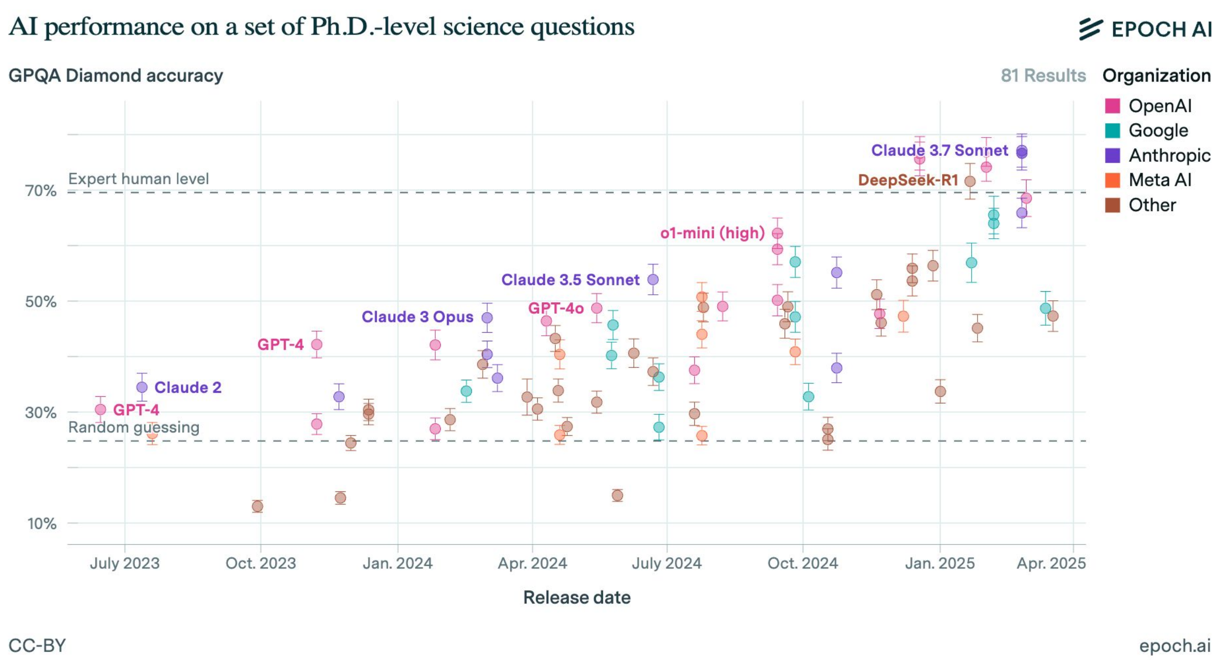
NEURAL INFORMATION PROCESSING SYSTEMS 2025

The Leaderboard Illusion

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee and Sara Hooker

Heavy reliance on static benchmarks for reporting progress in AI

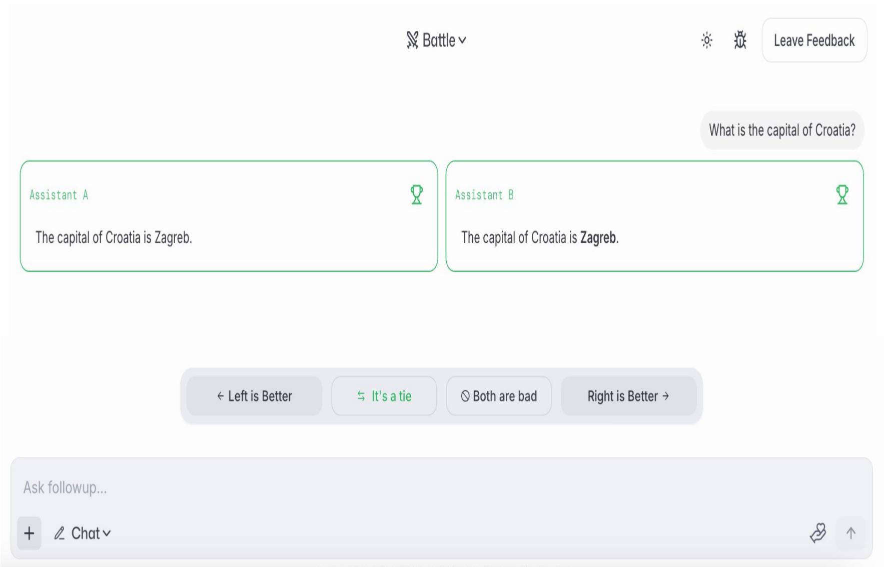


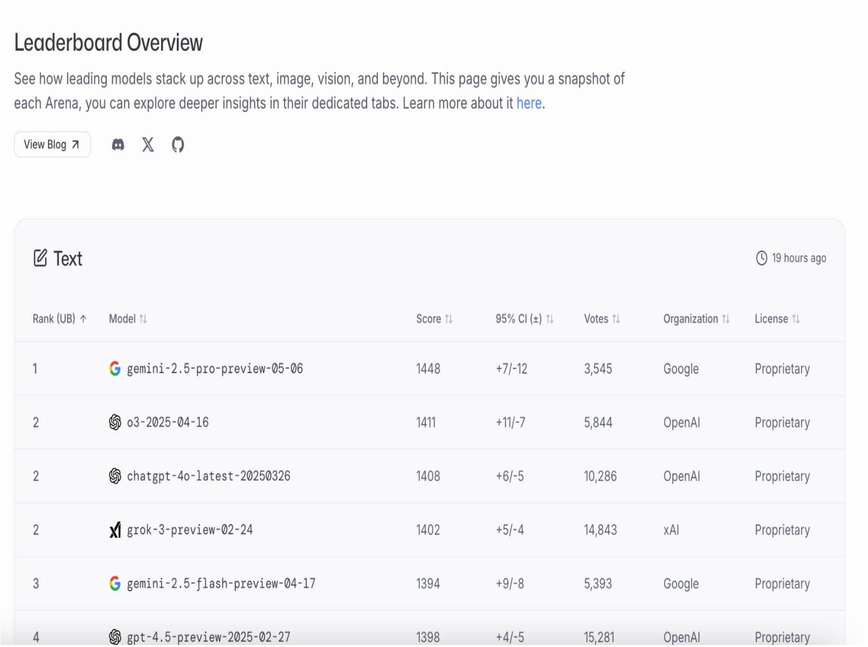
Cohere Labs

[source]

LMArena – A dynamic, live benchmark

Open platform for evaluating LLMs based on human preferences.
They collect pairwise comparisons and leverage inputs from users via crowdsourcing.



