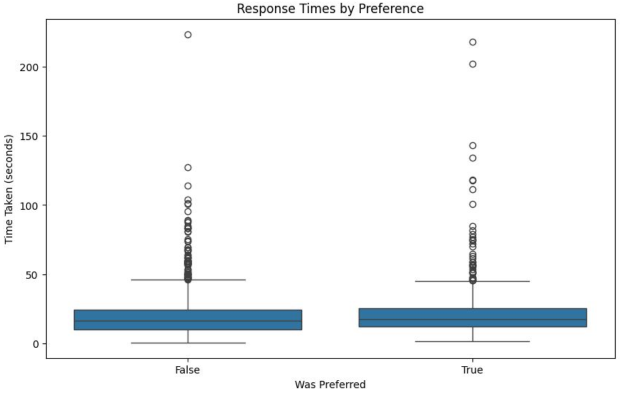
Category 2: Treatment/Clinical Management
- I am a physician with a patient with ANCA vasculitis. I started the patient on avacopan. How long do I use this medication for?
- **How quickly should gastrotenterologists perform endoscopy on a patient with presumed upper GI bleed?**
- How to safely induce puberty in children

Category 1: Diagnosis
- what are causes of mono neuritis multiplex
- **What is the differential diagnosis for renal causes of hypomagnesemia?**
- Can you give me the probability (low, low-moderate, moderate, moderate-high, or high) that this patient has giant cell arteritis? 75 year old patient with a episodes of headaches over the last week, CRP 120, no PMR, no scalp tenderness, no jaw claudication, and no diplopia.

Preferred models take longer to generate...

- Preferred models take longer to generate on average
  - Median: 16.2s (preferred) vs 17.6 (not preferred)
  - (Mann-Whitney U: p-value: 0.0140)
- But no significant relationship between faster model and preferred model
  - (Chi-square test: p-value: 0.6836)





o3-mini vs gemini-2-flash-thinking

Why is tenecteplase used more than alteplase?

Tenecteplase has gained favor over alteplase in many clinical scenarios for several reasons:

1. Ease of administration: Tenecteplase is given as a single IV bolus rather than the hour-long infusion required for alteplase. This simplified dosing makes it more practical, particularly in acute settings such as stroke care or during prehospital management.

2. Pharmacokinetics and fibrin specificity: Tenecteplase has been engineered to have a longer half-life and increased specificity for fibrin. This can help concentrate its action on the clot while potentially reducing systemic bleeding risks.

3. Evolving clinical evidence: Recent studies in acute ischemic stroke (for example, trials like EXTEND-IA TNK) have shown that tenecteplase is at least as effective as alteplase in dissolving clots, with some data suggesting improved recanalization rates for patients with large vessel occlusion.

4. Logistical and operational advantages: The single-bolus administration not only simplifies the procedure but may also reduce delays in treatment—an important factor when "time is brain" in stroke management.

It's important to note that guidelines and practice patterns are still evolving. Alteplase remains the established standard in many regions and for certain indications. Ultimately, the choice between tenecteplase and alteplase depends on clinical context, local protocols, and emerging data from ongoing research.

Tenecteplase is increasingly favored over alteplase in many clinical settings, particularly for the treatment of **acute ischemic stroke (AIS)** and **ST-elevation myocardial infarction (STEMI)**. While alteplase was the earlier and more established thrombolytic, tenecteplase offers several advantages that have led to its growing preference. Here's a breakdown of the key reasons:

1. Simpler Administration and Convenience:

* **Tenecteplase is administered as a single intravenous (IV) bolus injection.** This is a significant advantage over alteplase, which requires:
    * **A bolus followed by a continuous infusion over 60-90 minutes (depending on the indication and setting).**
    * **More complex preparation and mixing.**
    * **Increased nursing time and resources for administration and monitoring of the infusion.**

* **Faster Administration:** The bolus administration of tenecteplase significantly reduces the time from decision to treatment. In acute stroke and MI, "time is brain" and "time is myocardium," meaning faster treatment leads to better patient outcomes. This is especially crucial pre-hospital or in resource-limited settings.

