

Democratic Perspectives and Institutional Capture of Crowdsourced Evaluations

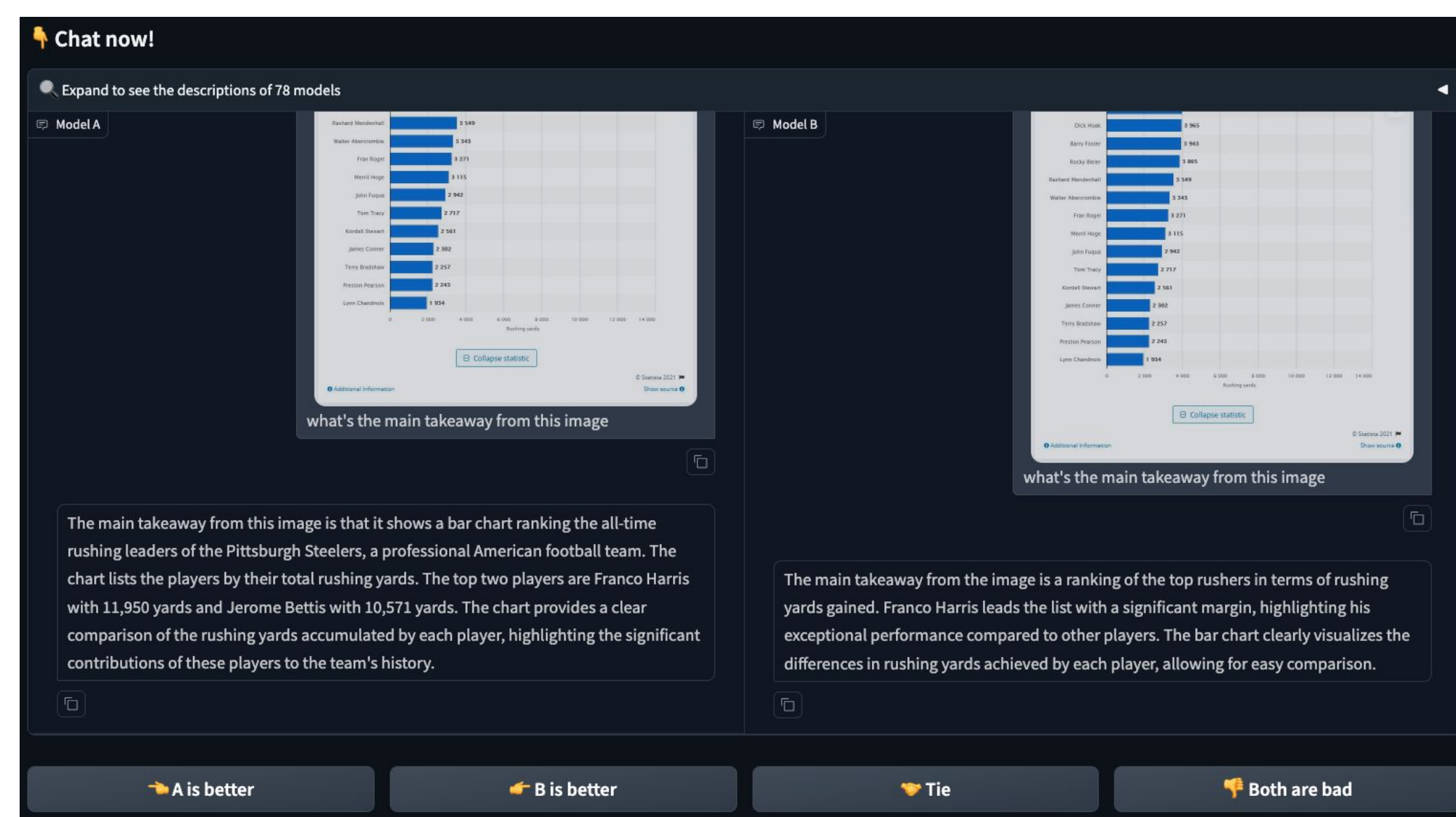
parth sarin, Michelle Bao

Democratic framings

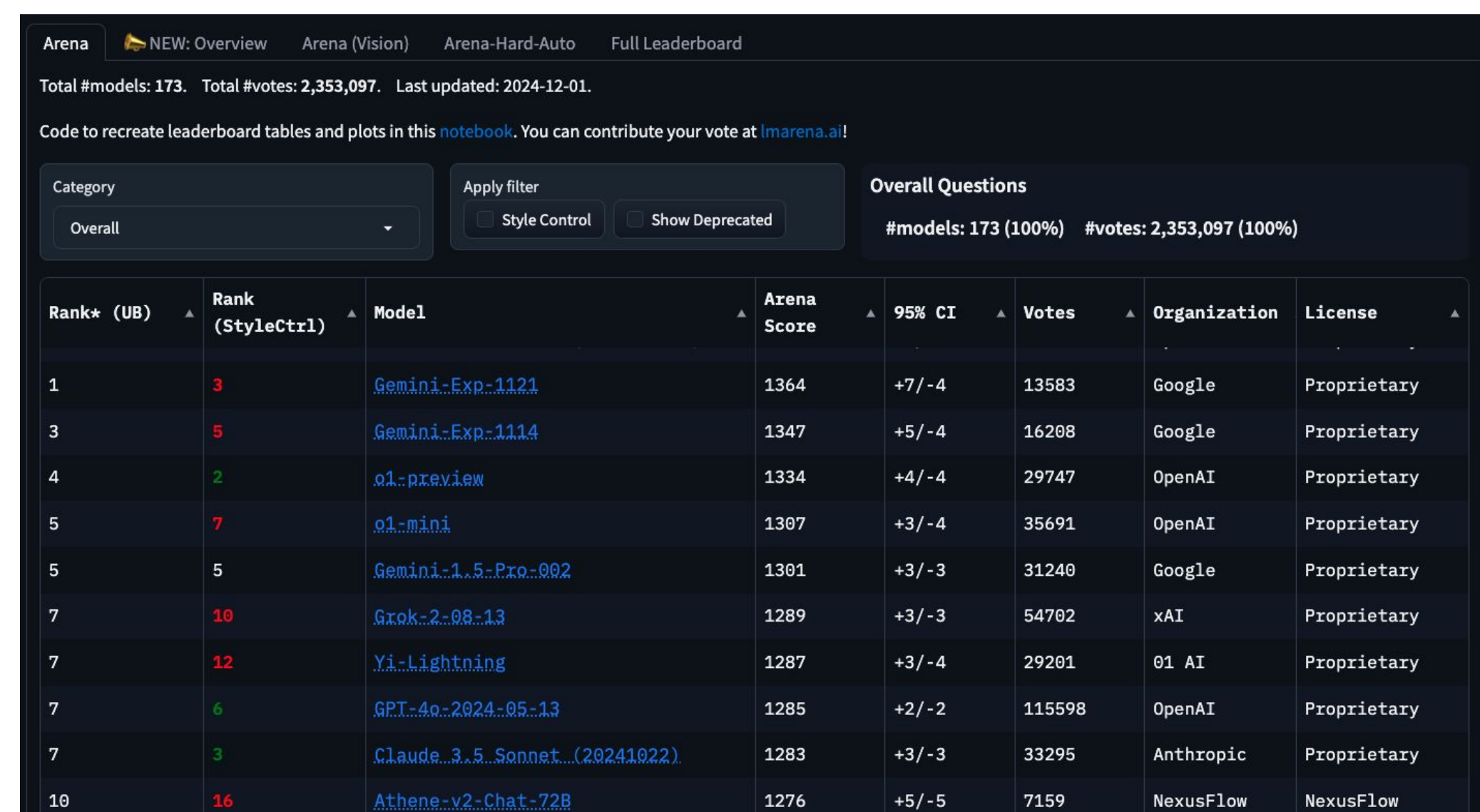
A growing trend in large language model (LLM) evaluation: AI companies and researchers framing the use of crowdsourced evaluations as a “democratization” of LLM development.

- OpenAssistant, a crowdsourced corpus of LLM conversations to “democratize research on aligning [LLMs]” [Köpf et al. 2023]
- “building an open [human] feedback platform” for anyone to contribute to evaluation [Don-Yehiya et al., 2024]

Crowdsourced evaluation programs that are framed as “democratic” mostly rate or rank LLM responses based on quality:



The user interface for Chatbot Arena

The image shows a screenshot of the Chatbot Arena LLM Leaderboard. It's a table with columns for Rank, Model, Arena Score, 95% CI, Votes, Organization, and License. The top models listed are Gemini-Exp-1121, Gemini-Exp-1114, o1-preview, o1-mini, Gemini-1.5-Pro-002, Grok-2-08-13, Yi-Lightning, GPT-4o-2024-05-13, Claude-3.5-Sonnet (20241022), and Athene-v2-Chat-72B.

