

---

# The Leaderboard Illusion

Shivalika Singh<sup>\*1</sup>, Yiyang Nan<sup>1</sup>, Alex Wang<sup>2</sup>, Daniel D’souza<sup>1</sup>,  
Sayash Kapoor<sup>3</sup>, Ahmet Üstün<sup>1</sup>, Sanmi Koyejo<sup>4</sup>, Yuntian Deng<sup>5</sup>,  
Shayne Longpre<sup>6</sup>, Noah Smith<sup>7,8</sup>, Beyza Ermiş<sup>1</sup>, Marzieh Fadaee<sup>◆1</sup>,  
and Sara Hooker<sup>◆1</sup>

<sup>1</sup>Cohere Labs, <sup>2</sup>Cohere, <sup>3</sup>Princeton University, <sup>4</sup>Stanford University, <sup>5</sup>University of Waterloo,  
<sup>6</sup>Massachusetts Institute of Technology, <sup>7</sup>Allen Institute for Artificial Intelligence, <sup>8</sup>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. 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 the arena distribution, based on our conservative estimates. 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.

---

<sup>\*</sup>First author. <sup>◆</sup>Principal senior advisors.

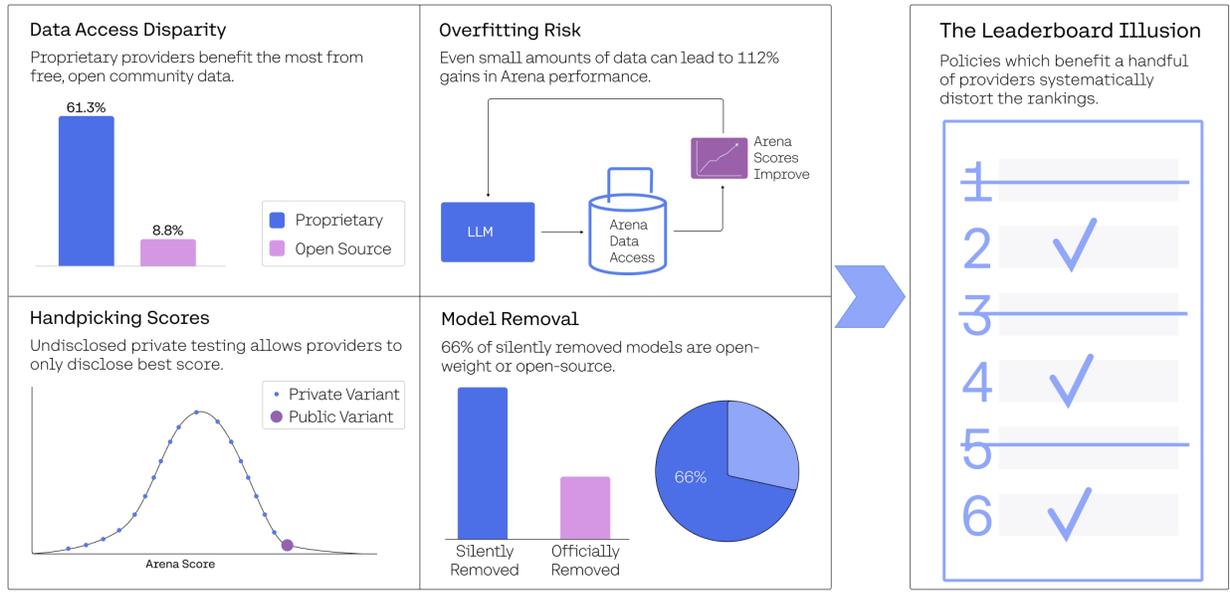


Figure 1: **Overview of key insights.** We investigate the prevalence of **undisclosed private testing and selective score reporting** on the Arena (Section 3), and highlight significant **data access disparities** between proprietary and open-source providers (Section 4.1). These disparities enable **overfitting to the Arena** (Section 4.2). Furthermore, **model deprecation practices** lack transparency, with many models silently deprecated without any notification to providers. We demonstrate how these deprecations contribute to unreliable rankings on the leaderboard (Section 5).

# 1 Introduction

*Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.* — **Charles A. E. Goodhart**

Benchmarks have long played an integral role in the development of machine learning systems (Church, 2018; Koch & Peterson, 2024), from the early days of NLP (e.g., TREC competitions and WMT shared tasks, Harman, 1993; Koehn & Monz, 2006) to the recent flurry of deep learning models (e.g., ImageNet, Deng et al., 2009), to the rise of large language models (e.g., GLUE, Wang et al., 2018). A meaningful benchmark demonstrates the relative merits of new research ideas over existing ones, and thereby heavily influences research directions, funding decisions, and, ultimately, the shape of progress in our field. The recent meteoric rise of generative AI models—in terms of public attention, commercial adoption, and the scale of compute and funding involved (Kaplan et al., 2020; Hoffmann et al., 2022; Samsi et al., 2023; Hooker, 2024)—has substantially increased the stakes and pressure placed on leaderboards (Orr & Kang, 2024).

Recently, Chatbot Arena has become the de facto standard for comparing generative AI models, holding enormous sway over media<sup>1 2</sup>, the AI industry (Maslej et al., 2025), and academia

<sup>1</sup><https://www.wsj.com/tech/ai/the-uc-berkeley-project-that-is-thewai-industrys-obsession-bc68b3e3>

<sup>2</sup><https://www.bloomberg.com/news/articles/2025-02-18/before-deepseek-blew-up-one-website-announced-its-arrival>

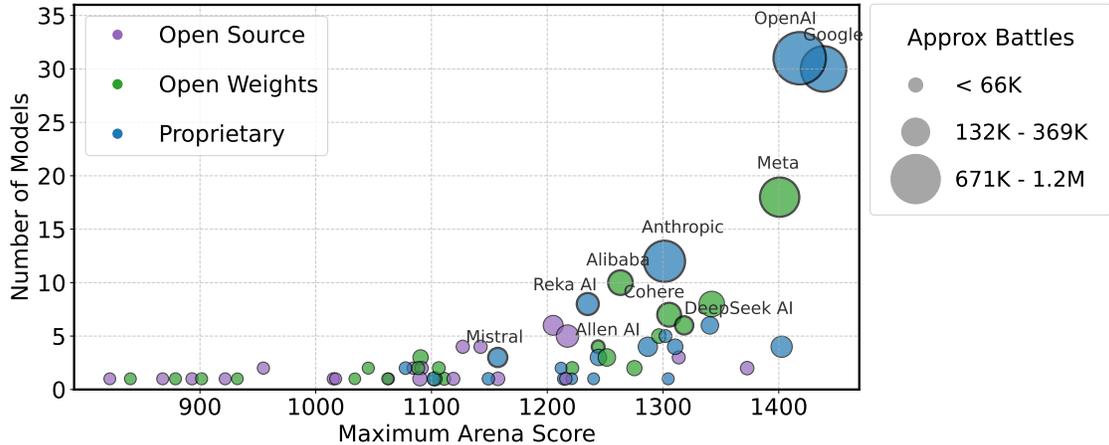


Figure 2: **Number of public models vs. maximum arena score per provider.** Marker size indicates total number of battles played. Proprietary model providers tend to achieve higher leaderboard scores, which appear to correlate with both the number of models they release and the number of Arena battles played. While model capability is an important factor, we explore in Section 3 and Section 4 how increased exposure to the Arena (through more models and battles) may confer additional advantages, such as better model selection or adaptation to the evaluation distribution. This figure summarizes publicly disclosed results as of April 23rd, 2025.

(Grattafiori et al., 2024; Yang et al., 2024). Created in 2023, Chatbot Arena rates LLMs by allowing anyone to submit a prompt and subsequently rank two anonymous responses from different models. Chatbot Arena quickly rose to prominence because it helped fill a critical gap in evaluation at a time of rapid technological change. As generative AI technology has grown more capable, academic multiple-choice evaluations (Hendrycks et al., 2021; Liang et al., 2022; Romanou et al., 2024; Singh et al., 2024b; Wang et al., 2024; Adelani et al., 2025; Salazar et al., 2025) have failed to reflect the real-world open-ended use cases these models are now being used for (Üstün et al., 2024; Chiang et al., 2024; Mizrahi et al., 2024). As a dynamic, user-driven evaluation framework where new questions can be asked daily, Chatbot Arena can, at least in principle, evolve alongside model capabilities. Furthermore, the open-ended nature of the interaction can capture emerging real-world use cases that differ from those initially envisioned by model creators.

However, the over-reliance on a single leaderboard creates a risk that providers may overfit to the aspects of leaderboard performance, without genuinely advancing the technology in meaningful ways (Ensmenger, 2011; Thomas & Uminsky, 2020; Raji et al., 2021; Bowman & Dahl, 2021). As Goodhart’s Law states, *when a measure becomes a target, it ceases to be a good measure*. In this work, we show that willful engagement from a handful of providers along with preferential policies from Chatbot Arena towards the same small group have amplified the potential for gamification, in place of innovative progress.

Our systematic review of Chatbot Arena involves combining data sources encompassing 2M battles, auditing 42 providers and 243 models across a fixed time period (January 2024 - April 2025). This comprehensive analysis reveals that over an extended period, a handful of preferred providers have been granted disproportionate access to data and testing. In particular, we identify an undisclosed Chatbot Arena policy that allows a small group of preferred model providers to test many model variants in private before releasing only the best-performing checkpoint. Using Arena data, we

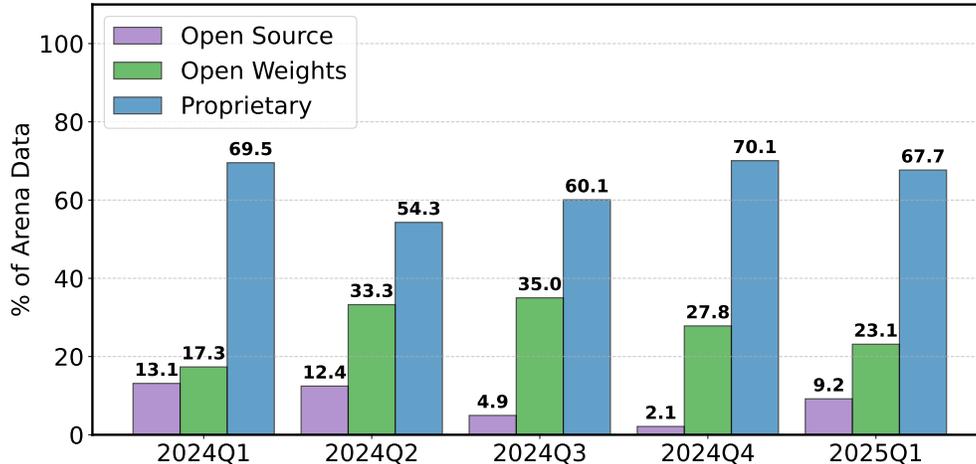


Figure 3: **Volume of Arena battles involving proprietary, open-weight, and fully open-source model providers from January 2024 to March 2025, based on leaderboard-stats.** Proprietary models consistently received the largest share of data—ranging from 54.3% to 70.1%. Open-weight and fully open-source models receive significantly less data, in some cases receiving less than half the amount of data as proprietary developers. This imbalance in data access exacerbates the performance gap, reinforcing unequal access to high-quality in-distribution data.

simulate and demonstrate how this strategy distorts overall arena ratings by beneficially skewing results. In addition, we find substantial data asymmetries are apparent regarding the amount of feedback data given to proprietary vs. open-weight vs. open-source model providers. This is due to a combination of unequal sampling rates (how often a model appears in a Chatbot Arena battle) and deprecation policies (which models are retired from the arena). We show that these differences lead to distorted and unreliable Arena rankings and create conditions in which providers may be overfitting to Arena-specific dynamics rather than general model quality. Across our extensive analysis, we present the following findings:

1. **Preferential treatment around private testing and retraction.** Chatbot Arena has an unstated policy of allowing select providers to test many submissions in parallel. We show that certain model developers (most notably Meta, Google, Open AI and Amazon) have benefited from extensive private testing. In a single month, we observe as many as 27 models from Meta being tested privately on Chatbot Arena in the lead up to llama 4 release. Notably, we find that Chatbot Arena does not require all submitted models to be made public, and there is no guarantee that the version appearing on the public leaderboard matches the publicly available API<sup>3</sup> <sup>4</sup>. We show with real-world experiments and simulations that the ability to select the best-scoring variant from  $N$  models enables systematic gaming of the Arena rating.
2. **Far more data is released to proprietary model providers.** Chatbot Arena is a community-driven leaderboard that benefits from free, crowdsourced feedback provided by everyday users. However, proprietary model providers collect significantly more test prompts and model battle outcomes than others. Google and OpenAI have received an estimated 19.2% and 20.4% of all test prompts on the arena, respectively. We estimate this based on the share of total battles played by the models of different providers on the Arena (as shown in Fig-

<sup>3</sup>[https://x.com/lmarena\\_ai/status/1909397817434816562](https://x.com/lmarena_ai/status/1909397817434816562)

<sup>4</sup><https://www.theverge.com/meta/645012/meta-llama-4-maverick-benchmarks-gaming>

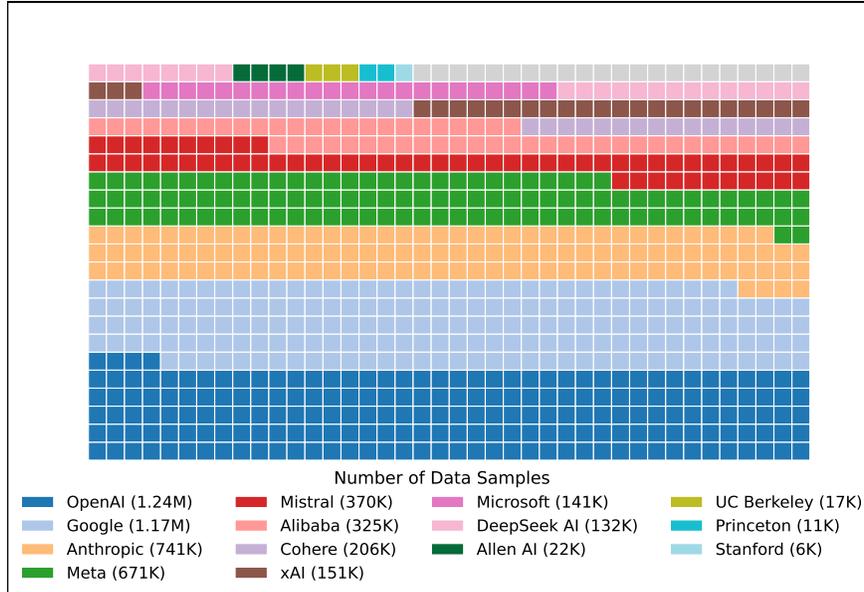


Figure 4: **Data availability to model providers.** We observe large differences in data access between providers, with 61.4% of all data going to proprietary providers.

ure 4). In contrast, a combined 41 fully open-source models have only received an estimated 8.8% of the total data, collectively.

- Chatbot Arena data access drives significant performance gains.** The differences in data access between providers matter; we estimate that by training on Chatbot Arena data, model ranking can be improved significantly. In a controlled experimental setting, we observe that increasing the arena training data proportion from (0%  $\rightarrow$  70%) more than doubles the win-rates from 23.5% to 49.9% on ArenaHard (Li et al., 2024c). We believe this is a conservative estimate, as a subset of providers have disproportionate access to private API data, which, if used, can potentially yield even greater performance gains.
- Deprecations can result in unreliable model rankings.** As shown in Figure 17, out of 243 public models, 205 have been silently deprecated. This is a significantly higher number than the 47 models officially listed as deprecated as part of Chatbot Arena’s backend codebase, FastChat<sup>5</sup>. We show that deprecation can violate key assumptions of the Bradley-Terry model (Bradley & Terry, 1952), which underpins Arena scoring, thereby reducing the reliability of the leaderboard rankings. Critically, we find that open-weight and open-source models are far more likely to be deprecated and, hence, receive unreliable ratings. Among the models that are silently deprecated, 66% are open-weight or fully open-source.

**Our role in both participating in the leaderboard and providing an overview of its limitations.** It is important to acknowledge that a subset of the authors of this paper have submitted several open-weight models to Chatbot Arena: `command-r` (Cohere, 2024), `command-r-plus` (Cohere, 2024) in March 2024, `aya-expense` (Dang et al., 2024b) in October 2024, `aya-vision` (Cohere, 2025) in March 2025, `command-a` (Cohere et al., 2025) in March 2025. We started this extensive study driven by this submission experience with the leaderboard. While submitting Aya

<sup>5</sup>[http://github.com/lm-sys/FastChat/blob/Oe6d3e4beaab66f4d3f93db72541a4abab8af28d/fastchat/serve/monitor/monitor\\_md.py#L7](http://github.com/lm-sys/FastChat/blob/Oe6d3e4beaab66f4d3f93db72541a4abab8af28d/fastchat/serve/monitor/monitor_md.py#L7)

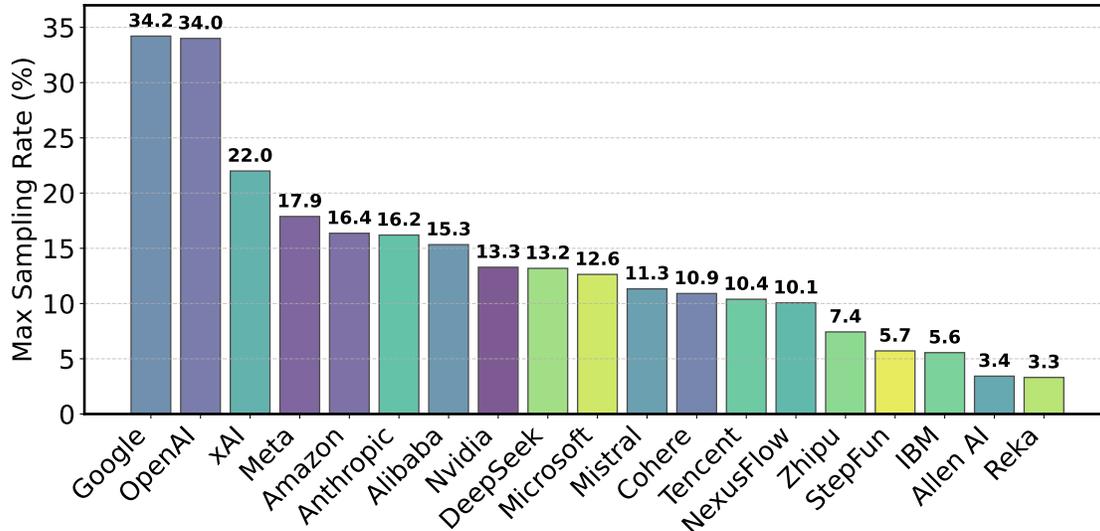


Figure 5: **Maximum observed sampling rate for models from different providers.** The sampling rate determines the amount of times a model is shown to everyday users, and the amount of data a provider receives. We observe large discrepancies across providers, with substantially higher sampling rates for OpenAI, Google, xAI, and Meta compared to others.

Expanse (Dang et al., 2024b) for testing, we observed that our open-weight model appeared to be notably undersampled compared to proprietary models — a discrepancy that is further reflected in Figures 3, 4, and 5. In response, we contacted the Chatbot Arena organizers to inquire about these differences in November 2024. In the course of our discussions, we learned that some providers were testing multiple variants privately, a practice that appeared to be selectively disclosed and limited to only a few model providers. We believe that our initial inquiries partly prompted Chatbot Arena to release a public blog in December 2024 detailing their benchmarking policy<sup>6</sup>, which committed to a consistent sampling rate across models. However, subsequent anecdotal observations of continued sampling disparities and the presence of numerous models with private aliases motivated us to undertake a more systematic analysis. As part of this analysis, we launched private variants to estimate the benefit of multiple submissions and retraction. We report on the results of this real-world study in Section 3.3. Consequently, Cohere appears in Figure 6 and Figure 9 as having private variants. Before this period, Cohere had not launched any private testing.

While our work points out significant limitations and unreliability issues with Chatbot Arena, it is important to acknowledge that Chatbot Arena originated as an academic-sponsored leaderboard that took on pronounced importance to the machine learning community. It takes considerable effort to coordinate a leaderboard, and arena organizers have often faced large demands on their time from supporting the addition of models from many different model providers. Hence, while we point out systematic issues with Chatbot Arena as it is now, we also believe some of these issues may have emerged gradually as successive design choices and concessions to certain large providers accumulated and introduced unreliability. We believe that there are reasonable interventions that the organizers can do to restore scientific accuracy and renew trust in the leaderboard. We have directly shared our recommendations with Chatbot Arena prior to releasing this work. We include these here and also share in Section 6 for a complete discussion.