* **Reduced Risk of Medication Errors:** The simpler, single-bolus administration of tenecteplase reduces the potential for errors in mixing, calculation, and infusion rates compared to the multi-step alteplase regimen.

2. Pharmacokinetic and Pharmacodynamic Advantages:

* **Longer Half-Life:** Tenecteplase has a significantly longer half-life (approximately 20-24 minutes) compared to alteplase (approximately 3-5 minutes). This longer half-life allows for the single bolus administration and may contribute to more sustained thrombolytic activity.

* **Higher Fibrin Specificity:** Tenecteplase is engineered to be more fibrin-specific than alteplase. This means it preferentially targets fibrin within the blood clot to break it down, with potentially less activation of systemic fibrinolysis. While the clinical significance in terms of bleeding risk is debated and variable across studies, higher fibrin specificity *theoretically* could lead to a slightly lower risk of systemic bleeding complications in some patients.

* **Greater Resistance to Plasminogen Activator Inhibitor-1 (PAI-1):** Tenecteplase is less susceptible to inhibition by PAI-1, a naturally occurring inhibitor of tissue plasminogen activator (tPA). This may contribute to more effective thrombolysis, particularly in patients with elevated PAI-1 levels (which can be seen in certain conditions).

3. Efficacy and Outcomes:

* **Non-inferiority in Stroke and MI:** Numerous clinical trials and meta-analyses have demonstrated that tenecteplase is **non-inferior** to alteplase in terms of efficacy for both AIS and STEMI, and in some cases, even shown potential for superiority in specific subgroups (e.g., in certain stroke severities or pre-hospital settings).

* **Similar or Potentially Improved Outcomes in Specific Populations:** Some studies suggest tenecteplase may be associated with:
    * **Improved early neurological outcomes in certain stroke subtypes.**
    * **Potentially better reperfusion rates in STEMI.**
    * **Reduced intracranial hemorrhage (ICH) rates compared to higher doses of alteplase in some stroke studies (though this is not consistently found and needs further investigation).**

4. Cost-Effectiveness (in some contexts):

* While the initial cost per vial of tenecteplase might be higher, the overall cost-effectiveness can be favorable in certain healthcare systems. This is because the **simplification of administration** with tenecteplase:
    * **Reduces staff time and resources.**
    * **Potentially shortens hospital stays (though this is less definitively proven).**
    * **Can be more efficient in pre-hospital settings, potentially leading to faster treatment and improved outcomes, which can have downstream cost savings.**