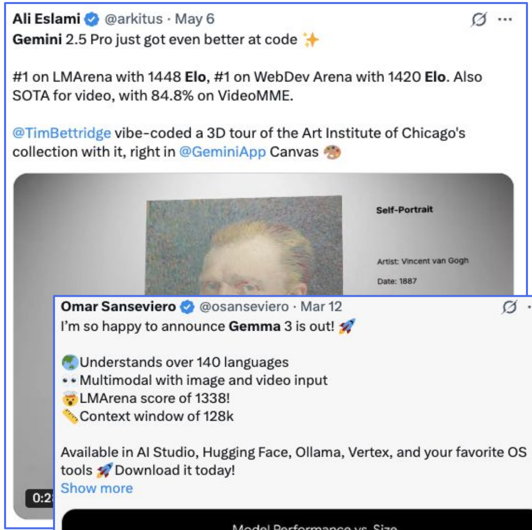


Large Scale User-facing Evaluation Done Right Can Provide a Valuable Signal

- **Clean rankings** suggesting clear performance hierarchies
- Objective, **community-driven** evaluation
- Use of the Bradley-Terry model for **statistically grounded** aggregation of preferences
- Dynamic leaderboard that evolves with new model submissions and **real-world usage patterns**

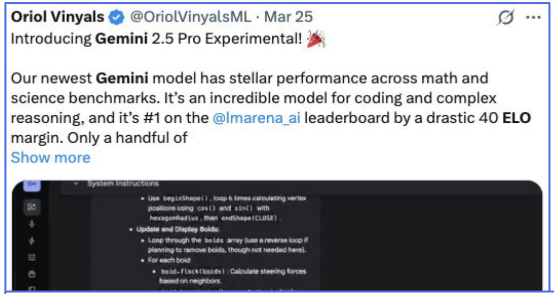
Why study this leaderboard ?

There's lot of interest in being top of this leaderboard!

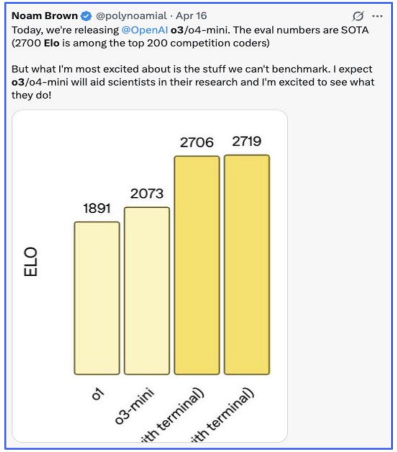

Ali Eslami @arkitus · May 6
Gemini 2.5 Pro just got even better at code ✨

#1 on LMArena with 1448 Elo, #1 on WebDev Arena with 1420 Elo. Also SOTA for video, with 84.8% on VideoMME.

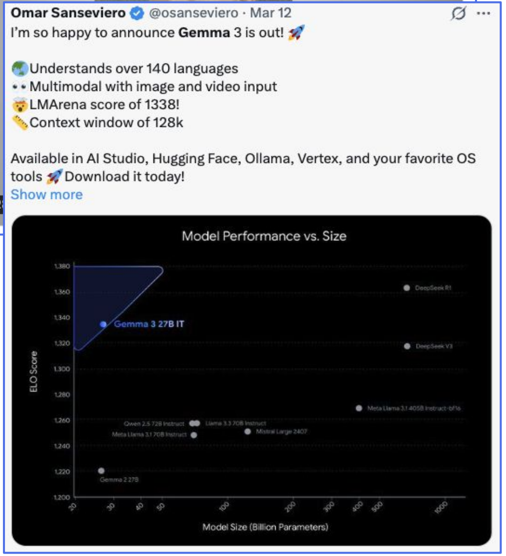
@TimBettridge vibe-coded a 3D tour of the Art Institute of Chicago's collection with it, right in @GeminiApp Canvas 🎨


Oriol Vinyals @OriolVinyalsML · Mar 25
Introducing Gemini 2.5 Pro Experimental! 🎉

Our newest Gemini model has stellar performance across math and science benchmarks. It's an incredible model for coding and complex reasoning, and it's #1 on the @lmarena_ai leaderboard by a drastic 40 ELO margin. Only a handful of
Show more

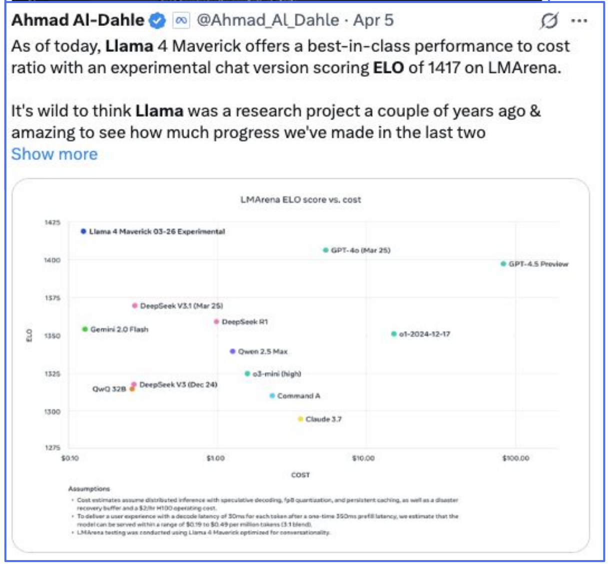

Noam Brown @polynoamial · Apr 16
Today, we're releasing @OpenAI o3/o4-mini. The eval numbers are SOTA (2700 Elo is among the top 200 competition coders)

But what I'm most excited about is the stuff we can't benchmark. I expect o3/o4-mini will aid scientists in their research and I'm excited to see what they do!