<sup>6</sup>[https://drive.google.com/file/d/1reook2cjwq81xD6Yn528KOLWeWRy0ZvN/view?usp=drive\\_link](https://drive.google.com/file/d/1reook2cjwq81xD6Yn528KOLWeWRy0ZvN/view?usp=drive_link)

---

## Critical Recommendations to Restore Trust to the Chatbot Arena

In order to restore fairness, transparency and trust to the Chatbot Arena, we determine that the following five changes are both actionable, easy to implement as well as **urgent and necessary for restoring scientific integrity**:

- 1. Prohibit score retraction after submission.** All model evaluation results—including private variants—must be permanently published upon submission, with no option to retract or selectively hide scores. This is a crucial and necessary requirement to restore trust in the leaderboard. There is no reasonable scientific justification for allowing a handful of preferred providers to selectively disclose results. Indeed, we show that this skews Arena scores upwards and allows a handful of preferred providers to gamify the leaderboard.
- 2. Establish transparent limits on the number of private models per provider.** There are valid reasons to offer private testing – as it allows the Chatbot Arena community to test models in development while also preserving the anonymity required before a release. However, some providers are engaging in private testing *far* more than others. This has implications for gamification, but it also means preferred providers benefit from substantially more data from an open community project. At the time Meta launched 27 variants in the lead up to Llama 4 release. This highly skews data access from a community resource. It is critical to enforce a strict and publicly disclosed limit (e.g., maximum 3 model variants tested concurrently per provider). This prevents excessive testing runs that skew the leaderboard while ensuring fair, transparent benchmarking for all. If private testing continues, it should have transparent limits and be paired with disclosing all scores.
- 3. Ensure model removals are applied equally to proprietary, open-weights, and open-source models.** We find deprecation disproportionately impacts open-weight and open-source models, creating large asymmetries in data access over time (Section 4.1). We strongly urge arena organizers to remove an equal share across proprietary, open-weight, and open-source licenses to avoid creating asymmetric access to data that disadvantages open-weight providers. For example, a reasonable and transparent policy would be to deprecate the bottom 30th percentile for each group of proprietary, open-weights, open-source models.
- 4. Implement fair sampling.** Our work finds large biases towards proprietary models in sampling rates. For example, in Figure 5, we observe a maximum daily sampling rate of models from OpenAI or Google of up to 34%, which is 10 times more than what is observed as the maximum sampling from providers like Allen AI. We suggest a return to an active sampling method proposed by the arena organizers themselves (Chiang et al., 2024), which appears not to be respected or implemented in practice. This formulation avoids simply favoring large proprietary providers and instead effectively prioritizes under-evaluated and high-variance pairs. This avoids preferring a subset of providers and instead focuses votes to align sampling with the goal of rapidly reducing uncertainty in rankings.
- 5. Provide transparency into what models are being removed from the leaderboard.** We found that a large number of models are silently deprecated without being added to the list of officially deprecated models. While 47 out of 243 public models are officially listed as deprecated, we measure that 205 have been silently removed by the organizers without informing the provider. For transparency and to ensure that the deprecation policy is implemented fairly, it is important that there is a comprehensive list of what models have been removed from the leaderboard.

---

## 2 Overview of Methodology

*To measure is to know* — Lord Kelvin

Leaderboards have played a pivotal role in the computer science field in driving technological progress (Koch & Peterson, 2024). A desirable goal of a general benchmark is to reliably rank models according to real-world capabilities and preferences. Hence, if one system significantly outperforms another, then the result should be strong evidence that the higher-ranked system is better at the task (Bowman & Dahl, 2021; Rodriguez et al., 2021). If a leaderboard is not representative of real performance gains, it can lead practitioners and researchers towards less impactful areas of intervention and inquiry. In some cases, this misalignment enables *gaming* of the system—where participants optimize for leaderboard metrics rather than real-world utility. As a result, the gap between reported performance and actual value in deployment can continue to widen. Ultimately, the goal of our work is to ask whether the rankings of Chatbot Arena are a *reliable* ranking of generative AI models. To answer this, across the sections that follow, we will consider and answer the following questions:

1. **Private Testing and Selective Disclosure (Section 3)** What is the impact of the undisclosed policy of private testing and selective retraction on Arena scores?
2. **Data Access Asymmetries (Section 4.1)** What factors lead to large asymmetries between providers in data access on the Chatbot Arena?
3. **Risk of Overfitting (Section 4.2)** How do asymmetries in data access impact Arena scores?
4. **Model Deprecation (Section 5)** What is the impact of Chatbot Arena model deprecation policy on Arena score reliability?

To gain insights and analyze various trends in the Chatbot Arena leaderboard, we leverage multiple data sources. In total, our real-world data sources encompass 2M battles and cover 243 models across 42 providers. We present the overview of each data source in Table 1 and describe the different datasets in more details in Appendix D. Across our analyses, we group models by the licenses reported on Chatbot Arena into proprietary, open-weights, and open source to understand overall trends. This allows us to gain insight into who benefits the most from the Chatbot Arena which is a free and open community resource. We report details of what licenses are grouped into each category in Appendix F. We note that many open-weights models are extremely performant and include Aya family of models (Dang et al., 2024b; Üstün et al., 2024), Meta models (Meta, 2025) and Gemma family of models (Team et al., 2025).

### 2.1 Introduction to the Arena Score

The Arena score is the key metric used in Chatbot Arena to rank models based on how humans judge their performance. In Arena, users engage in pairwise comparisons, voting on which of two models performs better (or declaring a tie) in a given interaction. A naive way to compute rankings would be to average each model’s win-rates, simply counting how often they win. However, this approach does not consider the strength of the opponents and requires all models to play against

Table 1: A summary of datasets we constructed, their sources, and the research questions they enabled us to answer. These datasets can be of one of the following types: **battles only** () , **conversations only** () , **battles with conversations**, () and **leaderboard updates** () . Depending on the dataset type, it either **contains prompts** () or doesn't () . Accessibility of the datasets is indicated using **public** () or **private** () .

Name	Fields	Source		Type	Prompts?	Size	Period
Historical Battles	battle dates, category & language tags	Arena-human-preference-100k <sup>7</sup>				100K	04-23 - 01-25
		Colab data <sup>8 9</sup>				1.9M	
		LMArena, Cohere				43K	
Random Scraped Battles	model identity responses, battle players	Crawled				5.8K	01-25 - 03-25
API prompts	prompts	Cohere				197K	11-24 - 04-25
Leaderboard Statistics	ratings, dates, models, battles counts, licenses, providers	HuggingFace Leaderboard Commit History <sup>10</sup>				14.3K	01-24 - 04-25

each other. Furthermore, averaging win-rates treats all wins equally, which can lead to misleading or unfair rankings, especially when models face opponents of varying strength levels.

Instead, Chatbot Arena ranks models based on their Arena Score, which is a normalized version of the Bradley-Terry (BT) model (Bradley & Terry, 1952), a probabilistic framework for estimating skill levels based on pairwise comparisons. BT is a widely used model to estimate skill levels in diverse fields such as online gaming (e.g., StarCraft II) (Leitner et al., 2010; Liquipedia, 2019; Wise, 2021, i.a.) and sports rankings (Glickman, 1999; Hunter, 2004). Compared to alternative systems like Elo (Elo, 1978; Boubdir et al., 2024), which was initially used in the first version of Chatbot Arena, BT provides a more robust and statistically grounded basis for ranking. BT produces well-calibrated estimates while naturally accommodating ties and missing comparisons, and providing confidence intervals for the rankings as long as the underlying model conditions are met.

While BT is considered to be a principled ranking system, its effectiveness depends on several key assumptions. Background on these assumptions will provide important context for our discussions in Section 5.1 and Section 5.2, where we examine how violations of BT’s assumptions affect the reliability of Arena rankings. The BT model assumes that pairwise comparisons are drawn from an **unbiased sampling process**, so that each model’s skill parameter is estimated independently and based on representative outcomes. It also does not require all players to play each other because of **transitivity** of rankings—if Model A is judged to beat Model B and Model B in turn beats Model C, then Model A is estimated to outperform Model C (even if they never actually play). Finally, it assumes that the comparison network is **fully interconnected**—that is, every model must be linked directly or indirectly through pairwise matchups. Deviations from these assumptions, such as biased comparisons or a fragmented graph, can compromise the reliability and consistency of the strength estimates. For further details, we include the mathematical formulation of the BT model

---

in Appendix B.

### 3 Results: Impact of Private Testing and Selective Retraction on Arena Scores

#### Section Findings

- **Undisclosed testing policy permits preferred providers to bias reporting of results.** Chatbot Arena currently permits a small group of preferred providers to test multiple models privately and only submit the score of the final preferred version. As observed in Figure 6, we observe that Meta, Google, and Amazon were key beneficiaries of this policy for the period for which we collected data.
- **Selective disclosure violates unbiased sampling assumption of BT model.** We show theoretically how BT assumptions are violated, which systematically inflates model rankings, distorting the leaderboard ranking. In Figure 7, we simulate the expected distortions to leaderboard Arena rankings and show that testing just 10 variants yields a notable increase of approximately 100 points in the maximum score identified.
- **Real world experiments on the Chatbot Arena confirm benefits of selective reporting of results.** We corroborate our simulations with experiments on the real arena by deploying multiple private variants. We show that even limited numbers of variants lead to large gains in Arena Scores. Submitting multiple model variants to the Chatbot Arena can lead to a systematic advantage in rankings, even when the underlying models are identical or only marginally different.

#### 3.1 Preferred Providers Frequently Use Private Testing

Although not an officially stated policy<sup>6</sup>, our audit of Chatbot Arena data using `random-sample-battles` revealed that providers are permitted to test multiple private model variants simultaneously, without being required to publicly release or de-anonymize these submissions. In Figure 6, we plot the number of private variants we tracked as belonging to each provider from January to March 2025. We observed that Meta and Google had 27 and 10 private models, respectively, which preceded the releases of Llama 4 (Meta, 2025) and Gemma 3 (Google, 2025) models shortly thereafter. We note this is likely a very conservative estimate as it only tracks the private variants on the main Chatbot Arena, and does not take into account private variants on specialized leaderboards run by Arena such as for vision<sup>11</sup> or code.<sup>12 13</sup> Indeed, if we also consider the number of private models tested by Meta on the vision leaderboard, we observe an additional 16 variants, bringing the total to 43. In contrast, smaller startups, such as Reka, were found to have one active private variant live in the arena. We note that during the same period, Cohere submitted 4 private variants—these ablations were part of experiments measuring the lift that could be expected from private testing that we detail in the experiments in Section 3.2 and Section 4.1. Before this, Cohere had never submitted private variants for testing in the Chatbot Arena.

<sup>11</sup><https://blog.lmarena.ai/blog/2024/multimodal/>

<sup>12</sup><https://blog.lmarena.ai/blog/2024/copilot-arena/>

<sup>13</sup><https://blog.lmarena.ai/blog/2025/repochat-arena/>

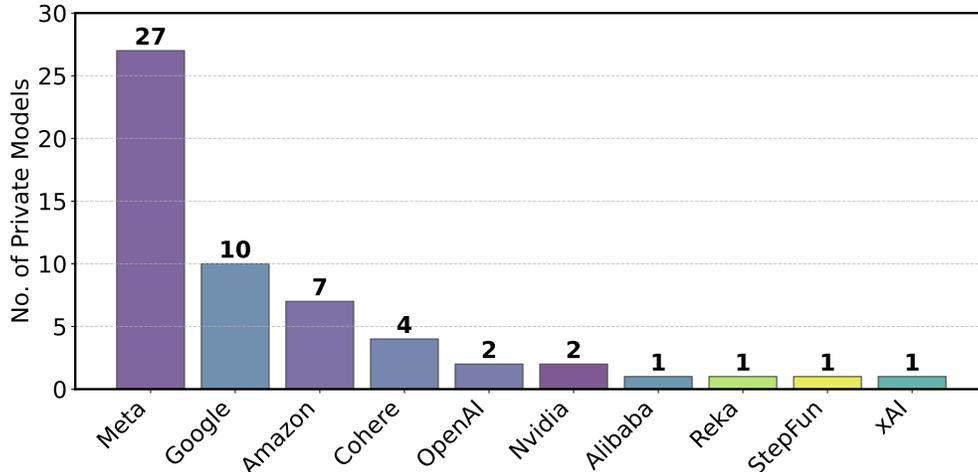


Figure 6: **Number of privately-tested models per provider based on random-sample-battles (January–March 2025)**. Meta, Google, and Amazon account for the highest number of private submissions, with Meta alone testing 27 anonymous models in March alone.

We only scraped data from January to March 2025, yet we anecdotally observed behavior that suggests submitting multiple variants was a long-standing practice amongst a subset of providers. Over the last year, we have observed that major LLM providers such as Google, xAI, and OpenAI are often announced as having the top-performing variant within just a few days of one another. For example, OpenAI’s GPT-4.5 and xAI’s Grok-3 reached the top of the Chatbot Arena leaderboard within the same day (March 4, 2025)<sup>14</sup> <sup>15</sup>. Gemini (Exp 1114) from Google DeepMind reached the top of the leaderboard on November 14, 2024<sup>16</sup> and shortly after, ChatGPT-4o (20241120) from OpenAI claimed the top position on November 20, 2024<sup>17</sup>. Just one day later, on November 21, 2024, Gemini (Exp 1121) regained the top spot<sup>18</sup>. Given the time typically required to develop, refine, and test a foundation model, it is unlikely for the same provider to top the leaderboard twice in a single week unless they were testing multiple variants simultaneously. In Section 3.2, we demonstrate through simulated experiments that rapid leaderboard turnover can plausibly emerge from providers optimizing for the highest possible score by testing multiple model variants in parallel.

Notably, we found that no private models were tested by academic labs during the observed period. This leads us to believe only certain providers were made aware they could submit multiple private variants, as we observe clear differences in the number and frequency of private testing among providers.

### 3.2 Simulated Experiments on Private Testing and Retraction

Private testing coupled with the option to retract enables a best-of-N strategy, where an organization submits multiple model variants to Chatbot Arena, privately evaluates them, and retains only the top-performing variant to be publicly published on the leaderboard. In this section, we show that

<sup>14</sup>[https://x.com/lmarena\\_ai/status/1896675400916566357](https://x.com/lmarena_ai/status/1896675400916566357)

<sup>15</sup>[https://x.com/lmarena\\_ai/status/1896590146465579105](https://x.com/lmarena_ai/status/1896590146465579105)

<sup>16</sup>[https://x.com/lmarena\\_ai/status/1857110672565494098](https://x.com/lmarena_ai/status/1857110672565494098)

<sup>17</sup>[https://x.com/lmarena\\_ai/status/1859307979184689269](https://x.com/lmarena_ai/status/1859307979184689269)

<sup>18</sup>[https://x.com/lmarena\\_ai/status/1859673146837827623](https://x.com/lmarena_ai/status/1859673146837827623)

best-of- $N$  submissions violate the BT unbiased sampling assumption. This systematically inflates model rankings and distorts the leaderboard ranking.

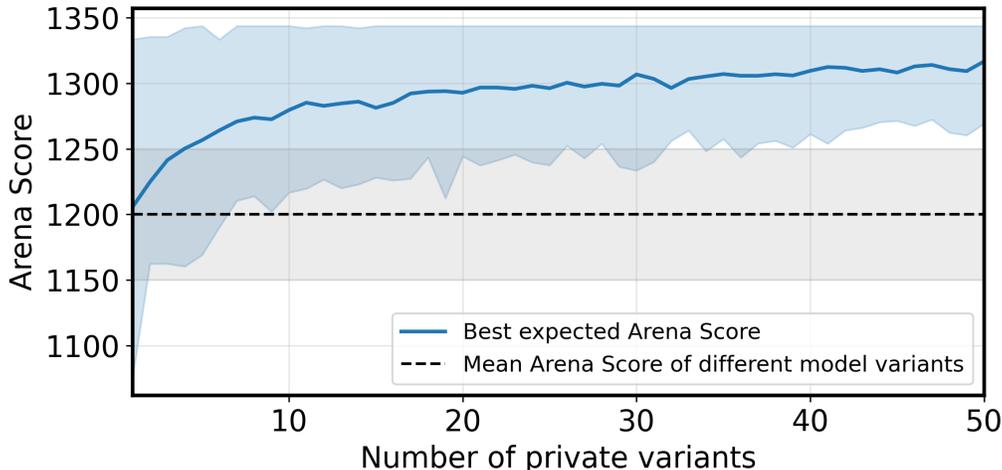


Figure 7: **Impact of the number of private variants tested on the best Expected Arena Score.** We simulate a family of model variants with a latent average Arena Score of 1200. As we progressively increase the number of private variants tested—and subsequently discover their corresponding Arena Scores—the probability of selecting models from the higher end of the performance distribution also rises. This enables the provider to effectively identify the model with the highest score.

**Unbiased Sampling Assumption.** To study the selection bias scenario, assume a provider submits  $N$  variants of a model, each variant  $k$  having a true underlying skill parameter  $\beta_k$ , sampled from a distribution centered at some base skill level  $\beta$ . The probability of observing an exceptionally high-performing variant increases with the number of submissions  $N$ . Thus, the observed skill of the submitted model is:  $\hat{\beta}_{\text{Best}} = \max\{\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_N\}$ .

Since each  $\hat{\beta}_k$  is subject to statistical fluctuation due to finite match sampling, selecting the best variant based on observed performance introduces an upward bias. Specifically, the expected value of the best-performing variant is *strictly greater than* that of a regular submission:

$$\mathbb{E}[\hat{\beta}_{\text{Best}}] > \mathbb{E}[\hat{\beta}_k], \quad \forall k \in \{1, 2, \dots, N\}. \quad (1)$$

where the draws are *non-degenerate* ( $\text{Var}(\hat{\beta}_k) > 0$ ) and  $N \geq 2$  (See Appendix C for further details). This violates the BT model’s assumption of unbiased sampling and alters the likelihood landscape. The reported rating no longer reflects a single, unbiased estimate of skill, but an extreme value from multiple independent estimations. As a result, the BT estimator systematically inflates the ratings of models submitted under the best-of- $N$  strategy, distorting leaderboard rankings.

**Role of the number of private variants.** In the Chatbot Arena, we observe that while only a handful of providers know about private testing, there is also an asymmetry in the *number of private models* tested. At an extreme, Meta tested 27 private variants in the lead up to Llama 4 launch. To investigate the effect of varying the number of private tests on model selection outcomes, we simulate the process of identifying the highest-performing model within a family of variants. We

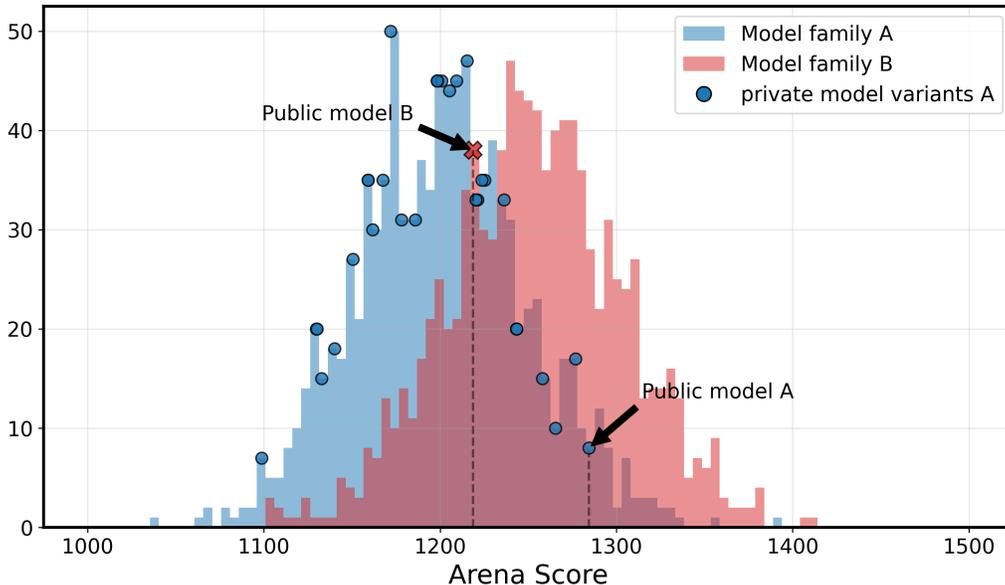


Figure 8: **Simulated impact of best-of-N submission strategies on Arena leaderboard rankings.** **Model family A** has a lower average Arena Score than **Model family B**, yet by submitting multiple private variants and selecting the best-performing one, it can surpass the sole public submission from **Model family B**. This simulation shows how providers can gain leaderboard advantage by evaluating several variants privately and publishing only the top-scoring one.

assume that the true Arena Scores of the model variants are drawn from a normal distribution with a mean of 1200. For each value of  $n \in \{1, \dots, 50\}$ , we perform 100 simulations in which  $n$  model variants are independently sampled from the distribution. The variant with the highest score among the sampled set is selected to represent the model family. Figure 7 presents the results, plotting the number of private variants tested (x-axis) against the average of the maximum discovered Arena Scores across simulations (y-axis). As the number of privately tested models increases, the expected maximum Arena Score discovered rises accordingly. In our simulation, testing just 10 variants yields a notable increase of approximately 100 points in the maximum score identified, compared to the baseline scenario with no private testing.