Rank	Model	Arena Score	95% CI	Votes	Organization	License
1	Gemini-Exp-1121	1364	+7/-4	13583	Google	Proprietary
3	Gemini-Exp-1114	1347	+5/-4	16208	Google	Proprietary
4	o1-preview	1334	+4/-4	29747	OpenAI	Proprietary
5	o1-mini	1307	+3/-4	35691	OpenAI	Proprietary
5	Gemini-1.5-Pro-002	1301	+3/-3	31240	Google	Proprietary
7	Grok-2-08-13	1289	+3/-3	54702	xAI	Proprietary
7	Yi-Lightning	1287	+3/-4	29201	01 AI	Proprietary
7	GPT-4o-2024-05-13	1285	+2/-2	115598	OpenAI	Proprietary
7	Claude-3.5-Sonnet (20241022)	1283	+3/-3	33295	Anthropic	Proprietary
10	Athene-v2-Chat-72B	1276	+5/-5	7159	NexusFlow	NexusFlow

Chatbot Arena's LLM Leaderboard

Critiques

1 Technocratic containment of democracy

Crowdsourced evaluations generally solicit feedback in forms that are consumable by AI companies for further LLM development, hence advancing a technocratic containment of democracy.

In particular, crowdsourced evaluations exclude participation via *deliberation* and *discourse*, which are central to other online crowdsourcing movements like open source [Benoit-Barné, 2007].

Others have advocated for abandoning the idea that democracy should aim for consensus, arguing that public spaces are constituted by conflict and that dissent, disobedience, and difference are essential [Mouffe, 1999; Fraser, 1990].

2 Institutional capture

Companies that solicit crowdsourced evaluations seemingly aspire to make AI systems more accessible in a manner that requires they ingest the data of more minoritized users.

A user looking to be included in the democratic vision of crowdsourced evaluations must be willing to underwrite AI extractivity and make sacrifices in terms of “time, labor, attention, and data” to submit their preferences as expected by these systems [Crooks, 2024].

In the majority of these cases, crowdworkers are excluded from the governance of what they help produce, with little control over how models ingest their data or are deployed.

Tensions between critiques

One avenue to address the first critique involves incorporating new modes of participation into evaluations — deliberation, discourse, dissent, and disobedience.

However, as more individuals participate in evaluations that are engaging and time-demanding, more free labor is captured and exploitation reinforced. The social factory of crowdsourced evaluations can only become less exploitative as contributors gain the power to meaningfully shape LLM systems at all sites of the pipeline.

Call to action

Expanding collective power-building and governance is a continual and ongoing project, which can be done alongside immediate calls to action, including:

- Improve working conditions
- Expand evaluations beyond models to applications and use cases
- Grapple with the limitations of evaluations with respect to representation, applicability, and political values
- Include perspectives generally excluded from the AI development process

At the same time, contemporary implementations of democracy often neutralize and disarm dissent through inclusion and legitimization [Brown, 2015, Selinger, 2024]. Technological governance must engage with anti-institutional counter-hegemonic movements towards justice as well.

The seeming impossibility of addressing all critiques is neither necessary nor universal: we imagine a world in which the power dynamics of language models are fundamentally restructured and evaluations can contribute meaningfully to the democratic governance of sociotechnical ecosystems.

References

- C. Benoit-Barné. Socio-technical deliberation about free and open source software: Accounting for the status of artifacts in public life. *Quarterly Journal of Speech*, 93(2):211–235, 2007.
- W. Brown. *Undoing the Demos: Neoliberalism's Stealth Revolution*. Zone Books, 2015. ISBN 9781935408536. URL <http://www.jstor.org/stable/j.ctt17kk9p8>.
- R. N. Crooks. *Access is Capture: How Edtech Reproduces Racial Inequality*. University of California Press, 2024.
- N. Fraser. Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Social Text*, (25/26):56–80, 1990. ISSN 01642472, 15271951. URL <http://www.jstor.org/stable/466240>.
- A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z.-R. Tam, K. Stevens, A. Barhoum, N. M. Duc, O. Stanley, R. Nagyfi, S. ES, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, and A. Mattick. OpenAssistant Conversations – Democratizing Large Language Model Alignment. 2023. URL <https://arxiv.org/abs/2304.07327>.
- C. Mouffe. Deliberative democracy or agonistic pluralism? *Social research*, pages 745–758, 1999.
- E. Selinger. Can “tech criticism” tame silicon valley? *Los Angeles Review of Books*, November 2024. URL <https://lareviewofbooks.org/article/can-tech-criticism-tame-silicon-valley/>.