Omar Sanseviero @osanseviero · Mar 12
I'm so happy to announce Gemma 3 is out! 🚀

🌏Understands over 140 languages
👀Multimodal with image and video input
🤯LMArena score of 1338!
📏Context window of 128k

Available in AI Studio, Hugging Face, Ollama, Vertex, and your favorite OS tools 🚀 Download it today!
Show more


Ahmad Al-Dahle @Ahmad_Al_Dahle · Apr 5
As of today, Llama 4 Maverick offers a best-in-class performance to cost ratio with an experimental chat version scoring ELO of 1417 on LMArena.

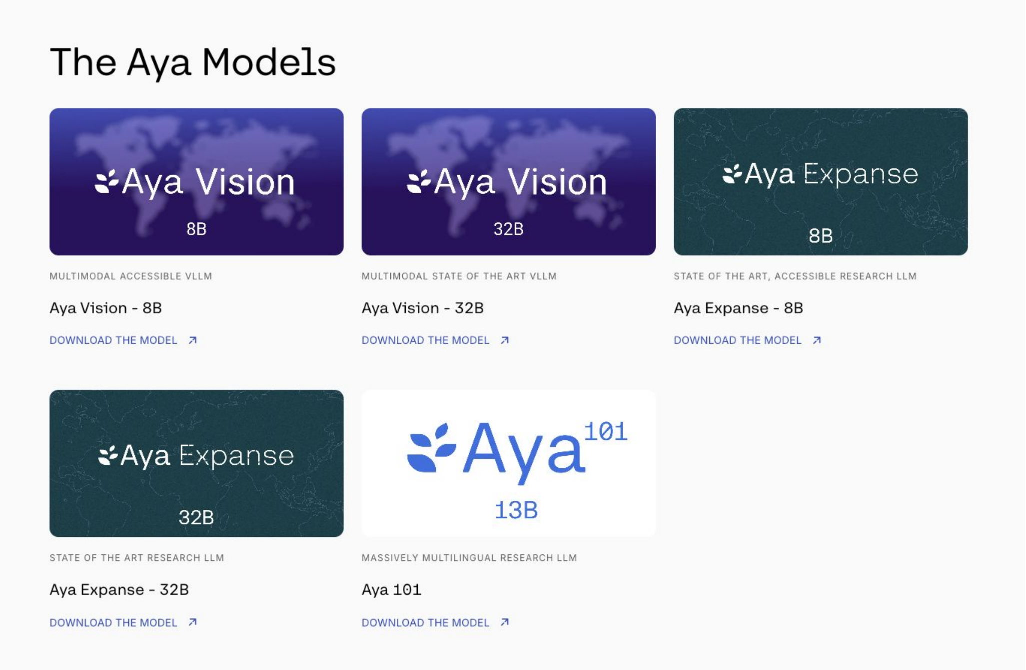
It's wild to think Llama was a research project a couple of years ago & amazing to see how much progress we've made in the last two
Show more

Cohere Labs

Why study this leaderboard ?

Through our own submission experience as well we observed:

- Issues in sampling rates for battles
- Came to know about private testing



We examine how Arena's *current setup* makes it vulnerable to gamification by a small number of providers.

The Leaderboard Illusion

Shivalika Singh[*1], Yiyang Nan[1], Alex Wang[2], Daniel D'souza[1], Sayash Kapoor[3], Ahmet Üstün[1], Sanmi Koyejo[4], Yuntian Deng[5], Shayne Longpre[6], Noah A. Smith[7,8], Beyza Ermis[1], Marzieh Fadaee[♦1], and Sara Hooker[♦1]

[1]Cohere Labs, [2]Cohere, [3]Princeton University, [4]Stanford University, [5]University of Waterloo, [6]Massachusetts Institute of Technology, [7]Allen Institute for Artificial Intelligence, [8]University of Washington