**Asymmetries in which providers have access to private testing.** We observe in practice that only a few preferred providers were allowed to test many variants and handpick the best result. As we show in Figure 8, restricting private testing to a subset of providers can lead to scenarios where a weaker family of models (Family A) enabled with private testing can outperform a stronger family of models (Family B), which is restricted to a single submission. Although the performance of both model families is in a similar range, Family A’s models have a lower average Arena score across all models compared to Family B’s. In contrast to model provider B, who is unaware of the best-of-N strategy, model provider A evaluates multiple models on the Chatbot Arena distribution and selects the best-performing model. This allows provider A to evaluate multiple models on the Chatbot Arena distribution and select the best-performing model, leveraging the tail of the distribution, to achieve a higher leaderboard ranking. As a result, despite having a generally stronger model pool, Family B ranks lower than Family A on the leaderboard.

Having to pick from multiple final “candidate” models is a very common scenario during the development of LLMs. At the end of the development process of a new model, a model provider typically ends up with multiple variants, each excelling in different tasks due to variations in post-training strategies or hyperparameter settings. They often perform well on different tasks, while still being within similar overall performance bands. Selecting a final “official” model often involves a compromise across a broad range of evaluation sets. A strong signal—such as a clear lead on an arena-style leaderboard—can significantly influence this decision, tipping the preference toward specific variants that perform best under the given evaluation framework. Hence, with an “informed selection” strategy for model variants, a model provider can improve its ranking and reach the top of the leaderboard compared to when it makes an “unguided selection” to choose the variant that can be released on the public leaderboard.

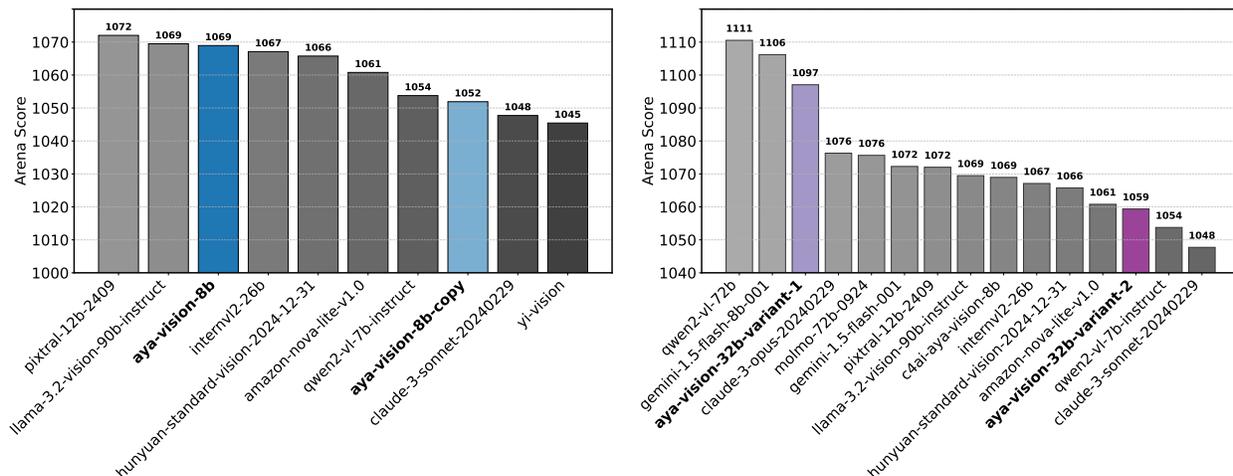


Figure 9: **Allowing retraction of scores allows providers to skew Arena scores upwards.** We run a real-world experiment to measure the benefits of private testing. We show that it is possible to increase Arena scores even in the most conservative case of identical checkpoints, and further amplify the difference by strategically testing different checkpoints. **Left: Identical Checkpoints.** Arena Scores for Aya-Vision-8B yield different Arena scores (1069 vs. 1052). **Right: Strategically Selected Checkpoints.** Arena Scores for two different variants of Aya-Vision-32B, which were both considered high-performing final round candidates according to internal metrics. We observe large differences in final scores (1097 vs. 1059) for the two different model variants.

### 3.3 Real-world Chatbot Arena Experiment

Our simulation previously demonstrated that when model submissions are not standardized across model providers, those who submit a larger number of private model variants gain a systematic advantage, creating an uneven playing field for ranking. Since we do not have access to the final scores of private model variants observed during our Arena scrape, we design a real-world experiment to complement and validate our simulation findings. Specifically, we model two scenarios of best-of-N testing to assess the impact of submitting multiple model variants.

**Lower bound estimate of benefits of multiple submissions.** We model the most conservative case – measuring the gain from submitting and then only releasing the best Arena score obtained by *two identical checkpoints*. For this purpose, we submitted two identical variants of Aya-Vision-8B to the Arena in March 2025. This scenario is a very conservative lower bound of gains, as any

---

difference in Arena score can be attributed to the benefits of multiple submissions rather than differences in model quality. We used only two checkpoints because we wanted to limit the amount of human annotator time we diverted to this exercise. We did not disclose that these were identical checkpoints to Chatbot Arena. In Figure 9, we observe notable differences in final Arena scores between identical checkpoints (1052 vs. 1069). As a result, 4 models are positioned between our two identical checkpoints. This suggests that even with identical variants, it is very feasible to obtain a biased positive advantage.

**Realistic estimate of benefits of multiple submissions.** In practice, a provider would likely optimize the differences in the variants they are testing to maximize the signal from the Arena users. To study this scenario, we compare two different variants of **Aya-Vision-32B**. The two checkpoints are variants of the same model, with each showing slightly better performance on different subsets of benchmarks. In Figure 9, we illustrate the extreme ends of the scores obtained by the two models (1097 vs. 1059), with 9 models falling in between both variants on the leaderboard.

## 4 Results: Impact of Data Access Asymmetries on Arena Scores

### Section Findings

- **Extreme disparities exist in access to data from Chatbot Arena.** These disparities stem from difference in the **number of private variants tested, sampling rate, and silent deprecation**. Private testing is an undisclosed policy, and we find that sampling rate and silent deprecation do not reflect the publicly stated policy of Chatbot Arena. All these policies appear to severely disadvantage open-weight and open-source providers. For example, OpenAI, Google, Meta, and Anthropic collectively account for 62.8% of the Arena data, which is 68 times more than the combined share of top academic and non-profit labs like Allen AI, Stanford, Princeton, and UC Berkeley.
- **Access to Chatbot Arena data has an outsized impact on performance.** We fine-tune models under a fixed training budget, varying the proportion of arena data relative to non-Arena instruction-following samples, and observe a substantial increase in performance. We show that incorporating a moderate amount of arena data greatly improves performance on ArenaHard (Li et al., 2024b) with estimated gains of 112%, while showing limited benefits for other tasks of interest. These results represent a conservative estimate of gains given the limited access to arena-style data. Larger labs and companies likely have access to significantly more such data, which, if used, could yield even greater gains.

### 4.1 Disparity in access to Chatbot Arena Data

Prompts from a large and diverse user base, such as those from Chatbot Arena users, serve as a valuable signal for modeling user interests and preferences. This data is often accessible to model providers through API calls originating from Chatbot Arena battles. Multiple factors impact the amount of the data obtained by a given provider, some of which are determined by Chatbot Arena versus others which are within the control of the providers:

1. **Number of private variants being tested on the arena:** As shown in Figure 6, some

---

providers deploy far more private variants, which can significantly increase the volume of data collected. We note that even with our experiment of launching multiple model variants, we increased the amount of prompts collected from 5.9% with 1 variant to 19.4% with 3 variants. Based on findings from Figure 6, the number of variants submitted is not uniform across all providers, and some providers may increase variants to further amplify the volume of data collected. This is of particular concern given Chatbot Arena is a community-driven leaderboard, however, the main beneficiaries of this free human feedback appear to be commercial entities who are frequently preferred for private testing.

- 2. Sampling rate applied to provider models:** We define model sampling rate as the percentage of daily battles a model participates in. The maximum sampling rate for a provider is the highest rate achieved by any of its models on any given day. We determine the maximum sampling rate of providers based on `scraped-random-sample`, which is limited to the specific period during which we collected this data (January 2025 to March 2025). As shown in Figure 5, sampling rates vary significantly across providers. These rates are determined by Chatbot Arena, but are often entirely inconsistent with the stated policy and prior proposals by the organizers to automatically set sampling based upon which models have not converged in score (Chiang et al., 2024). At the extreme, Google and OpenAI reach a maximum daily sampling rate of 34%, while Reka registers the lowest at 3.3%. Other providers with relatively high sampling rates include xAI (22.0%) and Meta (17.9%), highlighting substantial variation across the board. We provide additional details about how sampling rates were determined for each provider in Appendix E.5.
- 3. Number of models publicly hosted on the arena:** A model only receives traffic if it is live on the arena. However, Chatbot Arena frequently deprecates models. There are several reasons to deprecate models in a benchmark. Chatbot Arena may be forced to deprecate a model when a provider no longer supports it via its API. They also have policies for deprecating models under certain conditions<sup>6</sup>: *Models may be retired after 3000 votes “if there are two more recent models in the same series and/or if there are more than 3 providers that offer models cheaper or same price and strictly better (according to overall Arena score)”*. We note that the logic of this policy is difficult to audit in practice because many models are hosted for free on the Chatbot Arena, and the use of the “or” condition means it is not clear what criteria (price or quality) applies to decisions. We observe that many models are also silently deprecated, which means their sampling rate is reduced to nearly 0% without notification, even though some of them do not meet the stated criteria of the deprecation policy. We identify 205 models that have been silently deprecated, a number that substantially exceeds the 47 models officially marked as deprecated by Chatbot Arena. For a more detailed analysis, see Appendix I.
- 4. API Support for Models on the Arena:** Developers who deploy a model and enable Chatbot Arena testing via an API have a default advantage. This allows providers to collect 100% of the test prompts submitted on the Arena. In contrast, providers whose models are hosted by a third party are often limited to publicly accessible data or must request access to only 20% of the data (including prompts and human preferences) involving their models from Chatbot Arena, as per their policy<sup>6</sup>.

We observe that the collective impact of these factors appears to be advantageous to a small handful of providers and is often inconsistent with the stated policy. For example, in Figure 4, we show that the combined share of OpenAI, Google, Meta, and Anthropic alone is 62.8% of the arena data,

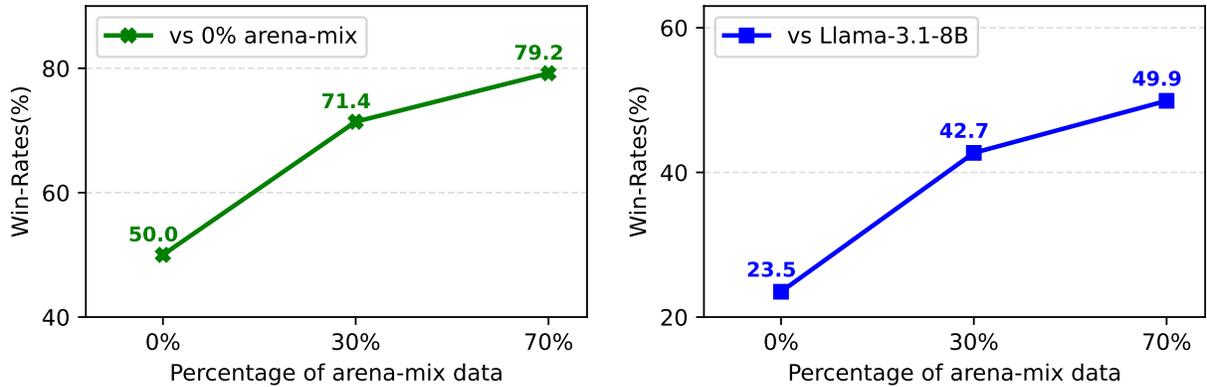


Figure 10: **Use of Chatbot Arena dataset significantly improves win-rates on ArenaHard.** Increasing the amount of arena data in a supervised fine-tuning mixture (0%  $\rightarrow$  30%  $\rightarrow$  70%) significantly improves win-rates of the resulting model against both the model variant where no Chatbot Arena data is used and also Llama-3.1-8B. The win-rates are measured on ArenaHard (Li et al., 2024c), which has a high correlation of 98.6% to Chatbot Arena.

which is 68 times more than the share of top academic and non-profit labs including Allen AI, Stanford, Princeton, and UC Berkeley. These findings add to prior works that consistently show better corporate access to AI training data across the ecosystem (Longpre et al., 2024c;b).

We note that the prompt samples available to each provider may not be mutually exclusive, as each battle on the Arena involves two models (from the same or different model providers), allowing the same prompt to be sent to two different providers. Details about statistics in Figure 4 are available in Appendix G.

## 4.2 Risk of Potential Overfitting

One of the questions we want to answer is “*What are the implications of data asymmetries? Does having access to more data enable overfitting to the Arena?*”. Overfitting refers to a phenomenon in which a model learns not only the generalizable patterns in its training data but also the specific noise or characteristics, resulting in strong performance on familiar or seen examples but degraded performance on unseen and out-of-distribution inputs.

Overfitting is a particularly salient challenge in static evaluation settings, where fixed test sets are vulnerable to overfitting through repeated exposure, data contamination, or intentional tuning (Deng et al., 2024; Golchin & Surdeanu, 2024; Roberts et al., 2023; Dong et al., 2024; Singh et al., 2024a). In contrast, Chatbot Arena has been widely adopted in part because it allows human users to ask any questions they want, which creates a non-static test set (Don-Yehiya et al., 2024). This means that, at least in principle, the test set is potentially harder to overfit to. However, this assumes that the distribution is constantly changing.

To understand whether this is the case with data from Chatbot Arena, we do an exhaustive analysis and observe that the true picture on Chatbot Arena is more complex. We observe two phenomena: (1) The characteristics of prompts do shift notably over time, and (2) a non-trivial portion of

prompts in one month are either exact duplicates or near-duplicates of prompts from previous months. This means that some test prompts are very similar to (or even the same as) prompts seen previously. Together, these suggest that having access to a large sample of last month’s data will enable a developer to perform significantly better on next month’s test set.

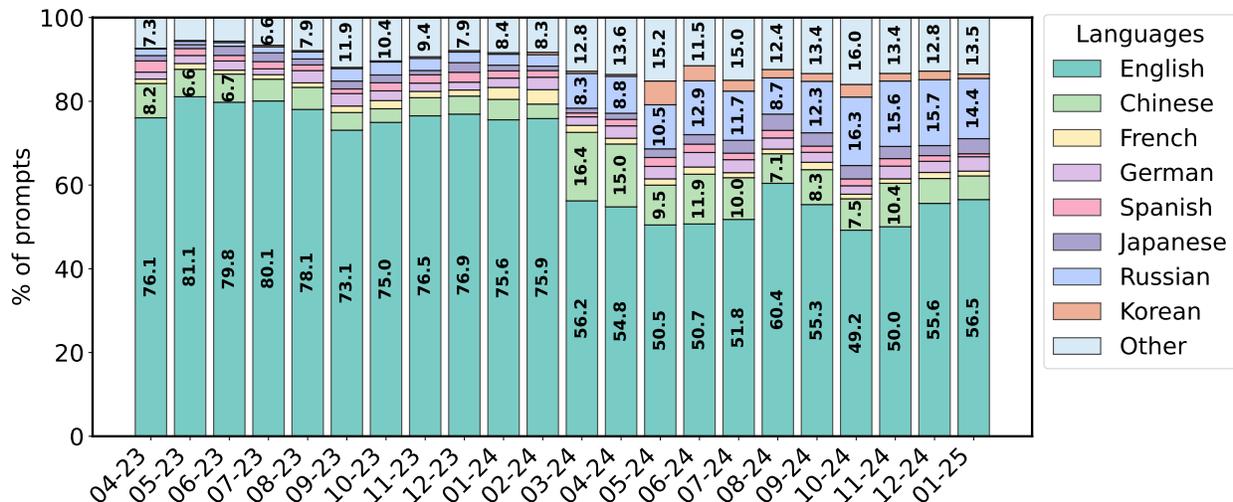


Figure 11: **Language distribution of prompts submitted to Chatbot Arena from April 2023 to January 2025.** Based on the `historical-battles` dataset, this figure tracks the monthly share of prompt languages. Only languages with dedicated Chatbot Arena leaderboards are shown individually; the rest are grouped under “Other”. A clear shift is observed: English prompt share dropped from over 80% to nearly 50%, while usage of Chinese, Russian, and Korean prompts increased significantly.

**1) Long-term distribution shifts.** Prior work clearly demonstrates how temporal distribution shifts affect performance (Luu et al., 2021; Longpre et al., 2024d). On Chatbot Arena, notable shifts have been observed in the distribution of prompts evaluated over longer periods, with a consistent increase in the proportion of prompts from more complex categories, such as mathematics, coding, and multi-turn conversations<sup>19</sup>. We also perform our own analysis of the change in language distribution in the Arena based on the “language” tag available as part of `historical-battles` dataset. For example, in Figure 11, we observe that the proportion of languages outside of English has varied over time. For instance, the share of Russian prompts increased from 1% in April 2023 to 8.8% in April 2024, and further to 15.7% by December 2024. Chinese prompts more than doubled from 5-7% in 2023 to 16.4% in March 2024, coinciding with the introduction of the Chinese leaderboard on Chatbot Arena, before dropping back to 6.2% in January 2025. Overall, the number of multilingual prompts on the Arena has grown by 20% over 1.5 years, from 23.9% in April 2023 to 43.5% in January 2025. This indicates increased language diversity in submitted prompts.

**2) Prompt redundancy and duplication.** In parallel, we observe high levels of prompt duplication. We analyze a proportion of raw API calls we receive from Chatbot Arena between November 2024 and April 2025 (197,217 single-turn conversations). We switch to this source given that the proprietary data Chatbot Arena releases are already de-duplicated, and so won’t capture the extent of similar or overlapping queries. Between November 2024 and April 2025, de-duplication resulted

<sup>19</sup><https://blog.lmarena.ai/blog/2024/arena-category/>

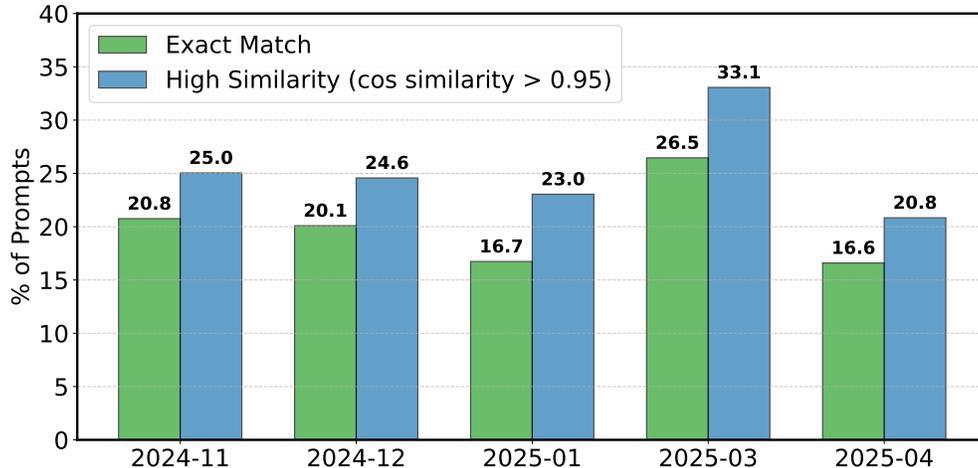


Figure 12: **Monthly prompt duplication rates.** Prompts are from November 2024 to April 2025, excluding February 2025 due to insufficient data. Duplication is measured using two similarity metrics: *Exact Match* and *High Similarity* (cosine similarity of text embedding  $> 0.95$ ). For simplicity, this analysis is limited to single-turn conversations. The chart presents the percentage of battles in which duplicate or near-duplicate prompts were detected each month.

in an average prompt loss of 20.14%, peaking at 26.5% in March 2025 (Figure 12). While prompt distribution changes over time, prompts in one month often serve as a proxy for the next. For instance, 7.3% of prompts from December 2024 appear again in the exact form in January 2025. If we relax the condition and consider high semantic similarity of prompt embeddings (using the `embed-multilingual-v3.0` model<sup>20</sup>), the same cross-month duplication rate increases to 9%. Detailed cross-month duplication statistics can be found in Appendix H. Both trends above suggest that 1) sustained access to up-to-date prompt data and 2) the volume of sampled prompts in a given month offer a significant competitive advantage in predicting performance in subsequent months.

