Democratic Perspectives and Institutional Capture of Crowdsourced Evaluations

parth sarin*
Stanford University
psarin@stanford.edu

Michelle Bao*
baom@cs.stanford.edu

This piece is a response to 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. A number of compensated and uncompensated crowdsourced LLM evaluations have emerged in the last two years. Köpf et al. [2023] introduced OpenAssistant, a crowdsourced corpus of LLM conversations to "democratize research on aligning [LLMs]." The 2023 DEF CON conference held the largest public AI red-teaming event [Cattell et al., 2023], and around the same time, several companies launched AI bug bounty programs [Page, 2023, OpenAI, 2023, Microsoft, 2023]. More recently, researchers built Chatbot Arena and ShareLM, tools that allow users to contribute LLM conversations to a crowdsourced dataset and vote on the best models [Chiang et al., 2024, Don-Yehiya et al., 2023]. Building on that work, [Don-Yehiya et al., 2024] have advocated for "building an open [human] feedback platform" for anyone to contribute to evaluation. Crowdsourced evaluations have been positively received by LLM development companies, many of which have used these platforms, datasets, and events to adapt their models [Dubey et al., 2024, Yang et al., 2024, Tong et al., 2024].

Crowdsourced evaluation programs that are framed as "democratic" mostly rate or rank LLM responses based on quality. On Chatbot Arena, contributors vote between two anonymized responses, a format that is easily adaptable to techniques like reinforcement learning from human feedback [Chiang et al., 2024, Ouyang et al., 2022]. OpenAssistant's contributors ranked responses and rated dimensions including quality, creativity, and humorousness on a Likert scale [Köpf et al., 2023].

Critiques of Crowdsourced Evaluations

AI companies often express an aspiration that "democratic," crowdsourced evaluations will make systems more *aligned* with the eventual user and use case. We seek to interrogate that claim by examining its presuppositions, i.e., the values advanced by these evaluation systems.

We offer two critiques of crowdsourced evaluations: First, because crowdsourced evaluations collect feedback in narrow modalities, they advance a technocratic containment of democracy which sacrifices diverse modes of participation in favor of modes designed to improve AI models. Second, current evaluations appeal to principles of social good to reframe the corporate capture of human labor.

A 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. This is consistent with what Subramonian et al. [2024] observed about how "democratization" is used in NLP research, namely that it signals broadened access to or use of technology. These evaluations prioritize scaling narrow modes of participation, neglecting modes that are enriching in democratic systems.

In particular, crowdsourced evaluations exclude participation via *deliberation* and *discourse*, which are central to other online crowdsourcing movements like open source [Benoit-Barné, 2007]. Allowing people in a democratic society to talk to each other enables them to build coalitions [Habermas, 1996, Steiner, 2012]; be in solidarity with one another against bias and harm [Calhoun, 2002]; and has been shown to reduce polarization [Fishkin, 2009]. In current evaluation arrangements, model behavior is generally guided by rules like ELO or the Borda count [Chiang et al., 2024, Siththaranjan et al., 2023], but these mathematical aggregation techniques are not neutral: they cannot be said to produce alignment when the evaluators' opinions were gathered without opportunity for deliberation.

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

*All authors contributed equally to this research.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

[Foucault, 1988, Brownlee, 2012, Mouffe, 1999, Arendt, 1972]. Radical and "agonistic" conceptions of democracy resist the deliberative democratic desire to squash dissent, depending on antagonism to challenge hegemonic power structures [Laclau and Mouffe, 2014]. For example, Fraser [1990] stresses the importance of "subaltern counterpublics" in renegotiating and altering power dynamics of exclusionary dominant discourses. There is a rich history of disobedience in the digital subaltern [Scheuerman, 2016, Gray and Suri, 2019]; and AI data workers have been protesting exploitative working conditions for years [Distributed AI Research Institute, 2024].

Contribution is 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]. The result is additional value from which evaluators are alienated, re-instantiating a common pattern of free labor in the digital economy: crowdsourcing allows AI companies to capture the value of the general public, creating a "social factory" wherein work processes are shifted onto society [Terranova, 2012, Scholz, 2017].[2]

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 on them. Model outputs that are more aligned with their preferences, without control over where or how these outputs are used, could result in increasingly effective nonconsensual applications of AI models.

Tensions between critiques. Our critique invites a reimagining of how collectives hold power to shape the development and usage of LLMs, while recognizing that such a process comes at a cost of efficiency and simplicity. One avenue to address the first critique involves incorporating new modes of participation into evaluations — deliberation, discourse, dissent, and disobedience — to expand the scope and impact of crowdworker input. Such a shift would indeed enrich the breadth and depth of contributions, particularly in challenging the presuppositions of LLM development.

However, doing so would exacerbate the negative impacts raised in the second critique. 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, development to deployment.

Call to Action

Expanding collective power-building and governance is a continual and ongoing project, and deeply dependent on institutional and community investment. This work can be done alongside immediate calls to action, some of which workers have been demanding for years like improved working conditions. Another tangible change involves expanding evaluations beyond models to applications and use cases, which are more related to AI models operationalized in sociotechnical contexts. Researchers and developers must also grapple with the limitations of evaluations — framed as "democratic" or not — with respect to representation, applicability, and underlying political values. Lastly, a democratic evaluation, requires the inclusion of perspectives generally excluded from the AI development process, especially marginalized groups subject to AI usage in carceral institutions, workplace surveillance, and at a country's borders.