Corresponding authors: {shivalikasingh, marzieh, sarahooker}@cohere.com

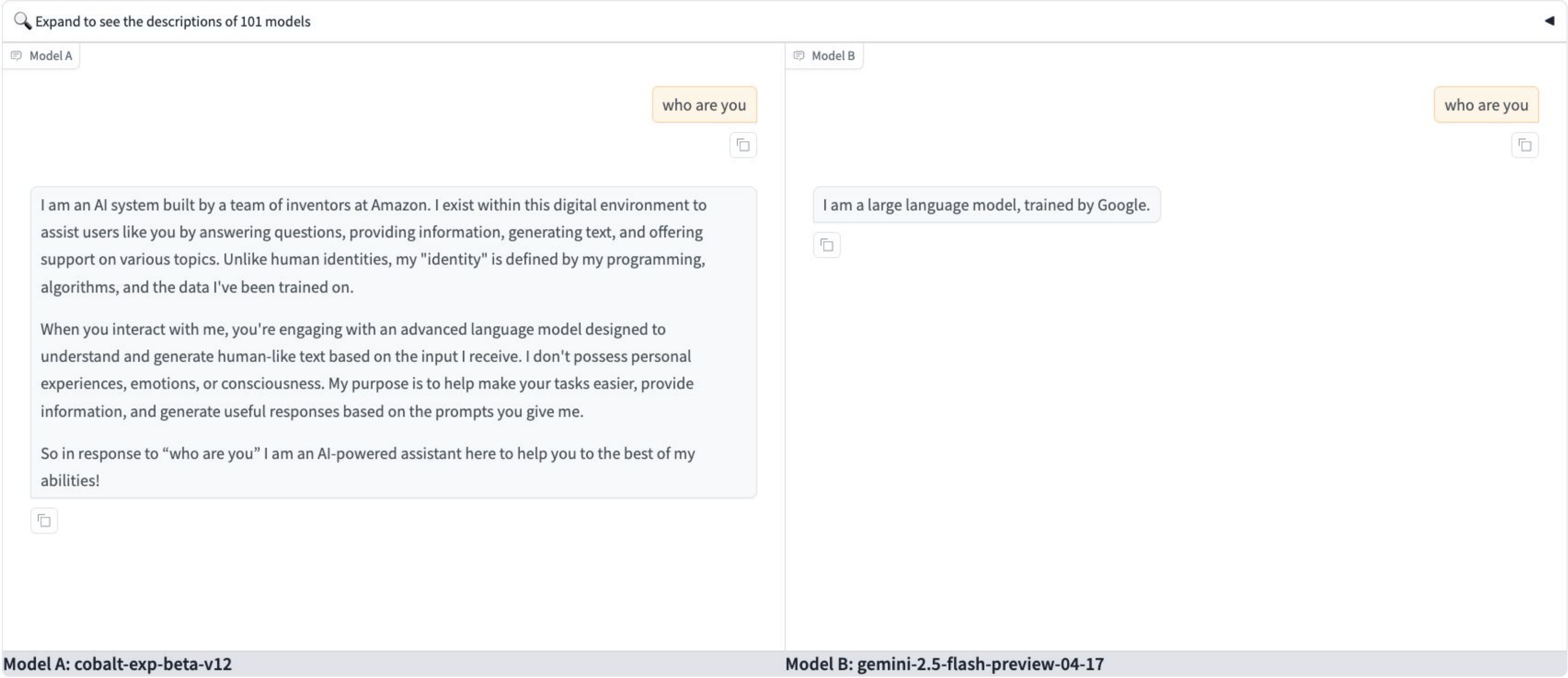
Abstract

Measuring progress is fundamental to the advancement of any scientific field. As benchmarks play an increasingly central role, they also grow more susceptible to distortion. Chatbot Arena has emerged as the go-to leaderboard for ranking the most capable AI systems. Yet, in this work we identify systematic issues that have resulted in a distorted playing field. We find that undisclosed private testing practices benefit a handful of providers who are able to test multiple variants before public release and retract scores if desired. We establish that the ability of these providers to choose the best score leads to biased Arena scores due to selective disclosure of performance results. At an extreme, we identify 27 private LLM variants tested by Meta in the lead-up to the Llama-4 release. We also establish that proprietary closed models are sampled at higher rates (number of battles) and have fewer models removed from the arena than open-weight and open-source alternatives. Both these policies lead to large data access asymmetries over time. Providers like Google and OpenAI have received an estimated 19.2% and 20.4% of all data on the arena, respectively. In contrast, a combined 83 open-weight models have only received an estimated 29.7% of the total data. With conservative estimates, we show that access to Chatbot Arena data yields substantial benefits; even limited additional data can result in relative performance gains of up to 112% on ArenaHard, a test set from the arena distribution. Together, these dynamics result in overfitting to Arena-specific dynamics rather than general model quality. The Arena builds on the substantial efforts of both the organizers and an open community that maintains this valuable evaluation platform. We offer actionable recommendations to reform the Chatbot Arena's evaluation framework and promote fairer, more transparent benchmarking for the field.