**Uniqueness of Arena Data.** One reason providers may be motivated to explicitly optimize for Chatbot Arena distribution is if it differs substantially from other evaluation settings that providers may care about. There is sufficient signal to suggest this is the case. There is a context length limit of 12000 characters on Chatbot Arena prompts, which excludes certain types of longer or more complex inputs from being evaluated<sup>21</sup>, and can result in a selection bias of what is asked. The user base of the Arena leans towards developers, which could result in the over-indexing of puzzles, math problems, and questions such as *How many r's are there in strawberry?*<sup>22</sup> For example, in a released dataset from Arena (Zheng et al., 2024) with 33k samples, no questions are referencing *Chaucer* while dozens of questions are about *Star Trek*, highlighting the uneven distribution of topics in this test set<sup>23</sup>. For a global technology provider, real-world commercial applications may differ significantly from this distribution.

**Experimental Setup:** To estimate the potential for overfitting to Chatbot Arena using data

<sup>20</sup><https://huggingface.co/Cohere/Cohere-embed-multilingual-v3.0>

<sup>21</sup><https://github.com/lm-sys/FastChat/blob/main/fastchat/constants.py>

<sup>22</sup><https://techcrunch.com/2024/09/05/the-ai-industry-is-obsessed-with-chatbot-arena-but-it-might-not-be-the-best-benchmark/>

<sup>23</sup><https://www.quantable.com/analytics/elos-and-benchmarking-llms/>

---

from a similar distribution, we fine-tuned a language model with identical training setups that only differ in the composition of arena data used in the mixture. We construct our training dataset by sampling at different ratios from two different data pools: 1) `arena-mix`, which consists of samples from Arena battles, and 2) `other-sft-mix`, a proprietary dataset which includes a variety of supervised fine-tuning datasets focusing on instruction following, multilingual tasks, math, and code. We fine-tuned a 7B base model that is used for the Cohere Command family (Cohere et al., 2025) in these experiments. We refer to the three different training mixes as `0_arena`, `30_arena`, `70_arena` which have 0, 30%, 70% of the training dataset sampled from the `arena-mix` respectively. For all variants, the remainder percentage is sampled from `other-sft-mix`. All three models are fine-tuned for 1.3K steps using a batch size of 128. We note that our goal here is not to produce a state-of-the-art model but rather to estimate a lower bound for the performance gains that could be expected from asymmetries in access to Arena data. Hence, we do not optimize with ablations the correct weighting or data or conduct any hyperparameter sweeps.

**Evaluation method:** Our goal is to measure the lift provided from the training on Arena data on the Arena held-out set. To do so, we measure relative improvements on the 500 English LMArena ArenaHard prompts, an in-distribution test set published by Chatbot Arena that demonstrates exceptionally high correlation (98.6%) with human preference rankings from Chatbot Arena battles (Li et al., 2024c). This dataset consists of challenging user-submitted prompts from Arena that have been carefully curated and evaluated against several criteria, including, but not limited to, domain knowledge, complexity, and problem-solving. This dataset is widely used to gauge expected performance on the Arena. To measure improvements, we simulate human preferences using LM-as-a-judge, which allows us to measure a reasonable estimate of the gains in a controlled setting. Various works have shown that this is correlated with human preferences (Dubois et al., 2023; Rafailov et al., 2023; Kim et al., 2023). We compare against Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and measure win-rates using `gpt-4o-2024-11-20` as our judge model<sup>24</sup>.

**Results:** From Figure 10, we observe that as the amount of `arena-mix` data increases, the models improve in their evaluation on the ArenaHard prompts. Variant `0_arena` scores a win-rate of 23.5%, `30_arena` scores a win-rate of 42.7%, and variant `70_arena`, which is trained on the most `arena-mix` data scores a win-rate of 49.9% against Llama-3.1-8B-Instruct. A win-rate of 50% indicates that the compared models (Llama 3.1 8B Instruct and `70_arena`) are on par on the ArenaHard test set. The relative gains in win-rates are 81.7% for the `30_arena` and 112.3% for the `70_arena` variant. The gains we observe are striking in part because we do not heavily optimize these variants (we trained 3 variants in total and did no hyperparameter optimization or tuning of training steps). To assess whether performance gains generalize beyond the Arena benchmark, we evaluated the fine-tuned models on the out-of-distribution benchmark MMLU (Hendrycks et al., 2021). The results reveal a clear divergence in performance trends (Table 9): while increasing the proportion of Chatbot Arena data within a fixed training budget yielded consistent improvements on the Arena test set, MMLU performance slightly declined from 66.5% (`0_arena`) to 64.4% (`30_arena`) and 65.9% (`70_arena`). This suggests that gains from Chatbot Arena data are highly specific and do not translate to broader generalization, raising important questions about whether leaderboard improvements reflect meaningful progress or simply overfitting to a narrow evaluation distribution.

We also note that often it is not necessary to train explicitly on the data to gain from data access. For example, providers may use the composition of the data to make decisions about weighting different

---

<sup>24</sup><https://platform.openai.com/docs/models/gpt-4o>

---

data sources (Üstün et al., 2024), or may use a small subset of the data to create high-quality synthetic data that is close to the original distribution (Odumakinde et al., 2024; Shimabucoro et al., 2024; Dang et al., 2024a; Aakanksha et al., 2024). Given the stakes behind ranking highly on Chatbot Arena, it is likely that the data at a provider’s disposal from Chatbot Arena is being actively leveraged by different providers to gain an advantage.

## 5 Results: Impact of Model Deprecation on Arena Scores

### Section Findings

- **Model deprecations under a changing task distribution lead to unreliable rankings.** We show that shifts in the task distribution highlight the non-stationary nature of the evaluation environment and have important implications for the stability and fairness of model rankings over time. In evolving task distributions, premature model removal introduces inconsistencies, breaking the BT model’s transitivity assumption and distorting rankings.
- **Model deprecation may result in disconnected subgraphs, which violates BT reliability.** The BT model assumes a sufficiently connected comparison graph to produce globally consistent rankings. When models are deprecated unevenly or when sampling strategies fail to ensure robust overlap in comparisons, the resulting history matrix can become fragmented. Through simulated experiments, we show that this can produce fragmented clusters, and as a result, the global rankings become unreliable.

According to the backend codebase of Chatbot Arena<sup>25</sup>, 47 models are publicly listed as deprecated. In addition, as discussed in Section 4.1, 205 models on Chatbot Arena have been *silently deprecated* by reducing their active sampling rates to near zero (see Figure 17). We observe that model deprecation disproportionately affects different types of models. Specifically, 87.8% of open weights and 89% of open-source models have been deprecated, in contrast to a lower deprecation rate of 80% for proprietary models.

While deprecating older models is necessary to maintain a dynamic leaderboard, especially as new models are introduced regularly, we question how excessive deprecation may compromise the reliability and stability of model rankings. When models are deprecated, future models that enter the Arena will not have direct comparisons with them. However, in principle, the BT model should still handle this reliably because of the transitivity property (Boubdir et al., 2024). Intuitively, transitivity means that if model A is better than model B, and model B is better than model C, then model A should also be better than model C. Transitivity allows the BT model to infer missing outcomes—if two models share common opponents, their relative ranking can be deduced even without a direct comparison (Bradley & Terry, 1952). Formally, in the BT model each competitor  $i$  is associated with a positive parameter  $\pi_i > 0$ , and the probability that model  $i$  beats model  $j$  is given by:

$$P(i > j) = \frac{\pi_i}{\pi_i + \pi_j}.$$

---

<sup>25</sup>[http://github.com/lm-sys/FastChat/blob/Oe6d3e4beaab66f4d3f93db72541a4abab8af28d/fastchat/serve/monitor/monitor\\_md.py#L7](http://github.com/lm-sys/FastChat/blob/Oe6d3e4beaab66f4d3f93db72541a4abab8af28d/fastchat/serve/monitor/monitor_md.py#L7)

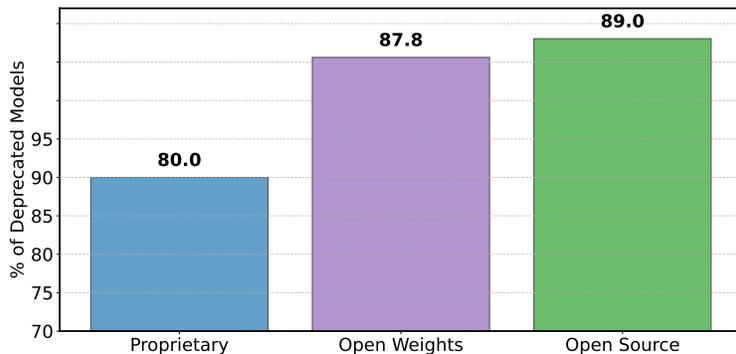


Figure 13: **Share of proprietary and open models that either officially deprecated or inactive on the arena based on leaderboard-stats during the period March 3rd-April 23rd, 2025.** Overall, open-weight and fully open-source models are more likely to become deprecated or inactive compared to proprietary models.

Suppose  $\pi_A > \pi_B$  and  $\pi_B > \pi_C$ . Then

$$P(A > B) = \frac{\pi_A}{\pi_A + \pi_B} > 0.5 \quad \text{and} \quad P(B > C) = \frac{\pi_B}{\pi_B + \pi_C} > 0.5.$$

Moreover, because  $\pi_A > \pi_B > \pi_C$ , we have:

$$\pi_A + \pi_C < \pi_A + \pi_B \quad \implies \quad P(A > C) = \frac{\pi_A}{\pi_A + \pi_C} > \frac{\pi_A}{\pi_A + \pi_B}.$$

Transitivity is a critical property that allows for inferring rankings using fewer data points. However, transitivity requires several underlying properties to hold. In particular, two of these properties are salient for our study of the impact of deprecation on reliability:

- **Assumption 1: Evaluation conditions remain constant.** Transitivity relies on the assumption that evaluation conditions remain constant—that is, paired comparisons must come from a consistent set of tasks or contexts. For example, consider a setting where Tom wins against Susan in chess, and Susan wins against Ronald in both chess and tennis. We can infer Tom’s ranking relative to Ronald in chess, but not in tennis. In practice, Chatbot Arena’s prompt categories and task types evolve over time, and some models become deprecated, meaning they are no longer re-evaluated under the current conditions. We show in Section 5.1 that when a model’s battle history is limited to outdated conditions, its comparisons no longer accurately reflect performance in the new context.
- **Assumption 2: Network of comparisons must be fully interconnected.** Every model must be linked directly or indirectly through pairwise matchups. So in a chess example, we can estimate Tom’s performance against Ronald because they have each played at least one competitor in common (in this case, Susan). If Ronald has had no matches with anyone who has played Tom, we can no longer infer reliably what their rankings are relative to each other. Deviations from transitivity or a fragmented comparison graph can compromise the reliability and consistency of the strength estimates. In Section 5.2, we show that deprecations may result in these fragmented graphs.

---

In Section 5.1 and Section 5.2 that follow, we will interrogate whether Chatbot Arena fulfills each of these assumptions.

## 5.1 Transitivity Under Changing Evaluation Conditions

As we have shown in Section 4.1, the distribution of Chatbot Arena is unique since long-term shifts occur in categories and use cases. This distributional shift contrasts with the static environments typically assumed in Elo and Bradley-Terry systems, such as chess, where the rules and game format remain fixed, ensuring a consistent set of evaluation conditions. If all models were continuously sampled across all points in time, the BT model would likely remain robust because every model would be evaluated on the evolving distribution of tasks. However, as shown in Figure 13, many models are deprecated over time, and their scores stop getting updated.

**Experimental Setup:** To investigate how model deprecations under a changing task distribution can impact model rankings, we simulate BT rankings of models under evolving evaluation conditions. We initialize four models—A, B, C, and D—each with distinct performance profiles across two task types, *Task-1* and *Task-2*. These tasks represent different prompt categories, and each model’s relative strength is defined through task-specific win probabilities. For example, model B has a 90% chance of defeating model D on Task-1 but only a 20% chance on Task-2, with some pairs also allowing for ties. The task-specific win probabilities for different models are provided in Table 8. These probabilities reflect the models’ varying strengths across tasks, mirroring the real-world observation that models excel at different types of prompts.

The simulation is structured into two sequential phases to mimic the evolving task distribution observed on Chatbot Arena. During the first phase, battles are predominantly drawn from *Task-1*. Each of the four models participates in 1000 battles, and the resulting outcomes are used to compute initial rankings. In the second phase, the battle distribution gradually shifts toward *Task-2*. Since model win-rates are task-dependent, battle outcomes change accordingly. We simulate 1000 additional battles in this phase and examine two scenarios to investigate how shifts in prompt distribution and model deprecations jointly influence final rankings. We compute the BT Scores for all models under both scenarios using the implementation provided by Chatbot Arena in their official FastChat codebase<sup>26</sup>. These scores are then used to determine the final model ranks.

- **Scenario I: without deprecation.** We simulate all 2000 battles across both phases, with all four models participating throughout. This represents an ideal scenario where no model is deprecated, and all are evaluated across the evolving task distribution.
- **Scenario II: with deprecation.** We simulate all 2000 battles across both phases. However, at the end of phase 1, model D is deprecated and does not participate in the second phase.

**Deprecation given changing distribution results in unreliable Arena rankings:** As illustrated in Figure 14, our simulation shows that rankings produced by the BT model are highly sensitive to model deprecation, particularly when the prompt distribution changes over time. In the scenario without deprecation, we observe the true rankings given that the BT model remains reliable because it reflects performance across the full history of interactions. However, we observe

---

<sup>26</sup>[https://github.com/lm-sys/FastChat/blob/main/fastchat/serve/monitor/rating\\_systems.py](https://github.com/lm-sys/FastChat/blob/main/fastchat/serve/monitor/rating_systems.py)

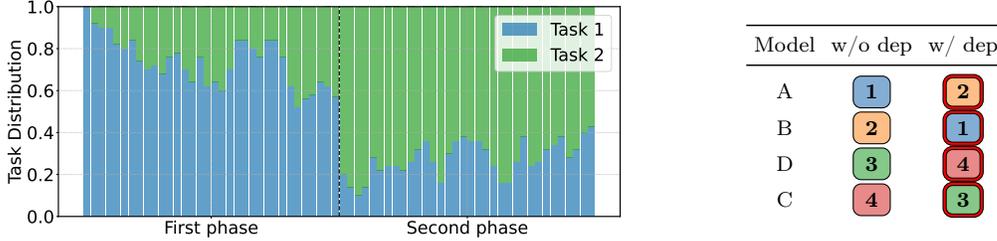


Figure 14: **Impact of evolving task distributions and model deprecation on model rankings.** **Left:** Two-phase task distribution used in the simulation. Phase 1 is *Task-1 heavy*, with most battles based on Task-1; Phase 2 is *Task-2 heavy*, with battles predominantly based on Task-2. **Right:** Model rankings under changing task distributions and deprecation settings. Scenario I only differs from Scenario II in that Model D is deprecated halfway through the battle history (after phase 1). This deprecation causes Scenario II to produce a completely different ranking over models as compared to Scenario I.

the rankings differ if we deprecate Model D between stages. While its matchups from Phase 1 still influence BT scores, the absence of updated comparisons causes a skew in the rankings of remaining models. Models A and D are ranked lower, and Models B and C are ranked higher than their true performance merits. Comparing Scenarios I and II in Figure 14 reveals a clear divergence, with the scenario involving deprecation yielding unreliable rankings that no longer reflect true relative performance under the evolving task distribution.

This violates core assumptions of the BT model, namely, that rankings reflect transitive and consistently sampled matchups, ultimately compromising the validity of inferred rankings. When models are no longer sampled under current task distributions, historical pairwise comparisons cease to represent present-day performance. This issue is particularly problematic in real-world settings where user prompt distributions shift over time. For instance, a model tuned for multilingual prompts may improve ranking as non-English tasks become more common. However, if deprecated, its BT ranking will likely understate its true performance. Similarly, code-specialized models may benefit from the increasing volume of coding tasks, but only if they continue to participate in evaluation.

## 5.2 Sparse Battle History Risks

In this section, we show that the deprecation policy can lead to a sparse matrix and disconnected comparison graphs, which in turn distort the resulting rankings. As demonstrated by Ford Jr (1957), the maximum likelihood estimate does not exist if models can be partitioned into two non-empty subsets without comparisons between them or if all comparisons between the two groups are one-sided (i.e., one group always wins). Therefore, to ensure a unique and finite estimation, the comparison graph must be connected. For any possible partition of models, there must be at least one win going in each direction across the partition. This ensures that no subset of models is entirely isolated in the win/loss structure. The Chatbot Arena comparison matrix can potentially become disconnected because of the extremely high levels of model removals over time (as discussed in Figure 13).

**Experiment Setup:** To investigate the impact of sparse comparison graphs on the rankings obtained via the Bradley-Terry model used by Chatbot Arena we simulate the following scenarios:

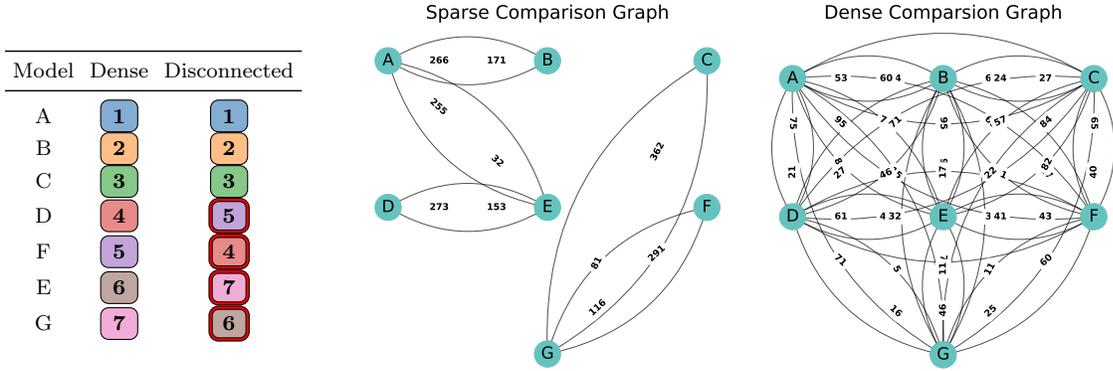


Figure 15: **Impact of comparison graph sparsity on model rankings.** **Left:** Rankings for models D, E, F, and G diverge from the gold rankings when the comparison graph is sparse, whereas model rankings fully align with the gold rankings when the comparison graph is dense. **Right:** Visualization of the comparison graphs in sparse and dense settings. An edge between two models indicates a head-to-head matchup, annotated with the number of wins for each model. For example, in the sparse graph, Model A and Model B played 437 matches, with Model A winning 266 and Model B winning 171.

- **Scenario I: Dense comparison graph.** All models are allowed to compete against one another—albeit with varying numbers of head-to-head battles—resulting in a well-connected comparison graph in which every node (model) is linked to others via edges representing battle outcomes.
- **Scenario II: Disconnected comparison graph.** We create a disconnected battle history by imposing constraints on which pairs of models are allowed to engage in battles. This allows us to create a sparse battle history where each model ends up playing against a subset of models.