At the same time, contemporary implementations of democracy can be fraught, as critical scholars have noted the neoliberal role that democratic institutions play in neutralizing, disarming, and suppressing dissent through inclusion and legitimization [Brown, 2015, Selinger, 2024]. While institutional reforms can alleviate the acute injustices of the present, technological governance must broadly engage with anti-institutional counter-hegemonic movements towards justice. These aspirations are complex, sometimes contradictory, and some more immediately realizable than others. The seeming impossibility of addressing all critiques, due to the narrowness of the normative conception of AI evaluation, 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.

[2] In the case of crowdsourced evaluations of non-market models (e.g. produced by governments or nonprofits), the dynamics of extraction remain largely the same. Many non-market projects are outsourced to companies [Electronic Privacy Information Center, 2023], and for in-house development, evaluators bolster nonprofit development or state control, violence, and care, with varying and limited agency over institutional power.

Limitations

Because we examine claims of democraticization proposed by researchers who do not specify the democratic principles they are advocating for, we explore multiple contrasting modes of participation, drawing from deliberative, consensus, and radical democratic principles. However, without advocating for a specific democratic perspective, we encountered a limitation in how concrete our calls to action could be, specifically in tangible suggestions for how evaluations can be less exploitative. Future work could build off ours by exploring the incentives, ownership, and labor dynamics of an evaluation and governance system under a particular framework, like radical democracy or democratic socialism.

Acknowledgments

We're incredibly grateful to Jenny Robinson, Alec Smecher, and John Willinsky for their thoughtful reflections on early drafts of this work. We'd also like to thank the three anonymous reviewers and EvalEval program chairs for their feedback. All of these revisions and insights were invaluable in strengthening this piece.

References

H. Arendt. *Crises of the republic: Lying in politics, civil disobedience on violence, thoughts on politics, and revolution*, volume 219, chapter Civil Disobedience. Houghton Mifflin Harcourt, 1972.

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.

K. Brownlee. *Conscience and conviction: The case for civil disobedience*. Oxford University Press, 2012.

C. J. Calhoun. Imagining solidarity: Cosmopolitanism, constitutional patriotism, and the public sphere. *Public culture*, 14(1):147–171, 2002.

S. Cattell, R. Chowdhury, and A. Carson. AI Village at DEF CON announces largest-ever public Generative AI Red Team, May 2023. URL https://aivillage.org/generative%20red%20team/generative-red-team/.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. 2024. URL https://arxiv.org/abs/2403.04132.

R. N. Crooks. *Access is Capture: How Edtech Reproduces Racial Inequality*. University of California Press, 2024.

Distributed AI Research Institute. Data Workers' Inquiry. https://data-workers.org/, 2024. (Accessed on 09/20/2024).

S. Don-Yehiya, L. Choshen, and O. Abend. ShareLM: Crowd-sourcing human feedback for open-source LLMs together. https://sharelm.github.io/, 2023. (Accessed on 09/16/2024).

S. Don-Yehiya, B. Burtenshaw, R. F. Astudillo, C. Osborne, M. Jaiswal, T.-S. Kuo, W. Zhao, I. Shenfeld, A. Peng, M. Yurochkin, et al. The Future of Open Human Feedback. *arXiv preprint arXiv:2408.16961*, 2024.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Electronic Privacy Information Center. Outsourced and Automated: How Government Agencies Are Using Private Contractors to Automate Decision-Making, 2023. URL https://epic.org/outsourced-automated/.

J. S. Fishkin. *When the people speak: Deliberative democracy and public consultation*. Oxford University Press, 2009.

M. Foucault. Interview. In J. Bernauer and D. Rasmussen, editors, *The Final Foucault*. MIT Press, 1988.

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.

M. L. Gray and S. Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

J. Habermas. *Between facts and norms: contributions to a discourse theory of law and democracy*. Polity Press, 1996.

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.

E. Laclau and C. Mouffe. *Hegemony and socialist strategy: Towards a radical democratic politics*, volume 8. Verso Books, 2014.

Microsoft. Microsoft AI Bounty Program. https://www.microsoft.com/en-us/msrc/bounty-ai, 2023. (Accessed on 09/16/2024).

C. Mouffe. Deliberative democracy or agonistic pluralism? *Social research*, pages 745–758, 1999.

OpenAI. Announcing OpenAI's Bug Bounty Program. https://openai.com/index/bug-bounty-program/, 2023. (Accessed on 09/16/2024).

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

C. Page. Google adds generative AI threats to its bug bounty program. https://techcrunch.com/2023/10/26/google-generative-ai-threats-bug-bounty/, 2023. (Accessed on 09/16/2024).

W. E. Scheuerman. Digital disobedience and the law. *New Political Science*, 38(3):299–314, 2016.

T. Scholz. *Uberworked and underpaid: How workers are disrupting the digital economy*. John Wiley & Sons, 2017.

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/.

A. Siththaranjan, C. Laidlaw, and D. Hadfield-Menell. Understanding Hidden Context in Preference Learning: Consequences for RLHF. In *Socially Responsible Language Modelling Research*, 2023.

J. Steiner. *Force of better argument in deliberation*, page 139–152. Cambridge University Press, 2012.

A. Subramonian, V. Gautam, D. Klakow, and Z. Talat. Understanding "Democratization" in NLP and ML Research. *arXiv preprint arXiv:2406.11598*, 2024.

T. Terranova. Free labor. In *Digital Labor*, pages 33–57. Routledge, 2012.

S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.