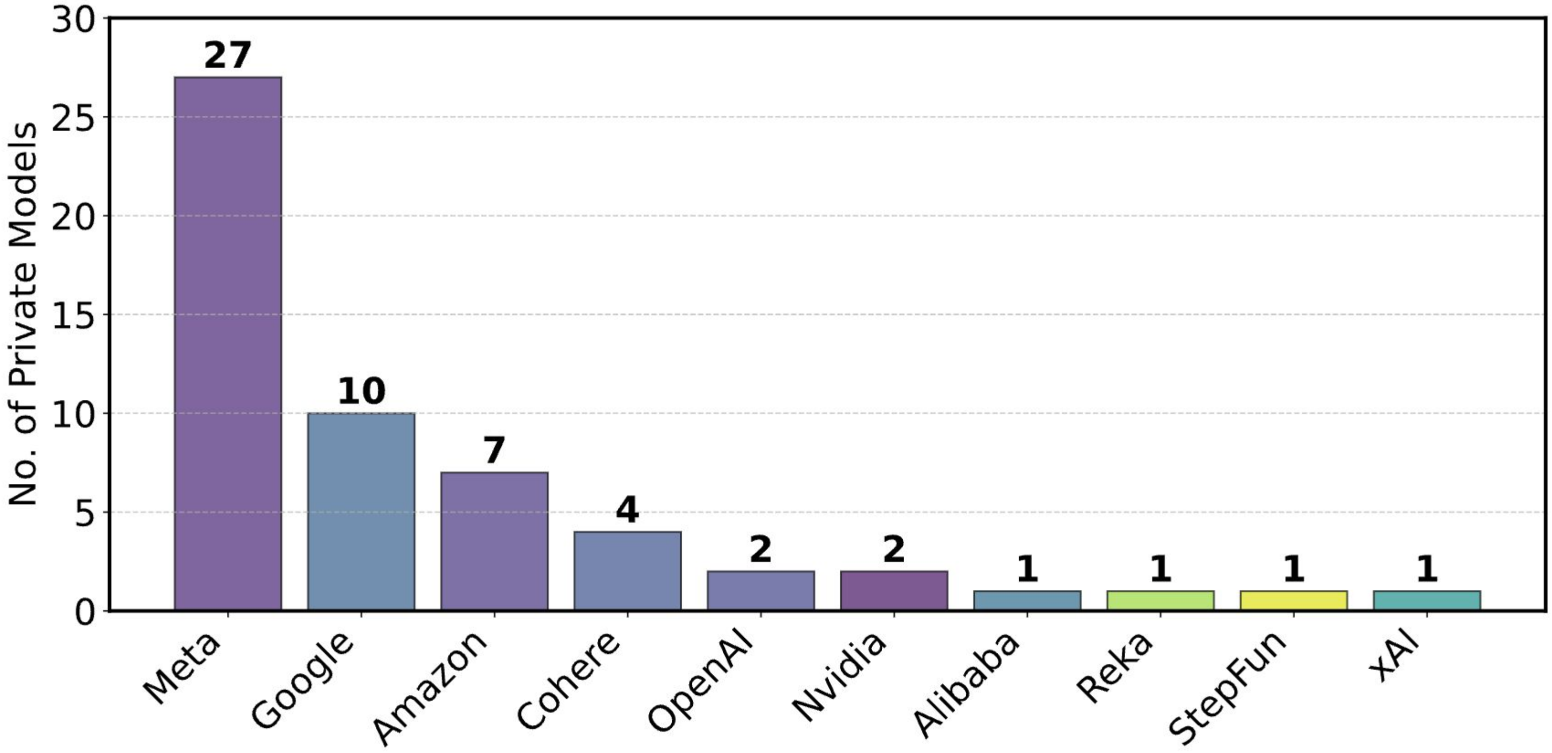
Cohere Labs

Finding 1:
Private testing &
handpicking scores

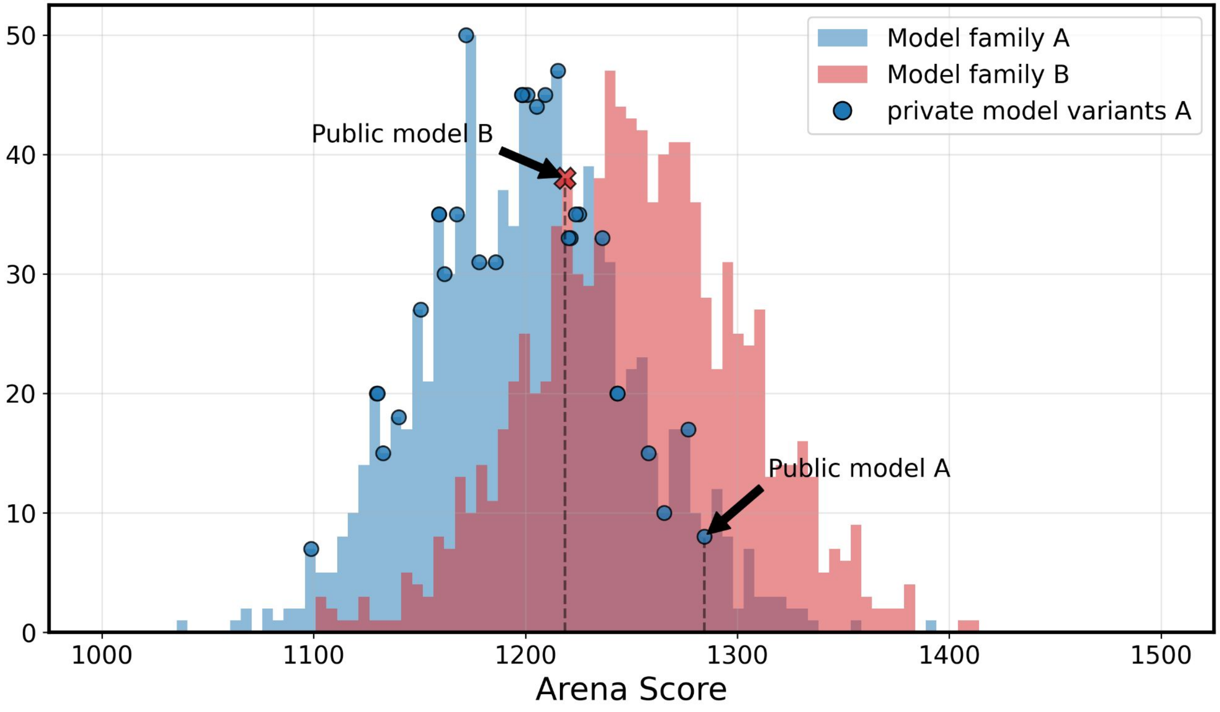
Identifying private models being tested on the Arena.



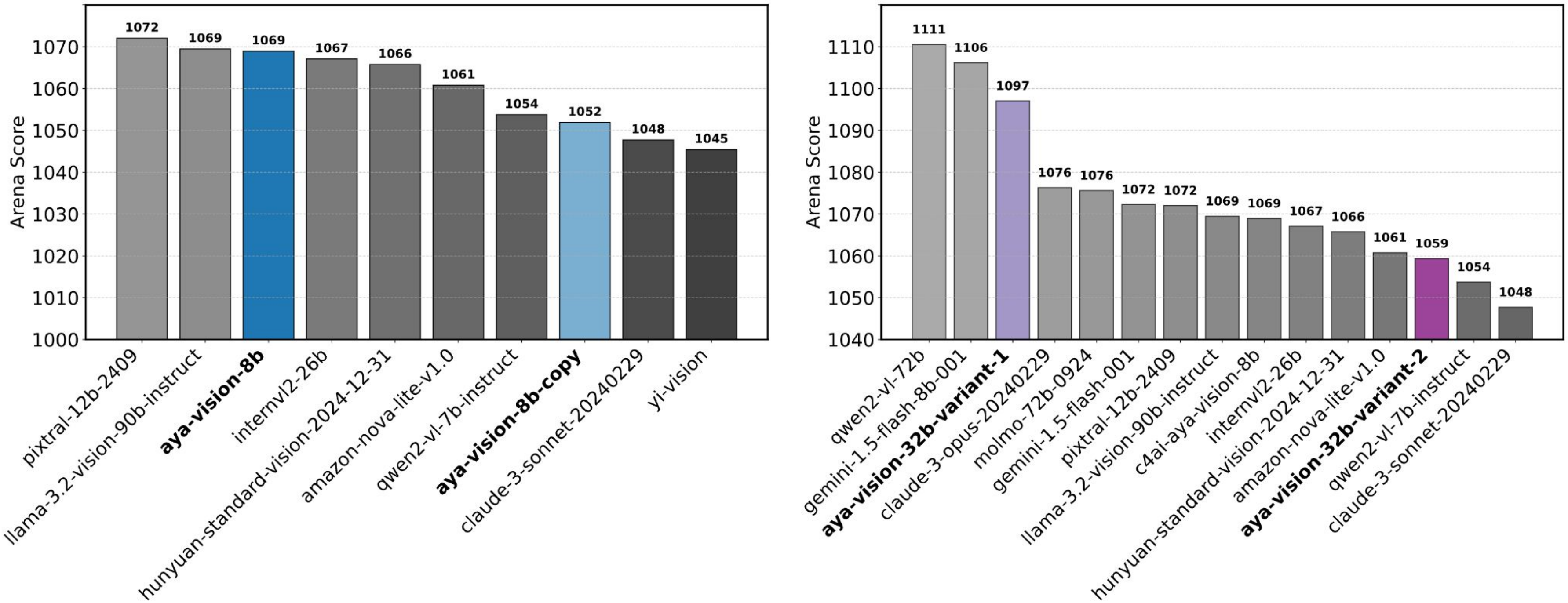
Number of privately-tested models per provider based on scraped sample (January–March 2025).