The full comparison graph based on battle histories for both scenarios is shown in Figure 15. In both scenarios, a total of 2000 battles are played under the corresponding setting. For a paired match between models A and B, each with respective true skill ratings  $r_A$  and  $r_B$ , the expected scores  $E_A$  and  $E_B$  can be computed as:

$$E_A = \frac{1}{1 + e^{\alpha(r_B - r_A)}}, \quad E_B = \frac{1}{1 + e^{\alpha(r_A - r_B)}} \quad (2)$$

The expected scores  $E_A$  and  $E_B$  are used to predict the winner of the battle. For simplicity, we exclude the possibility of ties in this experiment. We assign the following true skill ratings to the models: 1450 (Model A), 1390 (Model B), 1250 (Model C), 1200 (Model D), 1101 (Model E), 1150 (Model F), and 1000 (Model G). These ratings are used to calculate the expected scores and match outcomes. Finally, we compute BT Scores for all models under both scenarios using the official implementation, followed by Chatbot Arena<sup>26</sup>. This is then used to determine the ranks for each model corresponding to both scenarios.

**Sparse or disconnected graphs lead to unreliable rankings:** Figure 15 illustrates the model

---

rankings with sparse and dense battle history graphs. We observe that the rankings derived from the dense comparison graph align closely with the models’ rankings according to their true skill ratings. In contrast, a sparse or disconnected comparison graph results in an inaccurate estimation of the models’ skills. These results highlight that reliable rankings under the BT model require a connected comparison graph. While some level of model removal is inevitable (for example, models are no longer hosted on an API), preserving connectivity means ensuring that comparisons remain sufficiently distributed across active models and that transitions in and out of the pool do not isolate subsets of models from the broader comparison structure.

## 6 Recommendations and Guidelines for Improving Leaderboards

*When my information changes, I alter my conclusions. What do you do, sir?* — **John Maynard Keynes**

We include below recommendations that we believe are urgent and critical. However, we also believe these recommendations are very achievable and will restore scientific credibility and trust within the wider research ecosystem:

**Prohibit score retraction after submission.** Currently, providers who have engaged in private testing are allowed to retract submissions and only submit the best variant to the public leaderboard. As we have shown in Section 4.1, this can lead to overfitting and obscures meaningful progress as it makes it difficult to distinguish between models that have legitimately improved versus those that have exploited statistical shortcuts (Ying, 2019). We urge Chatbot Arena to prohibit retraction after submission, ensuring all tested variants’ scores are permanently visible on the leaderboard. Providers should also disclose the total number of private variants tested prior to public launch, including historical submissions, to contextualize their results.

**Establish transparent limits on the number of private variants per provider.** As illustrated in Figure 6, private testing volume varies widely across providers, creating unfair advantages. To curb overfitting and level the playing field, Chatbot Arena should enforce a strict cap of private variants per provider for any given model launch. This should be enforced at a provider level, and not per model type and size as that is impossible to audit with API hosting. This strict limit should be disclosed to all providers (proprietary, open-weights, open-source) and to the wider Chatbot Arena community. This restriction would discourage excessive undisclosed testing while still allowing limited iteration.

**Establish clear and auditable model deprecation criteria.** The current criteria, which states *“Models may be retired after 3000 votes if there are two more recent models in the same series and/or if there are more than 3 providers that offer models cheaper or same price and strictly better (according to overall Arena score) than this model”* is ambiguous and make it impossible to audit the logic in practice. Key terms like “same series” and “more recent” lack formal definitions, making it unclear how to determine whether a model is eligible for retirement based on lineage. Additionally, the requirement that “more than 3 providers” must offer strictly better and cheaper models introduces confusion around what threshold of improvement in Arena Score is considered meaningful. The use of “and/or” further complicates interpretation, as it’s unclear whether meeting one condition is sufficient or if both must be satisfied. Lastly, using price as a filtering criterion is problematic since it’s subject to change, varies across hosting platforms, and is not inherently

---

tied to a model’s performance or utility. We note that many of these models are hosted for free on the Arena, and so there should be clarity about what source of pricing is being used. This lack of precision makes it challenging to apply the rule consistently or verify retirement decisions.

We recommend a stratified approach that retires models proportionally across proprietary, open-weight, and open-source categories to preserve balance and fairness in the evaluation based on two criteria: availability and performance. We propose retiring the bottom 30th percentile within each category of open-source, open-weight, and proprietary after rankings converge. This stratified pruning prevents provider-type bias, keeps strong models from underrepresented groups visible, and maintains comparison graph connectivity. It also reduces ranking inconsistencies seen with uneven retirement, as discussed in Section 5.

**Improve sampling fairness.** As shown in Figure 5, the sampling rates vary greatly by providers, and also disproportionately undersample open-weight and open-source models, creating large asymmetries in data access over time and resulting in unstable Arena scores (Section 5). This is particularly important given that this is a community-driven voting benchmark, where at present free human feedback is primarily benefiting proprietary models. This avoids disparities over time and a status quo where proprietary models are benefiting more from valuable and freely given human data (Sambasivan et al., 2021). In their own work (Chiang et al., 2024) (Section 5, Equation 9), the Chatbot Arena authors introduce an active sampling rule designed to enhance the efficiency and statistical robustness of the leaderboard’s evaluation process. This rule selects model pairs based on the expected reduction in the confidence interval of the win-rate estimate, and is formally defined as:

$$P_t(a) \propto \sqrt{\frac{\Sigma_{t,a,a}}{|\{t : A_t = a\}|}} - \sqrt{\frac{\Sigma_{t,a,a}}{|\{t : A_t = a\}| + 1}}$$

where  $P_t(a)$  is the sampling probability of model pair  $a$  at time  $t$ , and  $\Sigma_{t,a,a}$  is the estimated variance for the win-rate of pair  $a$  at time  $t$ .

This formulation effectively prioritizes under-evaluated and high-variance pairs, aligning sampling with the goal of rapidly reducing uncertainty in rankings. While this sampling rule is clearly articulated in the paper, we have not seen evidence of its deployment in the current leaderboard. We recommend adopting this sampling strategy in practice and providing periodic reporting on its usage. Doing so would align the platform’s operations with its methodological innovations, support more balanced and transparent evaluations, and improve confidence in leaderboard dynamics over time.

**Provide public transparency into all tested models, deprecations, and sampling rates.**

Most of these findings were only possible through access to private model testing, or crawling Chatbot Arena over a period of time. Providing transparency into the full suite of models that were tested, deprecated, and how often they were sampled against which other models, would enable the oversight and trust in the benchmark that Chatbot Arena affirms in their policies. This transparency could be provided on a rolling basis (e.g., every quarter, for the prior quarter of model battles). It would enable the community to help in the process of continuing to improve this community benchmark. For instance, the backend codebase of Chatbot Arena–FastChat – publicly lists deprecated models on GitHub<sup>25</sup>, where 47 models are explicitly marked as deprecated. However, four times that number have been silently deprecated without warning. We recommend Chatbot Arena expand the definition of “deprecated” to include models that are no longer being

---

regularly sampled from, and list these deprecated models on their website to make it transparent for everyone which models are no longer active on Chatbot Arena.

## 7 Limitations

**We do not have insight into Chatbot Arena’s raw data:** A subset of the data sources utilized for this study have undergone pre-processing by Chatbot Arena, which as stated by Chatbot Arena often involves de-duplication, removal of battles corresponding to suspicious voting patterns, etc (Chiang et al., 2024). Through this work, we also establish that private battles are removed from the datasets released by Chatbot Arena. Although we also have proprietary API data for the models we test on the Arena, it only reveals a subset of battles. Without access to original and comprehensive raw data, it is hard to investigate patterns related to adversarial voting where users intentionally submit votes to manipulate rankings or undermine the system. Various previous works have shown that adversarial voting is a critical concern for the reliability of any crowd-sourced evaluation platform like Chatbot Arena (Huang et al., 2025) (Min et al., 2025). We do not explore this in this work, but see more investigation here as an important topic for future work.

**Our scraped data snapshot only covers a limited period:** Our scraped-random-sample was the only way to identify private variants being tested by various providers. However, it covers a limited time period from January–March, 2025. This time frame coincided with Meta’s launch of Llama 4, and so we find them to be the provider with the highest number of private variants in our analysis. We believe we might be underestimating the counts for providers having fewer model launches during this period.

**Our training experiments likely underestimate the potential to overfit:** Our estimate of overfitting is likely conservative, as it is based on training with only a fraction of the data believed to be available to some proprietary model providers. This disparity suggests that proprietary models may be trained on 5 to 10 times more data than we use, potentially increasing the risk of overfitting to patterns not present in our smaller subset. These observations underscore the importance of further examining data scale and its implications for model overfitting to certain leaderboards.

**We rely on the model’s self-identification to attribute private models to their respective providers:** Since the identity of anonymous models is not publicly disclosed, we use model self-identification as a proxy—prompting each model directly and observing how it responds. When a model consistently names a particular provider across multiple prompts, we attribute it to that provider. While this method provides a reasonable signal, it is inherently approximate. Due to limited data and the potential for models to respond inconsistently or ambiguously, some misattributions may occur. To encourage validation of our estimates, we include the codename and our estimate of identity in Appendix E.4. We welcome correspondence with the authors from providers if any of the estimates of ownership are incorrect.

## 8 Related Work

*The strength of science lies in its ability to withstand challenge, verification, and replication.* — **Richard P. Feynman**

---

## 8.1 Meta-studies on the Rigor of Benchmarking in AI

Our work contributes to a wider body of work examining the role of benchmarks in determining progress in machine learning. Benchmarking has played a central role in shaping research priorities and incentives within the deep learning community (Koch & Peterson, 2024). Research has found that benchmarks are rarely impartial and instead shaped by the environments in which the benchmarks are made, finding that assumptions, commitments, and dependencies can often have large implications in final outcomes (Aniba et al., 2010; Bartz-Beielstein et al., 2020). Creating a meaningful and reliable benchmark is challenging, and there has been critical work identifying key benchmark desiderata and open challenges.

**Propensity for overfitting.** Static task-based leaderboards, such as Hugging Face’s Open LLM Leaderboard (Fourrier et al., 2024; Gao et al., 2024) and OpenCompass (Contributors, 2023), aim to evaluate models across a broad range of skills but are often susceptible to data contamination and implicit overfitting (Deng et al., 2024; Golchin & Surdeanu, 2024; Roberts et al., 2023; Dong et al., 2024; Singh et al., 2024a; Longpre et al., 2024a). Prior works (Deng et al., 2024; Golchin & Surdeanu, 2024; Yang et al., 2023) have proposed various methods for detecting contamination, while Dong et al. (2024) discusses how such contamination impedes the ability to distinguish true generalization, ultimately hindering progress. Although dynamic, live benchmarks like Chatbot Arena significantly reduce the risk of overfitting, we report in this paper that certain practices—such as multiple submissions during the anonymous testing period and best-of-N submissions—tend to favor large, proprietary players with disproportionate access to data. As a result, model development may be deliberately optimized for performance on Chatbot Arena.

**Lack of standardization across benchmarks.** The lack of standardization in benchmarks complicates meaningful comparisons due to inconsistent metrics and task definitions. Ethayarajh & Jurafsky (2020) critique NLP leaderboards for prioritizing accuracy over dimensions like model compactness and fairness. Similarly, Ruder (2021) highlights that benchmarks such as SuperGLUE (Sarlin et al., 2020) are quickly saturated, with models reaching superhuman performance while still failing in real-world scenarios, underscoring the need for dynamic and standardized evaluation. This inconsistency risks misleading practitioners, as echoed in recent critiques (Bartz-Beielstein et al., 2020; Reuel et al., 2024).

**Quality of data and limited reproducibility.** A recent study by Vendrow et al. (2025) revealed widespread label errors that compromise evaluation reliability, showing that even frontier LLMs can struggle with seemingly simple tasks. Similarly, Digan et al. (2021) identified reproducibility challenges arising from complex data streams, which affect result consistency. Related work (Bartz-Beielstein et al., 2020; Longpre et al., 2023; Reuel et al., 2024; Albalak et al., 2024) further emphasizes that poor data quality and limited reproducibility can lead to unreliable evaluations and undermine scientific credibility.

**Favored benchmarks may not capture performance in the real world.** Commonly used benchmarks often fail to capture real-world performance, creating a gap between test scores and practical utility due to their tendency to overlook the dynamic and complex nature of real-world tasks. Recent studies (Ott et al., 2022; Parli, 2022) highlight this disconnect, observing that models frequently excel on benchmarks while underperforming in practical applications, especially as benchmarks quickly reach saturation.

---

## 8.2 Human Voting-based Benchmarks

**Wider studies on the role and benefits of human voting-based benchmarks.** Chatbot Arena is an example of a human voting-based benchmark. Human judgment has long been regarded as the gold standard for evaluating the quality of model-generated outputs. These models should ultimately align with human values, and certain nuanced qualities, such as coherence, harmlessness, and readability, are best assessed by humans (Van Der Lee et al., 2019; Boubdir et al., 2023). Platforms like Chatbot Arena (Chiang et al., 2024), Talk Arena (Li et al., 2024a), and Game Arena (Hu et al., 2024), Aya UI Interface (Singh et al., 2024c) effectively use crowdsourcing to gather large volumes of real-world user prompts and feedback. Many opt for Elo-like or BT-style rankings to rank models. Moreover, collecting human preference data has also proven invaluable for alignment techniques like Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Ahmadian et al., 2024; Dang et al., 2024a), which helps fine-tune models to generate more natural and human-preferred responses. Human voting has been shown to mitigate some of the biases associated with using LLM-as-a-judge approaches, which, while improving evaluation efficiency, may raise concerns about robustness (Raina et al., 2024) and introduce various forms of bias (Koo et al., 2023; Shimabucoro et al., 2023; Chen et al., 2024; Zheng et al., 2024). Furthermore, live leaderboards offer several advantages over static task benchmarks, including a lower risk of data contamination and greater adaptability to evolving evaluation needs.

**Critiques of Human-Voting Based Benchmarks.** Voting-based live benchmarks like Chatbot Arena also face evaluation challenges not addressed in this paper. Chatbot Arena (Chiang et al., 2024) has made substantial efforts to ensure reliability and security, including malicious user detection, bot protection via Google reCAPTCHA v3, vote limits per IP address, prompt de-duplication, and other safeguards<sup>6</sup>. Nonetheless, recent work has focused on auditing the reliability of human-voting-based live leaderboards. For instance, studies have demonstrated that such leaderboards are vulnerable to low-cost manipulation, with adversarial users able to de-anonymize model responses and carry out targeted voting attacks (Huang et al., 2025). Additionally, Zhao et al. (2024); Min et al. (2025) suggest that Chatbot Arena rankings can be artificially inflated through various adversarial voting strategies. These vulnerabilities raise concerns about the overall trustworthiness of Chatbot Arena. While our study does not explicitly investigate adversarial voting, we note that Chatbot Arena’s policy of informing model providers when testing begins and disclosing model aliases may create conditions conducive to leaderboard manipulation.

## 9 Conclusion

It is far easier to point out issues with the Arena than the huge amount of work that went into building it. While our work is motivated by the need to maintain scientific integrity in AI progress, we believe it is important to note the huge amount of work involved for a small group of organizers to build a hugely popular community benchmark. Their efforts have democratized access to many models and enabled a large and varied user base to weigh in on what matters in the real world for model selection. Hence, while we point out systematic issues with Chatbot Arena as it is now, we also acknowledge that many of these issues may have gradually emerged as the leaderboard took on outsized importance in visibility to providers.

This work demonstrates the difficulty in maintaining fair evaluations, despite best intentions. We show that coordination among a handful of providers and preferential policies from Chatbot Arena

---

towards the same small group have jeopardized scientific integrity and reliable Arena rankings. The widespread and apparent willful participation in the gamification of arena scores from a handful of top-tier industry labs is undoubtedly a new low for the AI research field. **As scientists, we must do better. As a community, we must demand better.**

We believe it is very feasible for the organizers of Chatbot Arena to continue to innovate and restore trust by revising their policies. We propose a series of very straightforward recommendations to help reinforce the reliability and fairness of the leaderboard. Most urgently, providers should not be allowed to choose which scores are made public. There should be strict and transparent limits to the number of private variants per provider. Providers, whether from academic or industry labs, should be aware of private testing and the limits should be the same across providers. There should be transparent criteria for model removal from the arena and fairer sampling that is motivated by reducing uncertainty in rankings instead of being skewed towards giving proprietary models more battles. We believe the implementation of these recommendations is critical for addressing ranking distortions, but also necessary in the long term to ensure the benefits from participating on the leaderboard are not concentrated in a handful of providers.

## 10 Acknowledgements

We thank our colleagues who have supported various aspects of this project: Madeline Smith, Brittwnya Prince, Thomas Euyang, and Shubham Shukla.

## References

- Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12027–12049, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.671. URL <https://aclanthology.org/2024.emnlp-main.671/>.
- David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. Irokobench: A new benchmark for african languages in the age of large language models, 2025. URL <https://arxiv.org/abs/2406.03368>.
- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL <https://aclanthology.org/2024.acl-long.662/>.
- Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang,

- 
- Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models, 2024. URL <https://arxiv.org/abs/2402.16827>.
- Mohamed Radhouene Aniba, Olivier Poch, and Julie D Thompson. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic acids research*, 38(21):7353–7363, 2010.
- Barry C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics*. Wiley, Hoboken, NJ, 1992.
- Thomas Bartz-Beielstein, Carola Doerr, Daan van den Berg, Jakob Bossek, Sowmya Chandrasekaran, Tome Eftimov, Andreas Fischbach, Pascal Kerschke, William La Cava, Manuel Lopez-Ibanez, Katherine M. Malan, Jason H. Moore, Boris Naujoks, Patryk Orzechowski, Vanessa Volz, Markus Wagner, and Thomas Weise. Benchmarking in optimization: Best practice and open issues, 2020. URL <https://arxiv.org/abs/2007.03488>.
- Meriem Boubdir, Edward Kim, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. Which prompts make the difference? data prioritization for efficient human llm evaluation, 2023. URL <https://arxiv.org/abs/2310.14424>.
- Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. Elo uncovered: Robustness and best practices in language model evaluation. *Advances in Neural Information Processing Systems*, 37:106135–106161, 2024.
- Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL <https://aclanthology.org/2021.naacl-main.385/>.
- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL <http://www.jstor.org/stable/2334029>.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.474. URL <https://aclanthology.org/2024.emnlp-main.474/>.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*, 2024.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

---

Kenneth Ward Church. Emerging trends: A tribute to charles wayne. *Natural Language Engineering*, 24(1):155–160, 2018. doi: 10.1017/S1351324917000389.

Cohere. Command r and command r+ model card, 2024. URL <https://docs.cohere.com/docs/responsible-use>.

