

DEMO!



# Chatbot Arena

## An Open Platform for Evaluating LLMs by Human Preference

Wei-Lin Chiang\*, Lianmin Zheng\*, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li  
Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, Ion Stoica



Demo: [leaderboard.lmsys.org](https://leaderboard.lmsys.org)  
Slides: <https://tinyurl.com/2s7jxv4h>  
Contact: [weichiang@berkeley.edu](mailto:weichiang@berkeley.edu)

the goal:

## better evals for LLMs

the ultimate end-to-end metric: **human satisfaction**

the platform:

### 2. Review — wrong interpretation

```
while True:
    result = math.sqrt(2)
    print("The square root of 2 is:", result)
    time.sleep(1) # This will print the result every second
```

### 3. Click: B is better!

### 2. Review — Correct, well organized

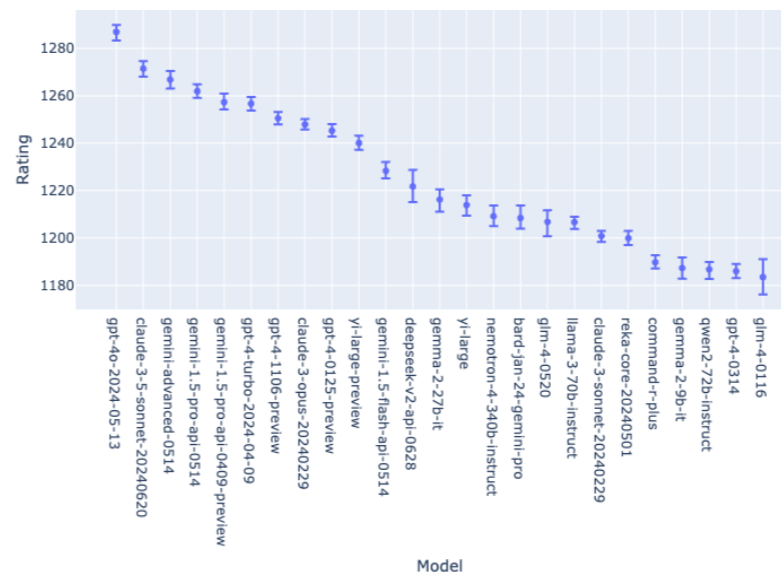
the results:

### 1. Ask any question

Bradley-Terry coefficients (logistic regression coefficients for win-prediction)

- ▶ Over **1.5 million** user votes, **100+** models
- ▶ 10+ categories (coding, math, multi-lingual)
- ▶ Open prompt & preference datasets
- ▶ Adoption by frontier AI labs at Google, OpenAI, Anthropic, Meta, ...

Overall — 1.5M votes



Coding — 280K votes (19%)