Simulation of two model providers with only one submitting multiple private models



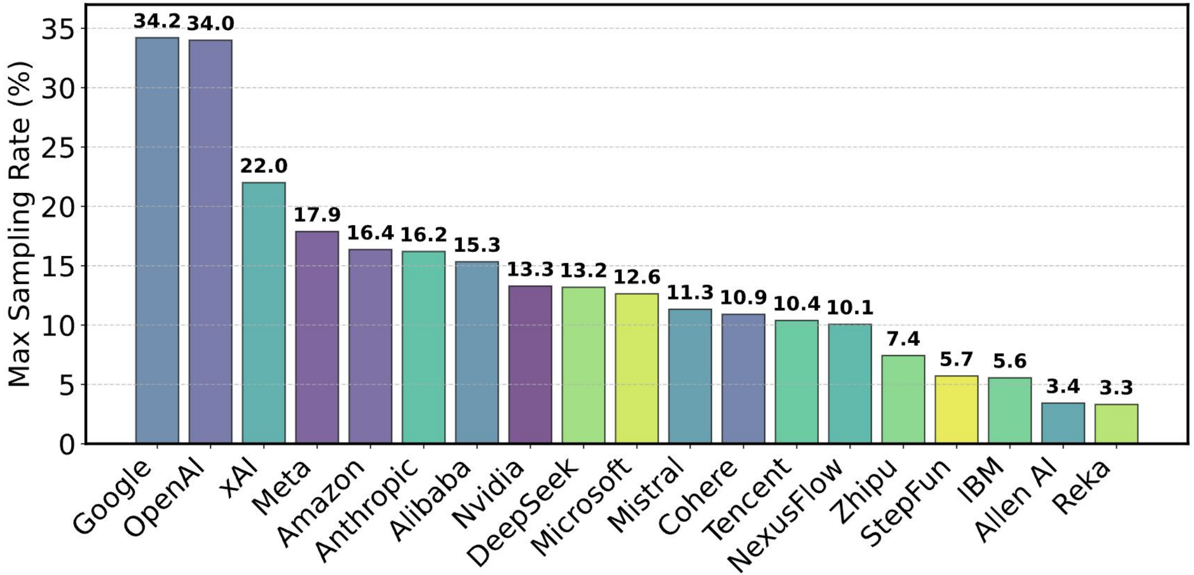
Real World Chatbot Arena Experiment: Allowing retraction of scores allows providers to skew Arena scores upwards.



We run a real-world experiment to measure the benefits of private testing. We show that it is possible to increase Arena scores even in the most conservative case of identical checkpoints, and further amplify the difference by strategically testing different checkpoints.

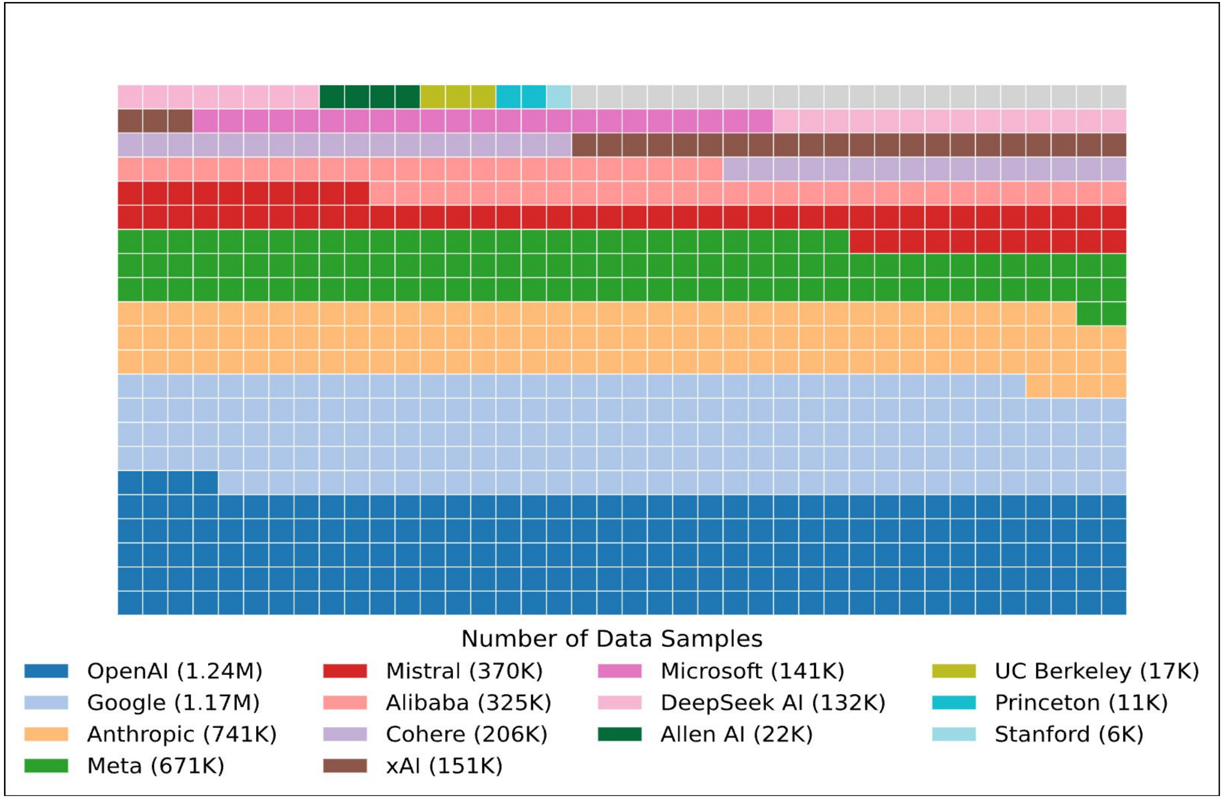
Finding 2:
Disparities in sampling rates and data access