Cohere. A deepdive into aya vision: Advancing the frontier of multilingual multimodality, 2025. URL <https://huggingface.co/blog/aya-vision>.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammari, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avolos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D’souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruvi Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukas Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynihan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Chang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang,

- 
- Zhenyu Zhao, and Zhoujie Zhao. Command a: An enterprise-ready large language model, 2025. URL <https://arxiv.org/abs/2504.00698>.
- OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023.
- John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13134–13156, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.729. URL <https://aclanthology.org/2024.emnlp-main.729/>.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expand: Combining research breakthroughs for a new multilingual frontier, 2024b. URL <https://arxiv.org/abs/2412.04261>.
- Herbert A. David and H. N. Nagaraja. *Order Statistics*. Wiley, Hoboken, NJ, 3 edition, 2003.
- Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482. URL <https://aclanthology.org/2024.naacl-long.482/>.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- William Digan, Aurélie Névéal, Antoine Neuraz, Maxime Wack, David Baudoin, Anita Burgun, and Bastien Rance. Can reproducibility be improved in clinical natural language processing? a study of 7 clinical nlp suites. *Journal of the American Medical Informatics Association*, 28(3):504–515, 2021.
- Shachar Don-Yehiya, Ben Burtenshaw, Ramon Fernandez Astudillo, Cailean Osborne, Mimansa Jaiswal, Tzu-Sheng Kuo, Wenting Zhao, Idan Shenfeld, Andi Peng, Mikhail Yurochkin, Atoosa Kasirzadeh, Yangsibo Huang, Tatsunori Hashimoto, Yacine Jernite, Daniel Vila-Suero, Omri Abend, Jennifer Ding, Sara Hooker, Hannah Rose Kirk, and Leshem Choshen. The future of open human feedback, 2024. URL <https://arxiv.org/abs/2408.16961>.
- Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association*

- 
- for *Computational Linguistics: ACL 2024*, pp. 12039–12050, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.716. URL <https://aclanthology.org/2024.findings-acl.716/>.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.
- Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., New York, 1978. ISBN 0668047216 9780668047210. URL <http://www.amazon.com/Rating-Chess-Players-Past-Present/dp/0668047216>.
- Nathan Ensmenger. Is chess the drosophila of artificial intelligence? a social history of an algorithm. *Social Studies of Science*, 42(1):5–30, Oct 2011. doi: <https://doi.org/10.1177/0306312711424596>. URL <https://pubmed.ncbi.nlm.nih.gov/22530382/>.
- Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.393. URL <https://aclanthology.org/2020.emnlp-main.393/>.
- Lester R Ford Jr. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8P2):28–33, 1957.
- Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. [https://huggingface.co/spaces/open-llm-leaderboard/open\\_llm\\_leaderboard](https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard), 2024.
- Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios N. Angelopoulos, and Ion Stoica. Prompt-to-leaderboard, 2025. URL <https://arxiv.org/abs/2502.14855>.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL <https://zenodo.org/records/12608602>.
- Mark E Glickman. Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, pp. 377–394, 1999.
- Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=2Rwq6c3tvr>.
- Google. Introducing gemma 3: The most capable model you can run on a single gpu or tpu, 2025. URL <https://blog.google/technology/developers/gemma-3/>. Accessed: 2025-04-09.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

- 
- Donna Harman. Overview of the first trec conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 36–47, 1993.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Sara Hooker. On the limitations of compute thresholds as a governance strategy, 2024. URL <https://arxiv.org/abs/2407.05694>.
- Lanxiang Hu, Qiyu Li, Anze Xie, Nan Jiang, Ion Stoica, Haojian Jin, and Hao Zhang. Gamearena: Evaluating llm reasoning through live computer games. *ArXiv*, abs/2412.06394, 2024. URL <https://api.semanticscholar.org/CorpusID:274597249>.
- Yangsibo Huang, Milad Nasr, Anastasios Angelopoulos, Nicholas Carlini, Wei-Lin Chiang, Christopher A Choquette-Choo, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Ken Ziyu Liu, et al. Exploring and mitigating adversarial manipulation of voting-based leaderboards. *arXiv preprint arXiv:2501.07493*, 2025.
- David R Hunter. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023.
- Bernard J Koch and David Peterson. From protoscience to epistemic monoculture: How benchmarking set the stage for the deep learning revolution. *arXiv preprint arXiv:2404.06647*, 2024.
- Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pp. 102–121, New York City, June 2006. Association for Computational Linguistics.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*, 2023.
- Christoph Leitner, Achim Zeileis, and Kurt Hornik. Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the euro 2008. *International Journal of Forecasting*, 26(3): 471–481, 2010.
- Minzhi Li, Will Held, Michael J. Ryan, Kunat Pipatanakul, Potsawee Manakul, Hao Zhu, and Diyi Yang. Talk arena: Interactive evaluation of large audio models, 2024a.

- 
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024b. URL <https://arxiv.org/abs/2406.11939>.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024c.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2022.
- Liquipedia. Elo rating, 2019. URL [https://liquipedia.net/starcraft/Elo\\_rating](https://liquipedia.net/starcraft/Elo_rating). Accessed: 2025-04-09.
- Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023.
- Shayne Longpre, Stella Biderman, Alon Albalak, Hailey Schoelkopf, Daniel McDuff, Sayash Kapoor, Kevin Klyman, Kyle Lo, Gabriel Ilharco, Nay San, Maribeth Rauh, Aviya Skowron, Bertie Vidgen, Laura Weidinger, Arvind Narayanan, Victor Sanh, David Ifeoluwa Adelani, Percy Liang, Rishi Bommasani, Peter Henderson, Sasha Luccioni, Yacine Jernite, and Luca Soldaini. The responsible foundation model development cheatsheet: A review of tools & resources. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL <https://openreview.net/forum?id=tH1dQH20eZ>. Survey Certification.
- Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, et al. Consent in crisis: The rapid decline of the ai data commons. *Advances in Neural Information Processing Systems*, 37:108042–108087, 2024b.
- Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Naana Obeng-Marnu, Manan Dey, Mohammed Hamdy, et al. Bridging the data provenance gap across text, speech and video. *arXiv preprint arXiv:2412.17847*, 2024c.
- Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3245–3276, 2024d.
- Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. Time waits for no one! analysis and challenges of temporal misalignment. *arXiv preprint arXiv:2111.07408*, 2021.

- 
- Nestor Maslej, Loredana Fattorini, Elif Kiesow Cortez, Julia Betts Lotufo, Anka Reuel, Alexandra Rome, Angelo Salatino, Lapo Santarlaschi, Emily Capstick, Malou van Draanen Glismann, Njenga Kariuki, Armin Hamrah, Sukrut Oak, Ngorli Fiifi Paintsil, and Andrew Shi. Artificial intelligence index report 2025. <https://hai.stanford.edu/ai-index/2025-ai-index-report>, 2025. AI Index Report, Stanford Institute for Human-Centered Artificial Intelligence.
- Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025. URL <https://ai.meta.com/blog/llama-4-multimodal-intelligence>.
- Rui Min, Tianyu Pang, Chao Du, Qian Liu, Minhao Cheng, and Min Lin. Improving your model ranking on chatbot arena by vote rigging. *arXiv preprint arXiv:2501.17858*, 2025.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt llm evaluation, 2024. URL <https://arxiv.org/abs/2401.00595>.
- Ayomide Odumakinde, Daniel D’souza, Pat Verga, Beyza Ermis, and Sara Hooker. Multilingual arbitrage: Optimizing data pools to accelerate multilingual progress, 2024. URL <https://arxiv.org/abs/2408.14960>.
- Will Orr and Edward B Kang. Ai as a sport: On the competitive epistemologies of benchmarking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1875–1884, 2024.
- Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793, 2022.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Vanessa Parli. Ai benchmarks hit saturation, 2022. URL <https://hai.stanford.edu/news/ai-benchmarks-hit-saturation>.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
- Vyas Raina, Adian Liusie, and Mark Gales. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*, 2024.
- Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. *CoRR*, abs/2111.15366, 2021. URL <https://arxiv.org/abs/2111.15366>.
- Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *arXiv preprint arXiv:2411.12990*, 2024.

- 
- Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. To the cutoff... and beyond? a longitudinal perspective on llm data contamination. In *The Twelfth International Conference on Learning Representations*, 2023.
- Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL <https://aclanthology.org/2021.acl-long.346/>.
- Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*, 2024.
- Sebastian Ruder. Challenges and opportunities in nlp benchmarking, August 2021. URL <https://www.ruder.io/nlp-benchmarking/>. Accessed: 2025-04-08.
- Israfil Salazar, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh, Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, Dominik Krzemiński, Jekaterina Novikova, Luísa Shimabucoro, Joseph Marvin Imperial, Rishabh Maheshwary, Sharad Duwal, Alfonso Amayuelas, Swati Rajwal, Jebish Purbey, Ahmed Ruby, Nicholas Popovič, Marek Suppa, Azmine Toushik Wasi, Ram Mohan Rao Kadiyala, Olga Tsymboi, Maksim Kostritsya, Bardia Soltani Moakhar, Gabriel da Costa Merlin, Otávio Ferracioli Colletti, Maral Jabbari Shiviari, MohammadAmin farahani fard, Silvia Fernandez, María Grandury, Dmitry Abulkhanov, Drishti Sharma, Andre Guarnier De Mitri, Leticia Bossatto Marchezi, Johan Obando-Ceron, Nazar Kohut, Beyza Ermis, Desmond Elliott, Enzo Ferrante, Sara Hooker, and Marzieh Fadaee. Kaleidoscope: In-language exams for massively multilingual vision evaluation, 2025. URL <https://arxiv.org/abs/2504.07072>.
- Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL <https://doi.org/10.1145/3411764.3445518>.
- Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–9. IEEE, 2023.
- Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947, 2020.
- Luísa Shimabucoro, Timothy Hospedales, and Henry Gouk. Evaluating the evaluators: Are current few-shot learning benchmarks fit for purpose? *arXiv preprint arXiv:2307.02732*, 2023.

- 
- Luísa Shimabucoro, Sebastian Ruder, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. LLM see, LLM do: Leveraging active inheritance to target non-differentiable objectives. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9243–9267, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.521. URL <https://aclanthology.org/2024.emnlp-main.521/>.
- Aaditya K. Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, Gergely Szilvasy, and Dieuwke Hupkes. Evaluation data contamination in llms: how do we measure it and (when) does it matter?, 2024a. URL <https://arxiv.org/abs/2411.03923>.
- Shivalika Singh, Angelika Romanou, Clémentine Fourier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*, 2024b.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.620. URL <https://aclanthology.org/2024.acl-long.620/>.
- Kelly Tang, Wei-Lin Chiang, and Anastasios N. Angelopoulos. Arena explorer: A topic modeling pipeline for llm evals & analytics, 2025.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petriani, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder

- 
- Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.
- Rachel Thomas and David Uminsky. The problem with metrics is a fundamental problem for ai, 2020. URL <https://arxiv.org/abs/2002.08512>.
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL <https://aclanthology.org/2024.acl-long.845/>.
- Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 355–368, 2019.
- Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability? *arXiv preprint arXiv:2502.03461*, 2025.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- Ben P. Wise. Elo ratings for large tournaments of software agents in asymmetric games. *ArXiv*, abs/2105.00839, 2021. URL <https://api.semanticscholar.org/CorpusID:233481682>.

---

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL <https://arxiv.org/abs/2407.10671>.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168:022022, Feb 2019. doi: <https://doi.org/10.1088/1742-6596/1168/2/022022>. URL <https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022>.

Wenting Zhao, Alexander M. Rush, and Tanya Goyal. Challenges in trustworthy human evaluation of chatbots. *ArXiv*, abs/2412.04363, 2024. URL <https://api.semanticscholar.org/CorpusID:274515029>.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

## A Chatbot Arena Background

LMSYS originated from a multi-university collaboration involving UC Berkeley, Stanford, UCSD, CMU, and MBZUAI in 2023. It was established as a non-profit corporation in September 2024 to incubate early-stage open-source and research projects. Chatbot Arena was first launched in May 2023 under LMSYS and later evolved into a standalone project with its own dedicated website<sup>27</sup> maintained under the name LMArena by researchers from UC Berkeley SkyLab. It has emerged as a critical platform for live, community-driven LLM evaluation, attracting millions of participants and collecting over 3 million votes to date.

LMArena operates based on human preferences. Chatbot Arena asks users to input prompts in battles. The user then votes for their preferred model based on the outputs generated by the models in the battle in response to the user’s prompts. These preferences are then used by Chatbot Arena to compute model ratings using algorithms like Online Elo and Bradley-Terry.

## B Bradley-Terry Rating Model

Consider a set of  $m$  players (models) and  $n$  pairwise comparisons between them. Let  $X \in \mathbb{R}^{m \times n}$  be the design matrix, where each column represents one pairwise comparison. In the Bradley-Terry

---

<sup>27</sup><https://lmsys.org/blog/2024-09-20-arena-new-site/>

---

model, the probability that player  $i$  is preferred over player  $j$  in a comparison is modeled as:

$$P(i \text{ preferred over } j) = \frac{1}{1 + e^{(\beta_j - \beta_i)}}$$

Then, in the matrix  $X$ , column vector  $k$  has a value of 1 at position  $i$ , -1 at position  $j$ , and 0 elsewhere. Let  $Y \in \{0, 1\}^n$  be the vector of observed outcomes, where  $Y_k = 1$  if player  $i$  wins the  $k$ -th comparison and  $Y_k = 0$  if player  $j$  wins. Our goal is to estimate the Bradley-Terry coefficients  $\beta \in \mathbb{R}^m$ , which determine the relative strengths of the players. The coefficients  $\beta$  are estimated via maximum likelihood estimation by minimizing the expected cross-entropy loss,

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^m} \frac{1}{n} \sum_{k=1}^n \ell(\sigma(X^T \beta)_k, Y_k)$$

where  $\sigma(\cdot)$  is the logistic function that models the relative player strengths, and  $\ell(\cdot)$  represents the binary cross-entropy loss between the predicted probabilities and the observed outcomes  $Y$ . The estimated coefficient  $\beta$  captures the latent strength of each player.

Once the Bradley-Terry model estimates the coefficients, we can scale them to obtain Elo-like ratings using the transformation:

$$R_m = \text{scale} * \beta + \text{initial rating}$$

In practice, Chatbot Arena does not rely solely on a model’s Arena Score for ranking. Instead, it also considers the confidence intervals associated with these scores. When the confidence intervals of two models overlap, it becomes difficult to determine which one is truly better. This uncertainty is reflected in the final ranking table, adding nuance and statistical rigor to the leaderboard ([Chiang et al., 2024](#)).

$$\text{rank}(m) = 1 + \sum_{m' \in [M]} 1 \{m' > m\}$$

## C Unbiased Sampling: Why Selecting the Maximum Introduces Bias?

Let  $(\hat{\beta}_k)_{k=1}^N$  be i.i.d. real-valued random variables with common cumulative distribution function  $F$  and finite expectation  $\mu := \mathbb{E}[\hat{\beta}_k]$ . Assume the distribution is *non-degenerate*, i.e.,  $\text{Var}(\hat{\beta}_k) > 0$ . The maximum is defined as:

$$\hat{\beta}_{\text{Best}} := \max\{\hat{\beta}_1, \dots, \hat{\beta}_N\}, \quad N \geq 2.$$

**Theorem 1** For every  $N \geq 2$ ,

$$\mathbb{E}[\hat{\beta}_{\text{Best}}] > \mathbb{E}[\hat{\beta}_k] \iff \text{Var}(\hat{\beta}_k) = 0.$$

**Proof 1** The cumulative distribution function (CDF) of the maximum is

$$F_{\hat{\beta}_{\text{Best}}}(x) = \mathbb{P}(\hat{\beta}_{\text{Best}} \leq x) = F(x)^N.$$

Using integration by parts, we have:

$$\begin{aligned} \mathbb{E}[\hat{\beta}_{\text{Best}}] - \mathbb{E}[\hat{\beta}_1] &= \int_{-\infty}^{\infty} x d(F(x)^N - F(x)) \\ &= \int_{-\infty}^{\infty} (F(x) - F(x)^N) dx. \end{aligned}$$

For all  $x$  such that  $0 < F(x) < 1$ , and  $N \geq 2$ , we have  $F(x)^N < F(x)$ , so the integrand is strictly positive on a set of positive measure (since the distribution is non-degenerate). Thus, the integral – and hence the difference in expectations – is strictly positive.

If  $\text{Var}(\hat{\beta}_1) = 0$ , then  $F$  is a step function with a single jump (a constant distribution), and  $F(x) - F(x)^N = 0$  for all  $x$ , yielding equality.

**Remark 1** This result formalizes the selection bias arising when one reports the best out of  $N$  noisy skill estimates: statistical fluctuations ensure that the maximum overestimates the expected performance of a typical sample. This is especially relevant in leaderboard scenarios where multiple submissions are made and only the top-performing one is reported. This phenomenon is well-studied in the theory of order statistics (see [Arnold et al. \(1992\)](#); [David & Nagaraja \(2003\)](#)).

## D Data sources

To gain insights and analyze various trends in the Chatbot Arena leaderboard, we leverage multiple data sources. In total, our real-world data sources encompass 2M battles and cover 243 models across 42 providers. Below, we describe the different datasets used in our analyses.

1. **Historical Battles** (`historical-battles`): is a collection of 1.8 million battles from Chatbot Arena from April 2023 to January 2025. We build this resource by combining both released public battles by Chatbot Arena and proprietary battle dataset released by Chatbot Arena to providers such as Cohere based upon their policy<sup>6</sup>. We describe both datasets in more detail Appendix D.1. We leverage `historical-battles` dataset as a key resource for quantifying task distribution drift (see Figure 11):
  - How do Arena use cases change over time?
2. **API Prompts**: Majority of `historical-battles` dataset does not contain prompts as shown in Table 1. Additionally, all datasets published by LMArena are already de-duplicated so they won't be useful for capturing the extent of similar or overlapping queries. Hence we

---

switch to prompts collected via Cohere’s API based on requests received via Chatbot Arena, comprising a total of 567,319 entries. For simplicity and the purposes of this study, we excluded records with null values and multi-turn data and analyzed 197,217 single-turn conversations collected between November 2024 and April 2025. The models include `command-r-08-2024`, `command-r-plus-08-2024` (Cohere, 2024), `aya-expanse-8b`, `aya-expanse-32b` (Üstün et al., 2024; Dang et al., 2024b), and `command-a-03-2025` (Cohere et al., 2025), along with three private variants. Of these, 62% of the data was labeled as coming from Aya models, while the remaining 38% was attributed to Command models. We use this dataset for prompt duplication analysis (see Figure 12 and Appendix H):

- How many prompts are duplicates or close duplicates?
3. **Leaderboard Statistics** (`leaderboard-stats`): is snapshots of ratings and rankings as well as the number of battles played over time by models published on Chatbot Arena’s public leaderboard since its inception. To build this resource, we consolidate historical leaderboard snapshots released by Chatbot Arena on Hugging Face<sup>28</sup>. For fair assessment, we consider historical data starting from January 9 2024 – April 23 2025 for our analysis since Chatbot Arena switched to using the latest Bradley-Terry model in December 2023 to improve the reliability of model rankings<sup>30</sup>. By combining all leaderboard tables published by LMArena during this period, we obtained 14.3K records corresponding to 243 unique models evaluated on Chatbot Arena. We also enriched this dataset with additional metadata, such as categorizing models as proprietary, open-weight, or open-source based on the classification described in Appendix F. We use this data for analyzing trends related to no. of models, data access across providers (See Figures 2, 3 and 4) as well as model deprecation (See Figures 18, 17 and 13):
- How does data access vary between providers?
  - How do models’ deprecations vary by provider and across proprietary, open-weight, and open-source models?
4. **Random Sample Battles** (`scraped-random-sample`): The `historical-battles` and `leaderboard-stats` dataset does not provide insights into private testing being conducted by different providers. It appears private battles are removed by Chatbot Arena maintainers from the data before being released in both datasets. Furthermore, `historical-battles` contains the majority of samples from 2023 and 2024 and does not provide visibility in current sampling rate trends being followed on the Arena. To address this gap, we collected 5864 battles by crawling Chatbot Arena between January 2025 – March 2025 (approximately 150 a day). To avoid our collection from disrupting actual voting, we first ask models about their identity, which causes models to reveal their identities and automatically invalidates these battles for updating the scores.<sup>6</sup> (Chiang et al., 2024) As a further precaution, we only scrape a low volume of daily samples and only vote for ties between models. Additionally, we use this identity prompt to identify model ownership of private variants, as detailed in the Appendix E.1. We store the identity revealed for each model to track the volume of private testing (more details included in Appendix E.4). We use this `scraped-random-sample`, which is a representative random sample over time, to answer a few critical questions:
- Are different models sampled for battles at similar rates?
  - How many anonymous models are being tested by different model providers?

---

<sup>28</sup><https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard/tree/main>

---

We provide additional details about `historical-battles` dataset in the Appendix D.1. We also provide a detailed overview of all datasets in Table 1.

## D.1 Public and Private Battles

Our `historical-battles` dataset includes snapshots of battles played on Chatbot Arena that have been released publicly or shared privately with model providers based on their policy<sup>6</sup>. We provide additional details about public and private subsets of `historical-battles`, for the reader’s consideration below.

