Democratic Perspectives and Corporate Captures of Crowdsourced Evaluations

Anonymous Author(s)
Affiliation
Address
email

This piece is a response to a growing trend in the evaluations of large language models (LLMs): AI companies and researchers are increasingly promoting evaluations that rely on crowdsourced labor and framing these developments as a "democratization" of LLM evaluation and development.

A number of 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].

Currently, crowdsourced evaluations generally solicit quality ratings for model responses. For example, on Chatbot Arena, 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]. Contributors to OpenAssistant rated model responses on a Likert scale across dimensions including quality, creativity, and humorousness, and also 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 produced an aligned system? Why should such an evaluation, with few parameters on participants and context, help 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 sacrifice diverse modes of participation in a manner that narrows the sociotechnical imaginaries for contribution to model development. And, second, this technocratic containment of democracy reframes the corporate

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

capture of human labor as a social good, appealing to principles of democratization while neglecting the cost of capture.

A technocratic containment of democracy. The design of "democratic" LLM evaluations currently solicit feedback in forms that are consumable by AI companies for further LLM development, hence advancing a technocratic containment of democracy. At the same time, these evaluations preclude the kinds of participation that are most 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 [Jürgen, 1996, Steiner, 2012]; be in solidarity with one another against bias and harm [Calhoun, 2002]; and has been shown to reduce polarization and increase the likelihood of consensus [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 consensus or alignment when the evaluators' opinions were gathered without opportunity for deliberation.

Arendt [1972] also argued that *dissent* and *disobedience* are important for democracy and should be valued as a form of democratic participation. There is a rich history of digital disobediance, but the exploitation of data workers often remains undervalued and unheard [Scheuerman, 2016, Gray and Suri, 2019]. As these workers dissent to their experiences of participation in LLM evaluation and development, those protests ought to be recognized as legitimate and necessary in a democracy [Distributed AI Research Institute, 2024].

Contribution is capture. Companies that solicit crowdsourced evaluations ostensibly aspire to make AI systems more accessible in a manner that necessarily 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.

Moreover, there is a stark gap between shaping preferred model outputs and providing an evaluation of a model that has the power to shape its development and use. Evaluators, representing themselves and their communites, 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 and incorporated into our day-to-day in an innumerable number of ways, and the vast majority of people cannot collectively contribute to key decisions. In this context, one may wonder how to formulate an evaluation that resists these dynamics and empowers communities, or if it is even possible to create an evaluation that does not continue to reify hegemonic power dynamics.

We believe it is critical that evaluations begin to incorporate new modes of participation prioritizing deliberation, discourse, dissent, and disobedience, in order to more meaningfully expand the scope and impact of crowdworker input. Such a shift would enrich the breadth and depth of contributions, particularly contributions that challenge the presuppositions of LLM development. At the same time, as more individuals are encouraged to participate in evaluations that are more engaging and time-demanding ways, this recommendation may further exacerbate contribution as capture. The seeming impossibility of addressing both critiques, due to a neoliberal enframing of AI development, is neither necessary nor universal: we imagine a world in which the power dynamics of language models are fundamentally restructured and a freeing, democratic AI evaluation can be realized.

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.

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.

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

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

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

H. Jürgen. *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`.

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

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.

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.

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.