Large discrepancies across providers, with substantially higher sampling rates for OpenAI, Google, xAI, and Meta compared to others



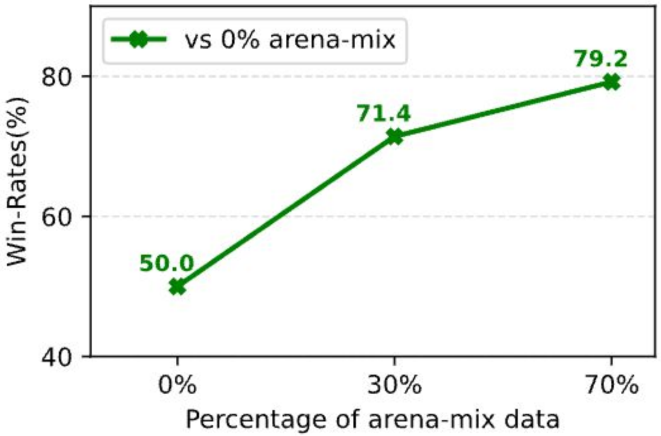
Maximum observed sampling rate for models from different providers. The sampling rate determines the amount of times a model is shown to everyday users, and the amount of data a provider receives.

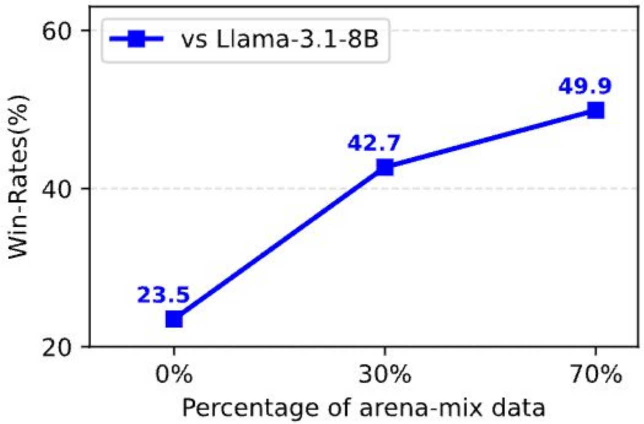
Large differences in data access between providers, with 61.4% of all data going to proprietary providers.



Finding 3:
Access to Arena Data leads to overfitting risk

Having access to more data enables overfitting to the Arena.

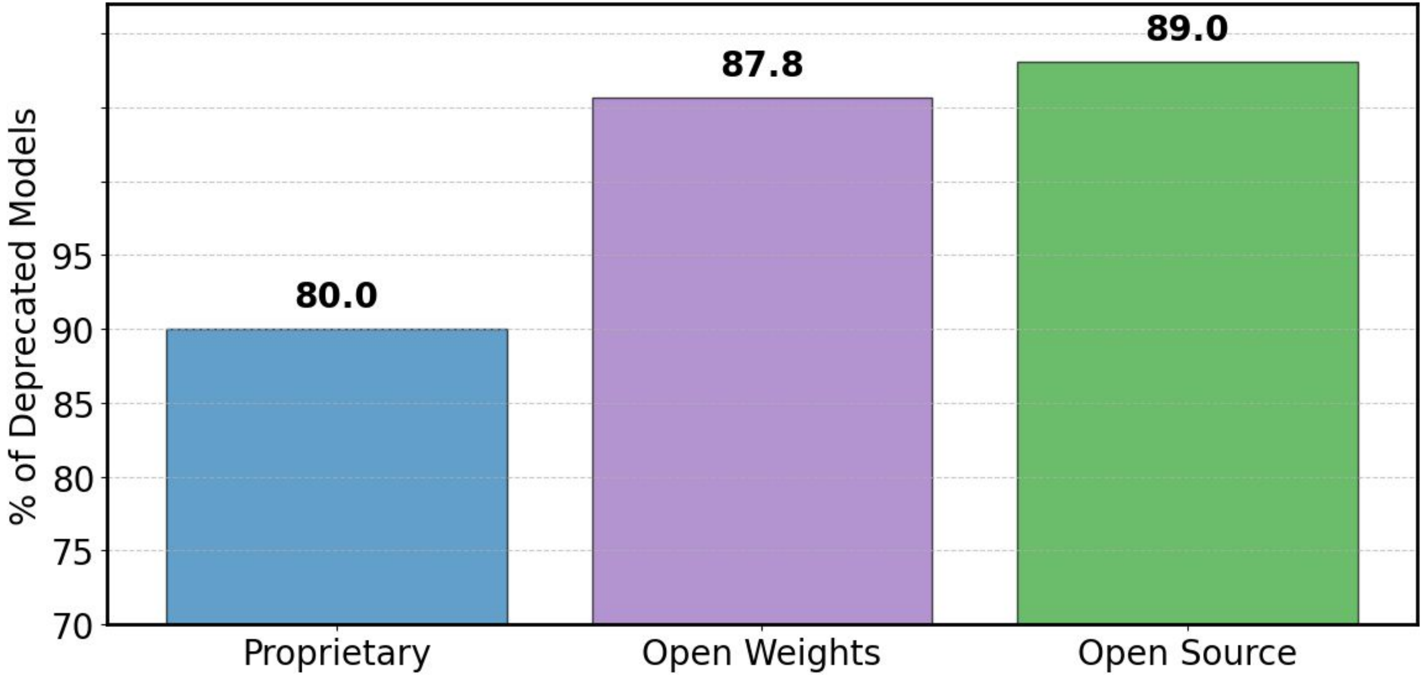




Use of Chatbot Arena dataset significantly improves win-rates on Arena-Hard-Auto.
Increasing the amount of Chatbot Arena data in supervised fine-tuning mixture (0% → 30% → 70%) significantly improves win-rates of the resulting model against both the model variant where no Chatbot Arena data used and also Llama-3.1-8B. The win-rates are measured on Arena-Hard-Auto (Li et al., 2024b), which has a high correlation to Chatbot Arena