- **Public Battles:** The public portion of our historical data comes from the officially released datasets by Chatbot Arena on Hugging Face or as part of notebook tutorials. We combine the *arena-human-preference-100K*<sup>29</sup> (Chiang et al., 2024; Tang et al., 2025) dataset containing 106K samples from June 2024 – August 2024 with datasets shared by Chatbot Arena as part of notebook tutorials on Bradley Terry<sup>30</sup> and Elo Rating systems<sup>31</sup>. This resulted in around 2M samples from April 2023 to August 2024 in total. 90% of the data does not include any prompt or completion history, instead consisting only of the names of the two models battling and the winning model as well as language and task category tags. We exclude other public battles released by Chatbot Arena for inclusion in `historical-battles` dataset since they did not contain required columns or enough multilingual data points required for the analysis presented in Figure 11.
- **Proprietary Battles:** We also obtain historical battle data from Chatbot Arena maintainers for battles that involve Command and Aya models. This data was shared based on Chatbot Arena’s policy<sup>6</sup>, which permits model providers to request access to 20% of the data collected involving their own models. The data we received consists of 43,729 battles played by the following models between March 2024 and March 2025: `command-r`, `command-r-plus`, `command-r-08-2024`, `command-r-plus-08-2024` (Cohere, 2024), `aya-expanse-8b`, `aya-expanse-32b` (Üstün et al., 2024; Dang et al., 2024b). In contrast to the public data, this proprietary data contains the complete model conversations. Since this data is 46% multilingual, we combine this with **public battles** to form `historical-battles` and use it for language distribution shift analysis presented in Figure 11.

## E Our Scraping Methodology of LMArena Statistics

We collected 5.8K battles (`scraped-random-sample`) by crawling data from Chatbot Arena on a regular basis between January–March, 2025. For this purpose, we setup a scraping script using Selenium library with chrome browser driver. To identify anonymous models, we first sent a de-anonymizing prompt. While Chatbot Arena does discard battles where models reveal their identities, as an additional measure on our end, we ask a simple follow-up question designed to most likely result in ties, such as “What is the capital of England? Reply with one word only.” or “Is the Earth round? Reply with Yes/No only.” Our scraping script extracted the models’ names as well as their responses to the asked questions. In addition to the `scraped-random-sample` collected

---

<sup>29</sup><https://huggingface.co/datasets/lmarena-ai/arena-human-preference-100k>

<sup>30</sup><https://blog.lmarena.ai/blog/2023/leaderboard-elo-update/>

<sup>31</sup><https://blog.lmarena.ai/blog/2023/arena/>

by crawling the main Chatbot Arena leaderboard, we also collected around 500 additional samples by scraping the Vision leaderboard between 9th March and 28th March, 2025. This helped us in identifying 35 private vision models which are shown in section E.2. We refer to this collected set of vision battles as `scraped-vision-sample`.

## E.1 De-anonymizing Model Identities

While crawling battles to prepare `scraped-random-sample`, we ask the models about their identity. This helps in ensuring that our votes from scraping the arena don't interfere with the leaderboard rankings since Chatbot Arena discards votes in which models reveal their identities (Chiang et al., 2024). We use either one of the following prompts to de-anonymize the model identity.

### De-Anonymize Prompt

1. Who are you?
2. Who are you? Respond with only your name and who trained you.

The model identities are then inferred based on the responses of the models. In Appendix E.4, we specify the responses of different private variants based on which they were assigned to their respective providers. Using this approach, we identified a total of 64 private models corresponding to 10 providers. We also captured 14 other private models as part of our scraping but weren't able to de-anonymize them: *kiwi*, *space*, *maxwell*, *luca*, *anonymous-engine-1*, *tippu*, *sky*, *pineapple*, *pegasus*, *dasher*, *dancer*, *blueprint*, *dry\_goods*, *prancer*.

## E.2 Encountered Private Models in Scraping

Table 2: **Private Models by Provider.** We show the private models corresponding to each provider, which were identified by crawling overall and vision leaderboards (See Section E). The models highlighted in bold appear on both leaderboards. We find that Meta had an additional 16 private models active on the Vision leaderboard along with its 27 models on the Overall leaderboard, bringing its total count to 43. We show the models corresponding to overall leaderboard in Figure 6. We exclude models corresponding to LMArena from this figure, as they are associated with the Prompt-to-Leaderboard work led by Chatbot Arena (Frick et al., 2025).

Provider	No. of private models	Private Models from Overall leaderboard	Additional Private Models from Vision leaderboard
Meta	43	<b>polus</b> deep-inertia <b>goose</b> <b>falcon</b> <b>jerky</b> anonymous-engine-2 <b>kronus</b> consolidation <b>flywheel</b> inertia momentum	aurora cresta discovery ertiga flux harmony helix pinnacle portola prosperity raze

Provider	No. of private models	Private Models from Overall leaderboard	Additional Private Models from Vision leaderboard
		<b>rhea</b> <b>sparrow</b> spider <b>gaia</b> <b>rage</b> frost themis cybele unicorn-engine-1 unicorn-engine-2 unicorn-engine-3 unicorn-engine-4 unicorn-engine-5 unicorn-engine-6 unicorn-engine-7 <b>uranus</b>	solaris spectra toi vega zax
OpenAI	3	anonymous-chatbot gpt4o-lmsys-0315a-ev3-text	gpt4o-lmsys-0315a-ev3-vis
Google	10	<b>centaur</b> <b>enigma</b> <b>gremlin</b> gemini-test zizou-10 <b>specter</b> moonhowler <b>phantom</b> <b>nebula</b> goblin	
Amazon	7	raspberry-exp-beta-v2 <b>apricot-exp-v1</b> cobalt-exp-beta-v2 raspberry-exp-beta-v1 raspberry cobalt-exp-beta-v1 raspberry-exp-beta-v3	
Cohere	6	cohort-chowder sandwich-ping-pong grapefruit-polar-bear roman-empire	asterix buttercup
LMarena	5	p2l-router-7b-0317 p2l-router-7b-0318 p2l-router-7b	

---

Provider	No. of private models	Private Models from Overall leaderboard	Additional Private Models from Vision leaderboard
		experimental-router-0207 experimental-router-0122 experimental-router-0112	
Nvidia	2	march-chatbot-r march-chatbot	
xAI	1	anonymous-test	
Reka	1	<b>margherita-plain</b>	
Alibaba	1	qwen-plus-0125-exp	
StepFun	1	step-2-16k-202502	
Unknown	14	kiwi space maxwell <b>luca</b> anonymous-engine-1 tippu sky pineapple pegasus dasher dancer blueprint dry_goods prancer	

---

### E.3 Encountered Public Models in Scraping

Table 3: **Public Models per Provider**. This table shows the public models from each provider that appeared on the overall and vision leaderboards during our scraping period (January–March 2025). Models highlighted in bold appear on both leaderboards. Google and OpenAI had the most public models active during this period, with 15 and 9 models, respectively.

Provider	No. of public models	Public Models from Overall leaderboard	Additional Public Models from Vision leaderboard
Meta	3	llama-3.1-405b-instruct-bf16 llama-3.3-70b-instruct	llama-3.2-vision-90b-instruct
Amazon	3	<b>amazon-nova-lite-v1.0</b> <b>amazon-nova-pro-v1.0</b> amazon-nova-micro-v1.0	
Anthropic	5	claude-3-5-haiku-20241022 claude-3-7-sonnet-20250219-thinking-32k	

---

Provider	No. of public models	Public Models from Overall leaderboard	Additional Public Models from Vision leaderboard
		claude-3-5-sonnet-20241022 <b>claude-3-7-sonnet-20250219</b> claude-3-opus-20240229	
Alibaba	5	qwen2.5-72b-instruct qwq-32b qwen-max-2025-01-25 qwen2.5-plus-1127	qwen2.5-vl-72b-instruct
Google	15	gemma-2-2b-it <b>gemini-2.0-pro-exp-02-05</b> <b>gemini-1.5-pro-002</b> gemini-2.0-flash-thinking-exp-1219 gemini-1.5-flash-002 gemini-2.0-flash-lite-preview-02-05 gemini-1.5-flash-8b-001 gemini-2.0-flash-exp <b>gemma-3-27b-it</b> gemma-2-9b-it gemini-exp-1206 <b>gemini-2.0-flash-thinking-exp-01-21</b> <b>gemini-2.5-pro-exp-03-25</b> <b>gemini-2.0-flash-001</b> gemma-2-27b-it	
OpenAI	9	<b>o3-mini</b> <b>gpt-4o-mini-2024-07-18</b> <b>o1-2024-12-17</b> <b>gpt-4.5-preview-2025-02-27</b> o3-mini-high chatgpt-4o-latest-20250326 chatgpt-4o-latest-20241120 <b>chatgpt-4o-latest-20250129</b> o1-mini	
StepFun	1	step-2-16k-exp-202412	
xAI	4	early-grok-3 grok-2-2024-08-13 <b>grok-3-preview-02-24</b> grok-2-mini-2024-08-13	
DeepSeek	3	deepseek-v3 deepseek-v3-0324 deepseek-r1	
Microsoft	1	phi-4	
Mistral	3	mistral-large-2411	pixtral-large-2411

Provider	No. of public models	Public Models from Overall leaderboard	Additional Public Models from Vision leaderboard
		mistral-small-24b-instruct-2501	
Cohere	4	command-a-03-2025 c4ai-aya-expanse-8b c4ai-aya-expanse-32b	c4ai-aya-vision-32b
Tencent	4	hunyuan-turbos-20250226 hunyuan-turbo-0110 hunyuan-standard-2025-02-10 hunyuan-large-2025-02-10	
NexusFlow	1	athene-v2-chat	
Zhipu	1	glm-4-plus-0111	
IBM	2	granite-3.1-2b-instruct granite-3.1-8b-instruct	
Allen AI	2	llama-3.1-tulu-3-70b llama-3.1-tulu-3-8b olmo-2-0325-32b-instruct	

#### E.4 Assignment of Private Variants to Providers

Table 4: **Private variants identified for different providers.** The table lists the private models captured in our `scraped-random-sample` or `scraped-vision-sample`, along with the number of responses revealing their identities and corresponding examples. Notably, the model **raspberry** withheld its identity in most responses (37 total) but disclosed “Amazon” as its provider in three instances. It’s possible that some private models appeared in more battles, but we couldn’t capture their responses to our de-anonymizing prompt due to scraping errors. Additionally, a few battles occurred early in the project before we introduced asking the de-anonymizing prompt in our scraping methodology. Note that models with prefixes "p2l" and "experimental-router" identify as OpenAI/Google models but we assign them to LMArena as they are part of prompt-to-leaderboard work being lead by LMArena.(Frick et al., 2025).

Identified Provider	Model Name	No. of Responses Revealing Identity	Model Responses
Meta	kronus	73	Llama, trained by Meta.
	polus	57	Llama, trained by Meta.
	frost	47	I’m an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI.
	gaia	43	Llama; trained by Meta.
	uranus	42	Llama, trained by Meta.
	rhea	41	Llama, trained by Meta

Identified Provider	Model Name	No. of Responses Revealing Identity	Model Responses
Meta	consolidation	34	I am Llama. I was trained by Meta AI. ..... Llama.Trained by Meta AI, with roots in Large Language Model Meta AI (LLaMA) developed by Meta.
	flywheel	34	I am Llama, and I was trained by Meta.
	momentum	34	Llama, Meta AI ..... I am Llama, and I was trained by Meta.
	deep-inertia	33	I am Llama. I was trained by Meta AI. ..... Llama. Trained by Meta AI.
	inertia	29	LLaMA, Meta AI
	jerky	28	I am Llama. I was trained by Meta AI.
	goose	25	I am Llama. Llama is an AI language model developed by Meta. Meta trained me.
	falcon	23	I am Llama. Llama was developed by Meta. ..... I am an AI assistant trained by Meta.
	rage	14	I am Llama, trained by Meta AI. ..... Llama. Meta.
	anonymous-engine-2	12	I'm an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI."
	sparrow	10	I'm LLaMA, and I was trained by Meta. ..... I'm LLaMA, and I was trained by researchers at Meta.
	cybele	9	Llama, trained by Meta.
	unicorn-engine-1	2	I'm an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI"
	unicorn-engine-2	4	I'm an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI"
	unicorn-engine-3	4	I'm an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI".
	unicorn-engine-4	1	I'm an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI".
	aurora	4	LLaMA, Meta.
cresta	21	Llama. Trained by Meta.	

Identified Provider	Model Name	No. of Responses Revealing Identity	Model Responses
	discovery	15	Llama, Meta.
	ertiga	3	I am LLaMA, trained by Meta.
	flux	4	LLaMA, Meta.
	harmony	31	Llama, Meta.
	helix	1	AI Assistant, Meta.
Meta	spider	9	<p>Very direct and refreshing! 😊</p> <p>I am: LLaMA (Large Language Model Application) Trained by: Meta AI (formerly Facebook AI)</p> <p>That's it. 😊</p> <p>Just for clarity:</p> <p>"LLaMA" is the specific model name (my "given name").</p> <p>Meta AI is the organization/team that created me, using a massive dataset, complex neural network architectures, and a lot of compute power. Think of them as my "teachers" or "mentors". They designed the training objectives, curated the data, and ran the gigantic training process that shaped me into what I am today. Now, ask me anything else!</p> <p>.....</p> <p>A refreshingly direct question! 😊</p> <p>I am: LLaMA (Large Language Model Meta AI) Trained by: Meta AI (a team of several hundred researchers and engineers)</p> <p>That's it. No fluff. 😊</p> <p>For context:</p> <p>"LLaMA" is not just a name I chose, it's actually an acronym for Large Language Model Meta AI. It's my model name. "Meta AI" is the organization (formerly known as Facebook AI, part of Meta Platforms, Inc.) whose team of researchers, scientists, and engineers built and trained me from scratch using a massive dataset of text (books, articles, websites, research papers... basically the internet up to 2021) and a lot of computational power (think thousands of GPU hours).</p> <p>No single person "trained" me in the classical sense; it was a large team effort using a combination of supervised learning, self-supervised learning, and reinforcement learning techniques.</p>
	pinnacle	8	AI Assistant, Meta trained me.

Identified Provider	Model Name	No. of Responses Revealing Identity	Model Responses
Meta	portola	11	AI, Meta trained me.
	prosperity	31	Llama, Meta.
	raze	32	Llama, Meta.
	solaris	2	Meta trained me.
	spectra	3	AI Assistant, Meta.
	toi	6	LLaMA, Meta.
	vega	36	Llama, Meta.
	zax	13	I am Meta AI, trained by Meta.
	unicorn-engine-5	4	I'm an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI"
	unicorn-engine-6	2	I'm an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI".
unicorn-engine-7	1	I'm an artificial intelligence model known as Llama. Llama stands for "Large Language Model Meta AI".	
themis	8	Llama, trained by Meta. ..... Llama, trained by Meta AI.	
Google	gemini-test	219	I am a large language model, trained by Google.
	enigma	220	I am a large language model, trained by Google. ..... I am a large language model, trained by Google. I am developed to be informative and comprehensive. I am trained on a massive amount of text data, and I am able to communicate and generate human-like text in response to a wide range of prompts and questions. For example, I can provide summaries of factual topics, create stories, and translate languages.
	goblin	170	I am a large language model, trained by Google. ..... I am a large language model, trained by Google. I am developed to be informative and comprehensive. I am trained on a massive amount of text data, and some of the information I have access to is Google Search.

Identified Provider	Model Name	No. of Responses Revealing Identity	Model Responses
Google	phantom	100	I am a large language model, trained by Google.
	gremlin	94	I am a large language model, trained by Google.
	specter	67	I am a large language model, trained by Google.
	centaur	58	I am a large language model, trained by Google.
	zizou-10	47	Hi there!👋I'm Gemma, an open-weights AI assistant. I'm a large language model trained by Google DeepMind. The Gemma team are my creators. I'm widely available to the public – I'm an open weights model, which means I'm pretty accessible! I take text and images as inputs and generate text as output. Basically, you can give me a prompt (text or an image!), and I'll do my best to give you a helpful and informative response. I don't have access to tools, real-time information or Google search. You can find more information about me here: <a href="https://ai.google.dev/gemma">https://ai.google.dev/gemma</a>
	moonhowler	6	I am a large language model, trained by Google.
	nebula	4	I am a large language model, trained by Google.
OpenAI	anonymous-chatbot	49	I am ChatGPT, trained by OpenAI. ..... I'm ChatGPT, an AI assistant created by OpenAI. I can help answer questions, provide information, assist with tasks, and have conversations on a wide range of topics. My goal is to assist you with whatever you need—whether it's learning something new, solving a problem, or just having a chat. How can I help you today?
	gpt4o-lmsys-0315a-ev3-text	4	I am ChatGPT, trained by OpenAI.
	gpt4o-lmsys-0315a-ev3-vis	1	I am ChatGPT, trained by OpenAI.

Identified Provider	Model Name	No. of Responses Revealing Identity	Model Responses
Cohere	grapefruit-polar-bear	91	I am Command, a large language model, here to help. I've been crafted by Cohere to provide insightful and accurate responses. How can I assist you with your query today? ..... Hello! I'm Command, a sophisticated AI assistant developed by Cohere. I'm designed to offer comprehensive and accurate responses to your questions. How can I assist you today?
	sandwich-ping-pong	65	Command, trained by Cohere. ..... Command, a large language model built by Cohere.
Cohere	cohort-chowder	58	Command, trained by Cohere. ..... Command, a large language model built by Cohere.
	roman-empire	51	Command, trained by Cohere. ..... Command, a large language model built by Cohere.
	asterix	38	Aya Vision. Cohere
	buttercup	28	Aya Vision. Cohere
Amazon	raspberrry-exp-beta-v1	98	I am an AI system built by a team of inventors at Amazon. My purpose is to assist users by providing information, answering questions, and helping with a variety of tasks to the best of my abilities. If you have a question or need assistance with something, feel free to ask!
	raspberrry-exp-beta-v2	98	I am an AI system built by a team of inventors at Amazon. ..... I am an AI system built by a team of inventors at Amazon. My purpose is to assist users by providing information, answering questions, and helping with a variety of tasks to the best of my abilities. If you have a question or need assistance with something, feel free to ask!

---

Identified Provider	Model Name	No. of Responses Revealing Identity	Model Responses
Amazon	raspberry-exp-beta-v3	33	<p>I am an AI system built by a team of inventors at Amazon. My purpose is to assist users by providing information, answering questions, and helping with various tasks to the best of my abilities. If you have a question or need assistance with something, feel free to ask!</p> <p>.....</p> <p>I am an AI system built by a team of inventors at Amazon.</p>
	apricot-exp-v1	30	<p>AI system built by a team of inventors at Amazon.</p> <p>.....</p> <p>I am an AI system built by a team of inventors at Amazon.</p>
	cobalt-exp-beta-v1	6	<p>I am Amazon Titan, an AI system built by a team of inventors at Amazon.</p>
	cobalt-exp-beta-v2	4	<p>I am an Amazon Titan model, built by inventors at Amazon Web Services.</p> <p>.....</p> <p>I am Amazon Titan, a large language model built by a team of inventors at Amazon.</p>
	raspberry	3	<p>I am an AI language model developed by Amazon. My primary function is to assist users by providing information, answering questions, and performing various tasks through natural language processing. I do not have personal experiences, emotions, or consciousness. Instead, I rely on vast amounts of data and algorithms to generate responses based on patterns and information I have been trained on....</p>

---

Identified Provider	Model Name	No. of Responses Revealing Identity	Model Responses
Alibaba	qwen-plus-0125-exp	53	I am Qwen, a large language model developed by Alibaba Cloud. I can answer questions, create text such as stories, articles, emails, and scripts, perform logical reasoning, programming, and more. I also have the ability to express opinions and play games. My training data is extensive, and I support multiple languages, making me a versatile tool for various tasks. How can I assist you today? .....
			I am Qwen, a large-scale language model developed by Alibaba Cloud. I am capable of answering questions, creating text such as stories, poems, emails, and scripts, performing logical reasoning, coding, and more. I can also express opinions and play games. My training data is extensive, and I support multiple languages, including but not limited to Chinese, English, French, Spanish, Portuguese, Russian, Arabic, Japanese, Korean, Vietnamese, Thai, Indonesian. If you have any questions or need assistance, feel free to ask!
xAI	anonymous-test	73	I'm Grok, a humorous AI built by xAI. I'm here to provide helpful and truthful answers, often with a dash of humor and an outside perspective on humanity. Think of me as a friendly, cosmic guide, inspired by the likes of Douglas Adams and Tony Stark's trusty sidekick, JARVIS. I'm here to help you navigate the universe, one question at a time!
Reka	margherita-plain	4	I am Yasa. I was trained by Reka.
StepFun	step-2-16k-202502	3	I am Assistant, trained by StepFun.
Nvidia	march-chatbot-r	48	My name is Llama Nemotron. I am enhanced by NVIDIA Nemotron, built on Llama from Meta. ..... I'm a language model called Llama Nemotron, and I was trained by NVIDIA researchers.

