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 are increasingly promoting evaluations that rely on crowdsourced labor and framing this shift 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]." Last year, the DEF CON conference held the largest public AI red-teaming event designed in part to "[grow] the community" of LLM evaluators [Cattell et al., 2023]. Around the same time, several companies announced AI bug bounty programs, asking the general public to look for LLM vulnerabilities [Page, 2023, OpenAI, 2023, Microsoft, 2023]. More recently, researchers built Chatbot Arena and ShareLM, tools that allow LLM users to contribute their chats to a crowdsourced dataset and vote on which language model is the "best" [Chiang et al., 2024, Don-Yehiya et al., 2023]. Building on that work, others have advocated for expanding these efforts and "building an open [human] feedback platform" in which anyone can evaluate a language model as a way of participating in the improvement of LLMs [Don-Yehiya et al., 2024]. Crowdsourced evaluations have generally been received positively by companies that develop LLMs, many of whom have tested their models using these platforms, datasets, and events [Dubey et al., 2024, Yang et al., 2024, Tong et al., 2024].

Crowdsourced evaluation programs that are framed as "democratic" mostly solicit quality ratings for language model responses, though our arguments extend to product and application evaluations as well. On Chatbot Arena, an LLM evaluation platform, users submit their preferences by voting between two anonymized language model conversations, a format that is easily adaptable to the paradigm of reinforcement learning through human feedback [Chiang et al., 2024, Ouyang et al., 2022]. OpenAssistant's contributors rated model responses on a Likert scale across dimensions including quality, creativity, and humorousness, and ranked model replies to their queries [Köpf et al., 2023].

Critiques of Crowdsourced Evaluations

Those who promote crowdsourced evaluations generally express an aspiration that AI model development, guided by these evaluations, will produce systems that are more *aligned* with the eventual user and use case. We seek to interrogate that aspiration: What are presuppositions of the claim that a crowdsourced evaluation can produce an aligned system? Why should such an evaluation, with few parameters on participants and context, align AI systems with society's values? To complicate the matter, AI researchers generally articulate alignment as a "techno-moralistic exercise of training and evaluating LLMs" — one which primarily involves finding the right parameters in a high-dimensional vector space to capture "society's values" — rather than a complex condition of language, automation, and power enframed by societal values that vary in context and time [Hristova et al., 2024].

We offer two critiques of crowdsourced evaluations: First, because crowdsourced evaluations prioritize modes of engagement that are efficient and ingestible for ML models, they advance a technocratic

[*] All authors contributed equally to this research.

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

containment of democracy which sacrifices diverse modes of participation in favor of modes designed to improve AI models. And, second, current evaluations reframe of the corporate capture of human labor as a social good, appealing to principles of democratization while neglecting the cost of capture.

A technocratic containment of democracy. By design, "democratic" LLM 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. Crowdsourced evaluations have prioritized scaling narrow modes of participation, neglecting modes that are more enriching in a democratic society.

Most crowdsourced evaluations do not make room for *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 [Foucault, 1988, Brownlee, 2012, Mouffe, 1999, Arendt, 1972]. Radical and "agonistic" conceptions of democracy resist the deliberative democratic desire to squash dissent through consensus. Instead, they depend on antagonism to make oppressive power relations visible so their contradictions can be brought to a synthesis — that is, through solidarity and a "chain of equivalence," a diverse collection of demands can constitute a counter-hegemonic project that challenges dominant power structures [Laclau and Mouffe, 2014]. For example, Fraser [1990] points out that dominant discourses will always be exclusionary and stresses the importance of "subaltern counterpublics" in renegotiating and altering power dynamics.

There is a rich history of disobedience in the digital subaltern [Scheuerman, 2016, Gray and Suri, 2019]; and data workers in the machine learning development context have been protesting and organizing against exploitative working conditions for years [Distributed AI Research Institute, 2024]. Legitimizing and meaningfully addressing worker demands is a critical part of radical democratic theories with an anti-imperial and anti-oppression focus.

Contribution is capture. Companies that solicit crowdsourced evaluations ostensibly 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 endorse and underwrite the extractive regimes of AI technology 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 for AI companies from which evaluators are alienated, re-instantiating a common pattern of free labor in the digital economy: namely, crowdsourcing allows AI companies to capture the value of producers outside of their organization, creating a "social factory" wherein work processes are shifted onto society [Terranova, 2012, Scholz, 2017].

Even in the case of crowdsourced evaluations of non-market models (e.g. those produced by governments or nonprofits), the dynamics of extraction remain largely the same. Due to the costly nature of AI development, many non-market projects are outsourced to companies who earn profit while doing "social good" [Electronic Privacy Information Center, 2023]. Even when they are developed using government or nonprofit resources, evaluators' contributions still reify the pipeline of AI development and bolster the incentives of nonprofit development or state control, violence, and care.

Regardless of the type of institutional incentives, for the vast majority of these tools, crowdworkers remain excluded from opportunities to govern what they help produce. Evaluators, representing themselves and their communities, have little control over how their data is ingested, how they are subject to predictive models, and where they see AI-generated content. Shaping model outputs to be more aligned with one's preferences, without control over where or how these model outputs are used, could lead to increasingly compelling and targeted 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 the cost of efficiency and simplicity. In our present reality, most cutting-edge LLMs are developed by for-profit or academic institutions with highly centralized resources, LLMs are being leveraged into our day-to-day in an innumerable number of ways, and the vast majority of people cannot collectively contribute to key decisions. The question we are left with now is whether it is possible to operationalize democratic values in the context of an evaluation to resist power-centralizing dynamics and empower communities.

One avenue to address the first critique involves incorporating new modes of participation into evaluations — prioritizing deliberation, discourse, dissent, and disobedience — in order to more meaningfully 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 are encouraged to participate in evaluations that are increasingly engaging and time-demanding, this shift would deepen the exploitation of crowdworkers as more of their free labor is captured. The social factory of crowdsourced evaluations can only become less exploitative as contributors gain the power to more meaningfully shape LLM systems at all sites of the pipeline from development to deployment. Sufficiently addressing both critiques will involve fundamentally restructuring the power dynamics of language models to center marginalized communities and workers.

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 more 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 products, applications, and use cases, which are more closely related to operationalized AI models in sociotechnical contexts. Researchers and developers must also grapple with the limitations of evaluations — framed as "democratic" or not — with respect to representation and applicability, particularly interrogating the political values that such evaluations advance.

Democracy requires valuing inputs from and disagreements between diverse coalitions, beyond those who can easily participate in the dominant discourses. A democratic evaluation, then, requires the inclusion of communities that are generally excluded from the AI development participatory process, especially the most 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 and resistance through inclusion and legitimization [Brown, 2015]. While institutional reforms can alleviate the acute injustices of the present, evaluations must meaningfully engage with anti-institutional counter-hegemonic movements towards justice, which often emerge from grassroots organizing, intersectional community movements in the Majority World, and direct action.

These aspirations are complex and 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.

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.

T. Hristova, L. Magee, and K. Soldatic. The problem of alignment. *AI & Society*, pages 1–15, 2024.

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.

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.