# Investigation alleges preferential access and leaderboard skewing at Chatbot Arena



A new investigative paper authored by researchers from AI lab Cohere, alongside Stanford, MIT, and the Allen Institute for AI (Ai2), has raised allegations against LM Arena, the organisation managing Chatbot Arena, a widely referenced crowdsourced AI benchmarking platform. The paper claims LM Arena provided preferential access to certain leading AI firms, including Meta, OpenAI, Google, and Amazon, enabling them to test multiple AI model variants privately and withhold lower-performing results, thus skewing leaderboard rankings in their favour.

Chatbot Arena, launched in 2023 as an academic endeavour by UC Berkeley, operates by organising direct comparative 'battles' between different AI model outputs. Users vote for the better answer, with cumulative votes determining each model’s ranking on the platform's leaderboard. The presence of unreleased models competing pseudonymously is a frequent feature. LM Arena has positioned Chatbot Arena as an impartial, community-driven benchmarking tool.

However, the new study contends that select companies were granted opportunities for extensive private testing on the platform, a benefit not extended to all AI developers. For example, Meta reportedly conducted private tests for 27 model variants between January and March 2024 ahead of its Llama 4 launch, only publicly disclosing the best-performing single variant’s result, which ranked highly on the leaderboard.

Sara Hooker, Cohere’s Vice President of AI research and a co-author of the study, told TechCrunch, “Only a handful of [companies] were told that this private testing was available, and the amount of private testing that some [companies] received is just so much more than others. This is gamification.”

The authors analysed over 2.8 million Chatbot Arena battles conducted over five months, finding evidence indicating that increased exposure in model battles gave certain companies an advantage by enabling more extensive data gathering. According to the study, such additional access could improve performance by as much as 112% on LM Arena’s Arena Hard benchmark. LM Arena, however, counters that Arena Hard scores do not directly correspond to Chatbot Arena results.

LM Arena’s co-founder and UC Berkeley professor Ion Stoica responded to TechCrunch, characterising the study as containing "inaccuracies" and "questionable analysis." The organisation reiterated its commitment to fairness, stating, “If a model provider chooses to submit more tests than another model provider, this does not mean the second model provider is treated unfairly.”

Additional scrutiny came from Armand Joulin, a principal researcher at Google DeepMind, who disputed some of the study’s figures, clarifying that Google had only submitted one Gemma 3 model for pre-release testing. Hooker acknowledged these points and indicated the authors would issue corrections.

The paper also critiques Chatbot Arena’s transparency, suggesting that LM Arena should publicly disclose all private test scores and limit the number of private tests per AI lab to create a fairer competitive environment. While LM Arena rejected the suggestion to publish pre-release model scores, arguing that it is impractical since those models are not publicly accessible, it expressed openness to revising the sampling algorithm to ensure all models appear equally often in battles.

The investigation into preferential access on Chatbot Arena follows earlier revelations that Meta had optimised a Llama 4 variant specifically for conversational testing to enhance benchmark scores, without releasing that optimised variant publicly. LM Arena previously remarked that Meta should have been more transparent regarding this approach.

The study arrives as LM Arena recently announced plans to form a company and seek investment, raising questions about the independence and impartiality of private AI benchmarks amid growing industry reliance on them to demonstrate AI model performance.

Meta, Google, OpenAI, and Amazon, all implicated by the study, have yet to publicly respond to the allegations. Meanwhile, The TechCrunch report emphasises the ongoing debates in the AI community about benchmarking fairness and transparency as AI development accelerates.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://arxiv.org/abs/2403.04132> - This paper introduces Chatbot Arena, an open platform for evaluating large language models (LLMs) based on human preferences, detailing its methodology and operational history.
2. <https://en.wikipedia.org/wiki/Chatbot_Arena> - This Wikipedia article provides an overview of Chatbot Arena, including its launch date, purpose, and the major companies involved in supplying their LLMs for evaluation.
3. <https://arxiv.org/abs/2504.20879> - This study analyzes systematic issues in Chatbot Arena, highlighting how undisclosed private testing practices benefit certain providers and lead to biased leaderboard scores.
4. <https://www.noahwire.com> - This is the source of the original article discussing the allegations against LM Arena and Chatbot Arena, including the preferential access granted to certain AI firms.
5. <https://techcrunch.com/2025/04/30/ai-benchmarking-fairness/> - This TechCrunch article reports on the study's findings regarding preferential access in AI benchmarking and includes comments from Sara Hooker and Ion Stoica.
6. <https://www.cohere.ai/blog/ai-benchmarking-fairness> - This blog post from Cohere discusses the implications of the study on AI benchmarking fairness and transparency, authored by Sara Hooker.
7. <https://news.google.com/rss/articles/CBMioAFBVV95cUxPelhiY1BSQ2dyTzR5RmJ2bGxNN2w3N05EVlExaW9pYm96elZPaVc0ejl3dFo3SVdiU29RaTNJYWdaSEFiellJdkN6Y0pLNEh0YmViWWlOTms5cmJmZTBwTWVBa2RDMjJkWWJpZmRScjFra0p6aVE3Z3ZGTjRuMDVydU53eVNka2FmczJ3ZUlxclZuN3VELWw4Q0lmZ2I5eGRx?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data