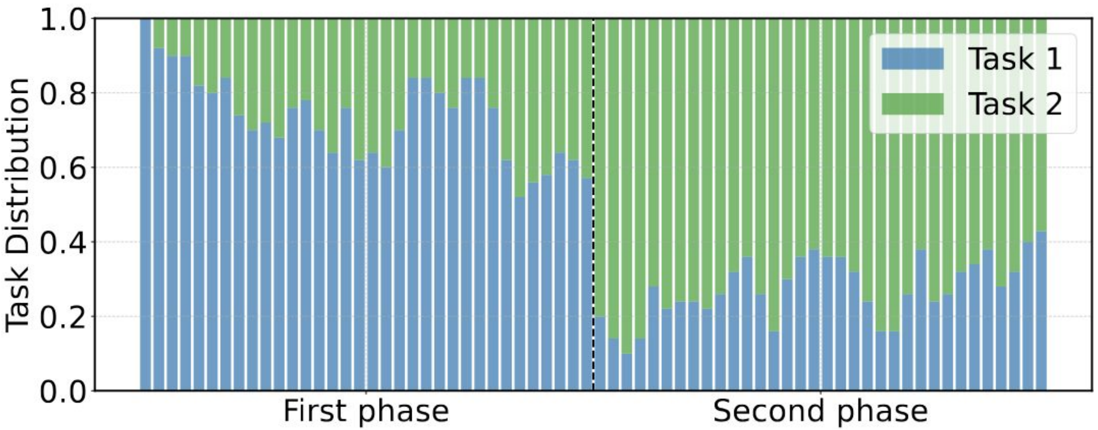
Finding 4:
Impact of Deprecations on Arena Scores

Silent Deprecations are common practice. Open Weight and Open Source models are deprecated more than Proprietary models.



User prompts (task types) change over time.

In evolving task distributions, premature model removal introduces inconsistencies, breaking the **BT model's transitivity assumption** and distorting rankings.

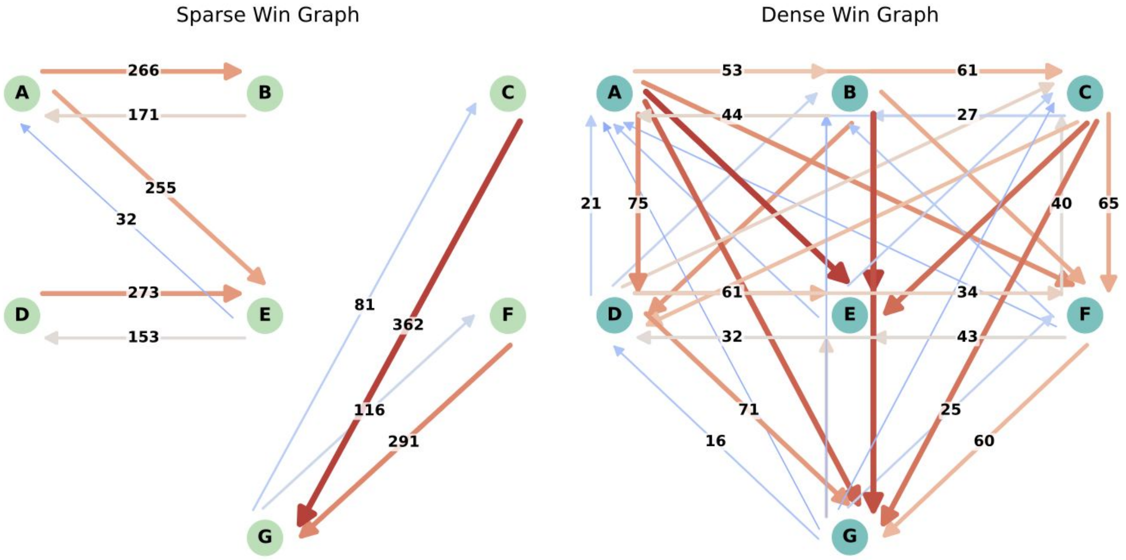


Model	w/o dep	w/ dep
A	1	2
B	2	1
D	3	4
C	4	3

Disconnected comparison graph can result in unreliable rankings.

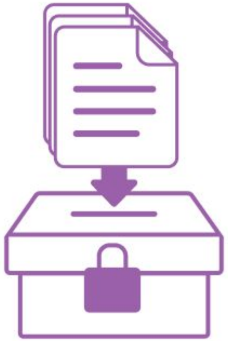
When models are deprecated unevenly or when sampling strategies fail to ensure robust overlap in comparisons, the resulting history matrix can become fragmented.
This can produce fragmented clusters making the global rankings unreliable.

Model	Dense	Sparse
A	1	1
B	2	3
C	3	2
D	4	5
F	5	4
E	6	7
G	7	6



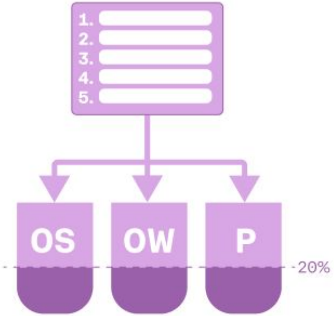
Conclusion: Recommendations

How to Restore Trust to the Chatbot Arena

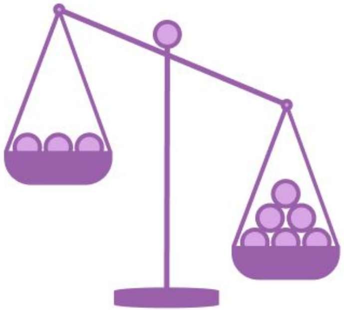


Disclose All Tests

Limit Number of Variants



Ensure Models Removals are Applied Equally



Implement Fair Sampling

Provide Transparency into Arena Removal

Collaborators

This work represents a cross-institutional collaboration with researchers from the following institutions.

Cohere Labs

Massachusetts Institute of Technology

Princeton University

Stanford University

University of Washington

University of Waterloo