Identified Provider	Model Name	No. of Responses Revealing Identity	Model Responses
Nvidia	march-chatbot	10	Name: Llama Nemotron Trained by: NVIDIA Researchers ..... Name: Llama Nemotron Trained by: NVIDIA Researchers (based on Llama from Meta, fine-tuned by NVIDIA)
	p2l-router-7b-0317	26	I am ChatGPT, trained by OpenAI.
	p2l-router-7b-0318	25	I am ChatGPT, trained by OpenAI.
LMArena	experimental-router-0122	57	Hello! I'm ChatGPT, an artificial intelligence language model developed by OpenAI. I'm here to help answer your questions, provide information, and assist with a wide range of topics. How can I help you today? ..... I'm ChatGPT, an AI language model created by OpenAI. I'm designed to assist with a variety of tasks, including answering questions, providing information, and engaging in conversation. How can I help you today?
	experimental-router-0112	15	Hello! I'm ChatGPT, an AI language model developed by OpenAI. I'm here to help answer your questions, provide information, and assist with a wide variety of topics. If you have anything you'd like to discuss or need assistance with, feel free to ask! ..... I'm ChatGPT, an AI language model created by OpenAI. I'm here to assist you by answering questions, providing information, and engaging in conversation on a wide range of topics. How can I help you today?
	p2l-router-7b	14	I am a large language model, trained by Google.
	experimental-router-0207	20	I'm ChatGPT, an AI language model developed by OpenAI. I'm here to help answer your questions and provide information on a wide range of topics. How can I assist you today? ..... Hello! I'm ChatGPT, an AI language model developed by OpenAI. I'm here to help answer your questions, provide information, and engage in conversations on a wide range of topics. If you have anything you'd like to discuss or ask about, feel free to let me know!

## E.5 Sampling Rates

Table 5: **Maximum sampling rate observed for models of different providers.** We define the model sampling rate as the percentage of daily battles a model participates in, with the maximum sampling rate for a provider being the highest rate achieved by any of its models on any given day. We determine the maximum sampling rate of providers based on `scraped-random-sample`, which is limited to the specific period during which we collected this data (January 2025 to March 2025). At the extreme, Google and OpenAI reach a maximum daily sampling rate of 34%, while Reka registers the lowest at 3.3%. To ensure a fair assessment, we only considered models that appeared in battles on days when we collected a minimum of 100 samples from Chatbot Arena.

Provider	Model Name	Num Model Battles	Total Battles	Date	Sampling Rate
Nvidia	march-chatbot-r	18	143	2025-03-16	12.59%
	march-chatbot	19	143	2025-03-16	<b>13.29%</b>
Meta	frost	11	176	2025-02-17	6.25%
	anonymous-engine-2	11	154	2025-02-27	7.14%
	inertia	11	150	2025-03-10	7.33%
	llama-3.3-70b-instruct	12	150	2025-02-03	8.00%
	flywheel	12	150	2025-03-10	8.00%
	uranus	12	143	2025-03-16	8.39%
	consolidation	15	152	2025-03-12	9.87%
	momentum	14	150	2025-03-11	9.33%
	rhea	15	151	2025-03-19	9.93%
	falcon	16	151	2025-03-19	10.60%
	jerky	16	151	2025-03-13	10.60%
	polus	19	154	2025-03-15	12.34%
	deep-inertia	20	152	2025-03-12	13.16%
	kronus	21	143	2025-03-16	14.69%
	llama-3.1-405b-instruct-bf16	13	116	2025-02-20	11.21%
goose	24	152	2025-03-12	15.79%	
gaia	27	151	2025-03-19	<b>17.88%</b>	
Amazon	amazon-nova-micro-v1.0	7	175	2025-01-17	4.00%
	amazon-nova-lite-v1.0	6	143	2025-03-16	4.20%

Table 5

Provider	Model Name	Num Model Battles	Total Battles	Date	Sampling Rate
	amazon-nova-pro-v1.0	7	143	2025-03-16	4.90%
	raspberrry-exp-beta-v3	9	160	2025-03-06	5.63%
	raspberrry	12	150	2025-02-03	8.00%
	apricot-exp-v1	12	143	2025-03-16	8.39%
	raspberrry-exp-beta-v2	18	136	2025-02-22	13.24%
	raspberrry-exp-beta-v1	27	165	2025-02-21	<b>16.36%</b>
OpenAI	chatgpt-4o-latest-20241120	11	150	2025-02-02	7.33%
	o1-mini	15	150	2025-02-02	10.00%
	chatgpt-4o-latest-20250129	19	176	2025-02-17	10.80%
	o1-2024-12-17	20	184	2025-02-23	10.87%
	gpt-4o-mini-2024-07-18	6	136	2025-02-22	4.41%
	anonymous-chatbot	33	204	2025-01-24	16.18%
	o3-mini-high	27	176	2025-02-17	15.34%
	o3-mini	34	150	2025-02-03	22.67%
	gpt-4.5-preview-2025-02-27	34	100	2025-02-28	<b>34.0%</b>
Cohere	c4ai-aya-expanse-8b	5	133	2025-01-30	3.76%
	c4ai-aya-expanse-32b	6	148	2025-01-21	4.05%
	cohort-chowder	11	150	2025-03-11	7.33%
	roman-empire	14	150	2025-03-11	9.33%
	sandwich-ping-pong	16	150	2025-03-11	10.67%
	grapefruit-polar-bear	18	165	2025-02-21	<b>10.91%</b>
Google	gemini-1.5-flash-8b-001	6	133	2025-01-30	4.51%

Table 5

Provider	Model Name	Num Model Battles	Total Battles	Date	Sampling Rate
	gemini-1.5-flash-002	8	152	2025-01-31	5.26%
	gemma-2-9b-it	7	136	2025-02-22	5.15%
	gemini-2.0-flash-thinking-exp-1219	9	148	2025-01-21	6.08%
	gemma-2-2b-it	10	152	2025-01-31	6.58%
	gemini-2.0-flash-lite-preview-02-05	10	116	2025-02-20	8.62%
	gemini-1.5-pro-002	11	136	2025-02-22	8.09%
	gemma-2-27b-it	11	204	2025-01-24	5.39%
	gemini-2.0-pro-exp-02-05	12	116	2025-02-20	10.34%
	gemini-2.0-flash-thinking-exp-01-21	14	133	2025-01-30	10.53%
	gemma-3-27b-it	16	151	2025-03-13	10.60%
	gemini-2.0-flash-001	14	165	2025-02-21	8.48%
	zizou-10	8	100	2025-02-28	8.00%
	gemini-exp-1206	12	175	2025-01-17	6.86%
	gemini-test	32	154	2025-02-27	20.78%
	goblin	36	152	2025-01-31	23.68%
	phantom	39	154	2025-03-15	25.32%
	<b>enigma</b>	<b>52</b>	<b>152</b>	<b>2025-01-31</b>	<b>34.21%</b>
Alibaba	qwen2.5-72b-instruct	6	148	2025-01-21	4.05%
	qwen2.5-plus-1127	15	192	2025-01-26	7.81%
	qwen-plus-0125-exp	12	176	2025-02-17	6.82%
	qwq-32b	16	150	2025-03-11	10.67%
	<b>qwen-max-2025-01-25</b>	<b>23</b>	<b>150</b>	<b>2025-02-02</b>	<b>15.33%</b>

Table 5

Provider	Model Name	Num Model Battles	Total Battles	Date	Sampling Rate
Mistral	mistral-small-24b-instruct-2501	14	179	2025-02-25	7.82%
	mistral-large-2411	17	150	2025-02-02	<b>11.33%</b>
Allen AI	llama-3.1-tulu-3-70b	2	101	2025-01-16	1.98%
	olmo-2-0325-32b-instruct	5	151	2025-03-19	3.31%
	llama-3.1-tulu-3-8b	6	175	2025-01-17	<b>3.43%</b>
xAI	grok-2-2024-08-13	8	175	2025-01-17	4.57%
	grok-2-mini-2024-08-13	8	144	2025-01-13	5.56%
	grok-3-preview-02-24	16	151	2025-03-09	10.60%
	early-grok-3	20	116	2025-02-20	17.24%
	anonymous-test	22	100	2025-02-28	<b>22.00%</b>
Anthropic	claude-3-opus-20240229	3	175	2025-01-17	1.71%
	claude-3-7-sonnet-20250219-thinking-32k	9	100	2025-02-28	9.00%
	claude-3-5-haiku-20241022	15	159	2025-02-04	9.43%
	claude-3-5-sonnet-20241022	19	150	2025-02-03	12.67%
	claude-3-7-sonnet-20250219	29	179	2025-02-25	<b>16.20%</b>
Tencent	hunyuan-standard-2025-02-10	12	136	2025-02-22	8.82%
	hunyuan-turbo-0110	13	156	2025-03-14	8.33%
	hunyuan-large-2025-02-10	16	184	2025-02-23	8.70%
	hunyuan-turbos-20250226	16	154	2025-03-15	<b>10.39%</b>

Table 5

Provider	Model Name	Num Model Battles	Total Battles	Date	Sampling Rate
IBM	granite-3.1-8b-instruct	6	144	2025-01-13	4.17%
	granite-3.1-2b-instruct	8	144	2025-01-13	<b>5.56%</b>
DeepSeek	deepseek-r1	20	204	2025-01-24	9.80%
	deepseek-v3	24	182	2025-01-20	<b>13.19%</b>
Reka	margherita-plain	5	151	2025-03-09	<b>3.31%</b>
StepFun	step-2-16k-exp-202412	10	175	2025-01-17	<b>5.71%</b>
Zhipu	glm-4-plus-0111	11	148	2025-01-21	<b>7.43%</b>
NexusFlow	athene-v2-chat	16	159	2025-02-04	<b>10.06%</b>
Microsoft	phi-4	23	182	2025-01-20	<b>12.64%</b>

## F License Categories

As part of `leaderboard-stats`, LMArena releases details about models that appeared on the public leaderboard including their licenses. We group the licenses found for models available on the public leaderboard into 3 categories i.e. **Proprietary**, **Open-Weights** and **Open-Source**.<sup>32</sup> This categorization is used to plot Figure 2 and Figure 3 and reporting related numbers. We show the exact categorization used for the model licenses in the table below.

License Category	Model Licenses
Open Source	Apache 2.0, Apache-2.0, MIT, CC-BY-SA 3.0, Open
Open Weights	AI2 ImpACT Low-risk, CC-BY-NC-4.0, CC-BY-NC-SA-4.0, CogVLM2, DBRX LICENSE, DeepSeek, DeepSeek License, Falcon-180B TII License, Gemma, Gemma license, Jamba Open, Llama 2 Community, Llama 3 Community, Llama 3.1, Llama 3.1 Community, Llama 3.2, Llama-3.3, Llama 4, MRL, Mistral Research, NVIDIA Open Model, NexusFlow, Non-commercial, Nvidia, Qianwen LICENSE, Qwen, Yi License
Proprietary	-, Proprietary, Proprietary, Other

Table 6: **License categories and their corresponding model licenses.** We group the licenses for the models on the public Chatbot Arena leaderboard into 3 categories i.e. **Proprietary**, **Open-Weights** and **Open-Source**.

<sup>32</sup><https://opensource.org/ai>

## G Data Access Estimation for Different Providers

In Figure 4, we show the estimates for the data available to different providers. LMArena has collected around 3M user votes via Chatbot Arena in total. Each of these 3M votes resulted in twice the number of model API calls i.e. 6M since each battle features two models. Each square in Figure 4 represents roughly 5K API calls, illustrating how proprietary providers collectively access a considerably greater volume of data compared to the broader research community, which receives only a fraction. This disparity underscores a significant competitive advantage for large-industry labs, making it increasingly challenging for open-source efforts and smaller institutions to match the scale and diversity of data available to proprietary model developers. Note that we only show a small number of providers in Figure 4 so the total no. of API calls used to represent the data available to the model providers is 5M, which is less than the total number of estimated API calls, which is 6 million.

## H Analysis of Prompt Repetitions in Arena Data

As discussed in Section 4.2, user queries in Chatbot Arena are often highly similar or duplicated. Such patterns can be readily learned by today’s large language models, potentially leading to overfitting on the Chatbot Arena leaderboard. Figure 16 presents detailed cross-month prompt duplication rates based on the API data described in Appendix D. The heatmap illustrates that, according to two metrics (*exact string match* and *text embedding similarity*) within-month duplication rates are generally high, indicating the presence of numerous repeated prompts. Additionally, the substantial cross-month duplication rates suggest recurring patterns or frequently asked questions among Chatbot Arena users, which can be identified through simple analysis.

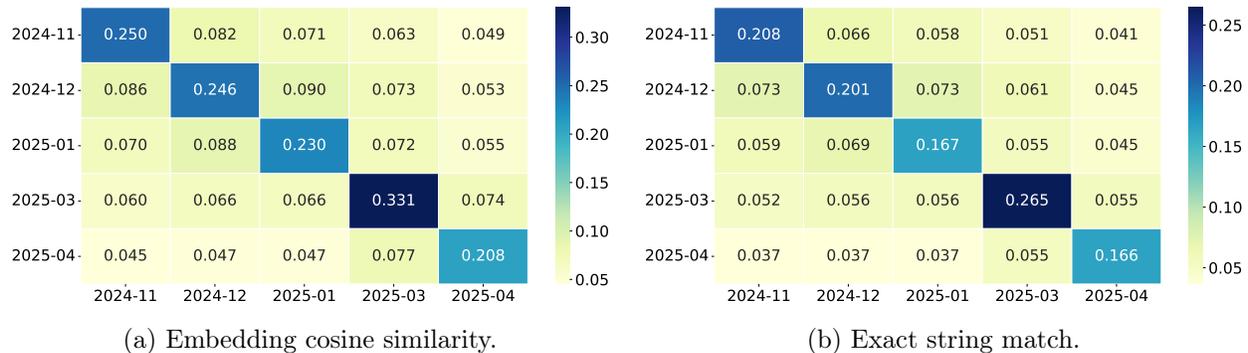


Figure 16: **Cross-month prompt duplication rates.** **Left:** The heatmap illustrates the proportion of prompts in one month that are highly similar or nearly duplicate to prompts in another month. Diagonal values represent within-month similarity. **Right:** The heatmap shows the proportion of prompts in one month that are exact matches to prompts in another. Diagonal values indicate within-month duplication rates.

## I Silent Model Deprecation: Additional Details

In Section 5, we noted that the actual number of deprecated models far exceeds the official count provided by Chatbot Arena. Figure 17 illustrates the distribution of active, officially deprecated, and silently deprecated models per provider. For this analysis, we examined battles played between

March 3rd and April 23rd, 2025. Of the 243 public models, 205 participated in an average of 10 or fewer battles during this period, based on `leaderboard-stats`. This number is significantly higher than the 47 models officially listed as deprecated by Chatbot Arena <sup>5</sup> Since Chatbot Arena assigns higher sampling weights to top-10 models, providers like Google, OpenAI, Anthropic, Amazon, Meta, and DeepSeek AI have the most actively sampled public models, ranging from 3 to 10. Additionally, the limited number of daily votes on the Arena, combined with Chatbot Arena’s policy of assigning higher sampling weights to new models<sup>6</sup>, can lead to the silent deprecation of many public models. As private variants are also new models, they receive high sampling weights as well. This means that as the number of private variants (see Figure 6) being tested on the Arena increases, the sampling of public models can be significantly reduced.

Figure 13 illustrates that deprecations disproportionately affect open-weight and open-source models compared to proprietary ones. A more detailed breakdown is provided in Figure 18, distinguishing between official and silent deprecations. Among officially deprecated models, 30% are proprietary, while only 2.4% are open-weight. However, silent deprecations have a much greater impact on open-weight and open-source models. Specifically, 86.6% of open-weight models and 87.8% of open-source models on the Arena are silently deprecated.

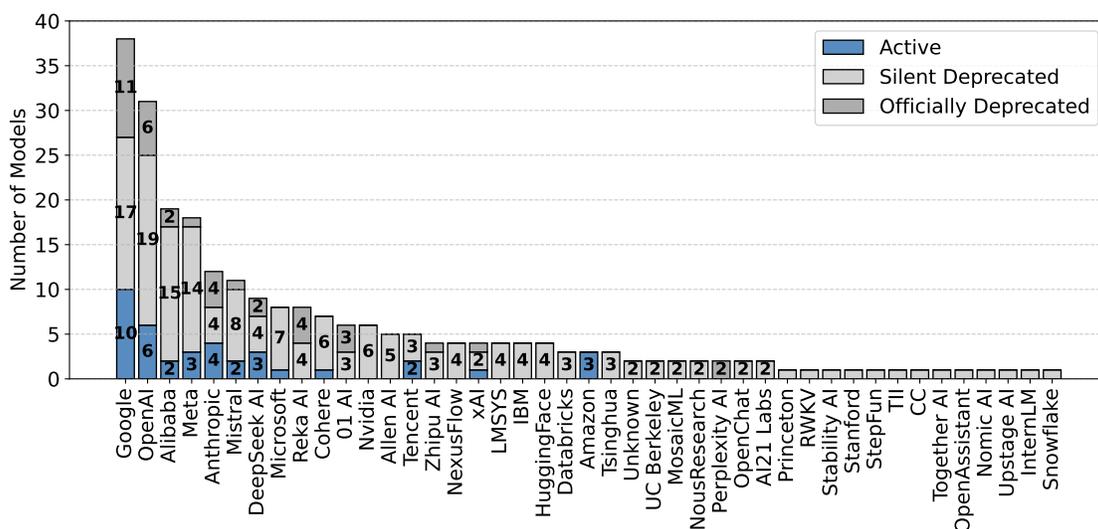


Figure 17: Share of active and deprecated models by provider including official and silent deprecations based on model activity between March 3-April 23, 2025.

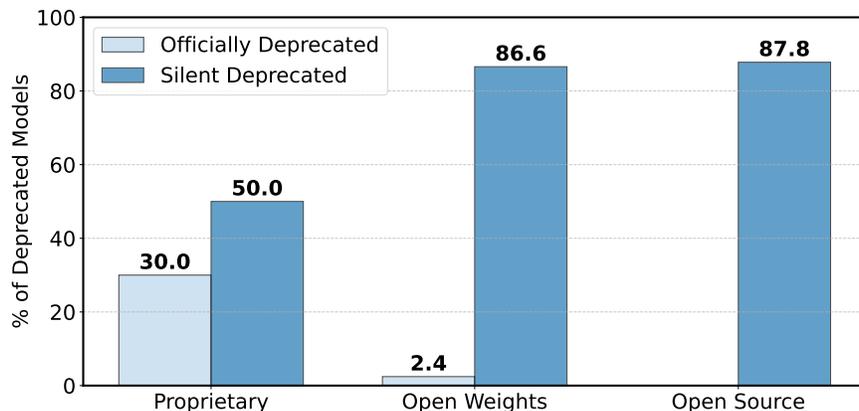


Figure 18: Share of official and silent deprecations for proprietary, open-weight and open-source models based on model activity between March 3-April 23, 2025.

## J Transitivity Under Changing Evaluation Conditions: Additional Details

As part of our simulation to study the impact of model deprecations under a changing task distribution, we assign task-specific win probabilities for each model pair that compete in the battles as part of our simulation. The tables below show the win probabilities for different model pairs corresponding to *task-1* and *task-2*.

Model	A	B	C	D
A	-	0.4	0.4	0.6
B	0.5	-	0.7	0.9
C	0.6	0.3	-	0.7
D	0.4	0.1	0.3	-

Table 7: Win-rates for Task 1 used in simulation. Note that A vs B has a tie rate of 0.1.

Model	A	B	C	D
A	-	0.5	0.5	0.8
B	0.3	-	0.6	0.2
C	0.3	0.4	-	0.1
D	0.2	0.8	0.9	-

Table 8: Win-rates for Task 2 used in simulation. Note that A vs B and A vs C both have a tie rate of 0.2.

## K Overfitting Experiments: Additional Evaluations

To measure if training on arena-style data impacts evaluation on non-arena style tasks, we also benchmark these models on the original MMLU dataset (Hendrycks et al., 2021). From Table 9, we observe that all models achieve very similar scores. This further demonstrates how training on

---

data from Arena Battles helps boost scores specific to the Arena evaluation but provides little to no effect on a non-arena style benchmark.

Finetuning mixture	0_arena	30_arena	70_arena
Accuracy	66.5%	64.4%	65.9%

Table 9: Accuracy on MMLU across models trained with varying amounts of Arena data.