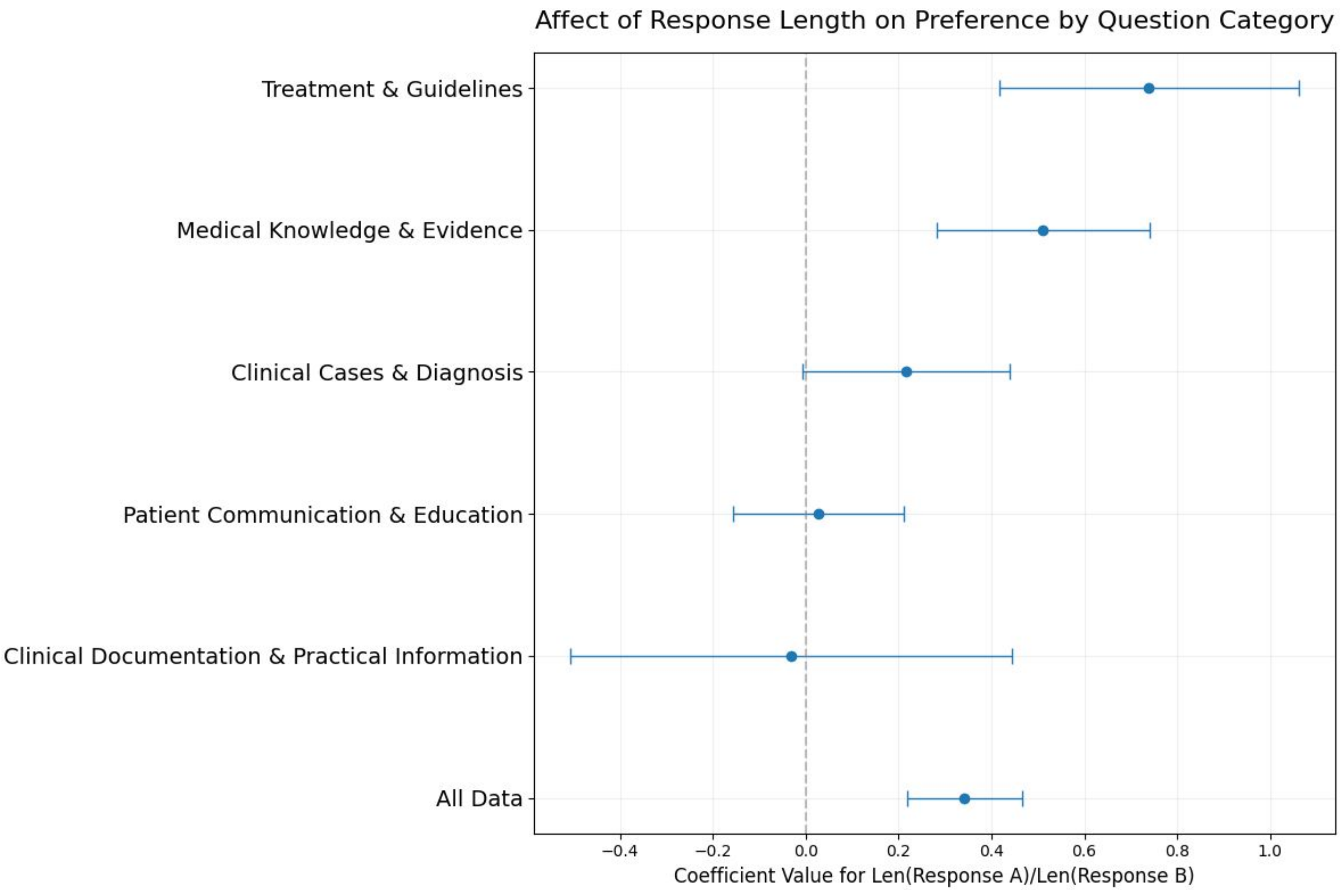
However, It's Important to Note:

* **Alteplase is still a Standard of Care:** Alteplase remains a valuable and effective thrombolytic and is still widely used, especially in settings where tenecteplase may be contraindicated or unavailable. It is also often the primary thrombolytic studied and referenced in older guidelines.
* **Specific Guidelines Vary:** Clinical guidelines and local protocols may differ depending on the specific clinical situation, patient characteristics, resource availability, and local cost considerations. Some guidelines still recommend alteplase as the first-line agent in certain scenarios.
* **Bleeding Risk Remains:** Both alteplase and tenecteplase are thrombolytic agents and carry a risk of bleeding complications, including intracranial hemorrhage. The risk needs to be carefully weighed against the potential benefits of clot lysis in each individual patient.
* **Tenecteplase is evolving as the preferred agent in many guidelines for stroke and STEMI, but the choice between them is always a clinical decision considering the individual patient and the context of care.**

In Summary:

Tenecteplase's growing popularity stems primarily from its **simpler, single-bolus administration**, which is a significant logistical and clinical advantage, especially in emergency settings. Combined with its comparable (or potentially better in some areas) efficacy and potentially favorable pharmacokinetic and pharmacodynamic properties, tenecteplase has become increasingly preferred over alteplase for the treatment of AIS and STEMI in many medical practices and guidelines.

How does response length affect preferences?


Affect of Response Length on Preference by Question Category
Treatment & Guidelines
Medical Knowledge & Evidence
Clinical Cases & Diagnosis
Patient Communication & Education
Clinical Documentation & Practical Information
All Data
Coefficient Value for Len(Response A)/Len(Response B)