Responsible Crowdsourcing for Responsible Generative AI: Engaging Crowds in AI Auditing and Evaluation

Wesley Hanwen Deng [1], Mireia Yurrita [2], Mark Díaz [3], Jina Suh [4], Nick Judd [5], Lara Groves [6], Hong Shen*[1], Motahhare Eslami*[1], Kenneth Holstein*[1]

[1] Carnegie Mellon University
[2] Delft University of Technology
[3] Google Research
[4] Microsoft Research
[5] Digital Safety Research Institute
[6] Ada Lovelace Institute

hanwend@cs.cmu.edu, m.yurritasemperena@tudelft.nl, markdiaz@google.com, jinsuh@microsoft.com, Nick.Judd@ul.org, lgroves@adalovelaceinstitute.org, hongs@andrew.cmu.edu, meslami@cs.cmu.edu, kjholste@cs.cmu.edu

Abstract

With the rise of generative AI (GenAI), there has been an increased need for participation by large and diverse user bases in AI evaluation and auditing. GenAI developers are increasingly adopting crowdsourcing approaches to test and audit their AI products and services. However, it remains an open question how to design and deploy responsible and effective crowdsourcing pipelines for AI auditing and evaluation. This workshop aims to take a step towards bridging this gap. Our interdisciplinary team of organizers will work with workshop participants to explore several key questions, such as how to improve the output quality and workers' productivity for GenAI evaluation crowdsourcing tasks compared to discriminative AI systems, how to guide crowds in auditing problematic AI-generated content while managing their psychological impact, ensuring marginalized voices are heard, and setting up responsible and effective crowdsourcing pipelines for real-world GenAI evaluation. We hope this workshop will produce a research agenda and best practices for designing responsible crowd-based approaches to AI auditing and evaluation.

Introduction

Human computation and crowdsourcing approaches have played important roles in advancing AI research and practices, especially in areas such as data generation and annotation (Bigham, Bernstein, and Adar 2014; Vaughan 2017; Russakovsky et al. 2015), as well as human- and application-level evaluation (Vaughan 2017; Anastasiou and Gupta 2011; Zaidan and Callison-Burch 2011; Zhou et al. 2019). With the recent rise of generative AI (GenAI), many calls have been made to engage large samples from diverse populations in evaluating and auditing generative AI systems (Feffer et al. 2024; Anthropic 2023; Kenthapadi, Lakkaraju, and Rajani 2023). This is largely due to the proven ability of non-experts with relatively low tech-savviness and AI literacy to uncover problematic AI behaviors that might be overlooked by small groups of AI researchers and developers (Shen et al. 2021; DeVos et al. 2022; Lam et al. 2022).

Recent work has suggested that AI developers often leverage crowdsourcing platforms to engage more diverse populations in testing and auditing their AI systems (Deng et al. 2023a; Wang et al. 2023). With the emergence of generative AI, which can produce a wide variety of outputs, the need for crowd workers and diverse citizens to take on the role of auditors has become even more critical. This necessity is further underscored by AI red teaming efforts aimed at systematically testing AI systems to identify toxicities, hallucinations, vulnerabilities, biases, and potential misuse (The White House 2023; Microsoft 2023). However, it remains unclear how to establish effective crowdsourcing pipelines in real-world contexts (Deng et al. 2023a), how to manage the psychological impact on crowd workers and citizens when conducting AI red teaming (Zhang et al. 2024; Pendse et al. 2024), how to amplify minority voices when aggregating results (Sap et al. 2019), how to support AI practitioners in integrating crowdsourcing pipeline into their existing working flows (Deng et al. 2022; Yildirim et al. 2023), and how to fairly compensate crowd workers for auditing and evaluating GenAI systems (Gray and Suri 2019).

To this end, this workshop aims to explore responsible and effective crowdsourcing in generative AI evaluation and auditing. In particular, our goal is to collectively develop a research agenda and a set of best practices to design and develop a new generation of responsible crowdsourcing systems that can effectively engage diverse crowd workers in evaluating and auditing generative AI.

To achieve this goal, our workshop will bring together an interdisciplinary organizational team, across academia, industry, and civil society, with diverse research backgrounds and practical experience in designing, developing, and deploying crowdsourcing pipelines for AI evaluation and audit. We will build upon previous workshops on relevant topics successfully organized by the current organizers in other HCI venues such as CHI, CSCW, and FAccT (Deng et al. 2023b; Xiao et al. 2024). We also plan to take notes dur-

*These authors contributed equally.

ing the workshop and share the insights generated as a white paper.

In this workshop, we will aim to explore the following questions:

- How might we effectively guide and scaffold crowds and crowd workers in prompting and surfacing harmful AI-generated content, while managing the psychological impact of generative AI auditing and red-teaming?
- What are some appropriate ways to aggregate the auditing and evaluation results from crowds to ensure marginalized voices are heard and promote the value of diversity?
- How should we operationalize diversity in ways that constitute responsible practices across the wide range of auditing and evaluation tasks?
- How can we set up the crowdsourcing pipeline so that crowd workers from the crowdsourcing platforms can conduct GenAI audits and evaluations in more realistic contexts that are specific to their real world use cases?
- What is the fair compensation for crowd workers who participate in different types of generative AI audit and evaluation?
- How might we incentivize crowds to participate in generative AI audit and evaluation?
- For which kinds of tasks are crowdsourcing pipelines particularly ill-suited, and how might they be complemented by other auditing or evaluation approaches?

Workshop Format

We plan to accept 15 - 20 participants for our workshop. This six hours workshop (including two hours lunch and coffee break) consist of:

- **Welcome and Introduction** (10:00-10:10am): Organizers will welcome the participants, present the topic, and outline the format of the workshop session.
- **Lightning talk #1** *(10:10-10:20am)*: Psychological aspects for crowd workers conducting AI audits.
- **Lightning talk #2** *(10:20-10:30am)*: Diversity in AI audits and evaluation.
- **Lightning talk #3** *(10:30-10:40am)*: Mechanisms for scaffolding AI audits and evaluation.
- **Lightning talk #4** *(10:40-10:50am)*: Ecological validity for using crowdsourcing for AI audits and evaluation.
- **Interactive Q/A** *(10:50-11:30am)*: Workshop participants will engage in an interactive Q/A session with the four lightning talk speakers.
- **Lunch break** *(11:30-1:00pm)*: Participants will join different break out lunch groups with organizers.
- **Focus group activity #1** *(1:00-2:00pm)*: Identify current practices and challenges.
- **Coffee break** *(2:00-2:30pm)*.
- **Focus group activity #2** *(2:30-3:30pm)*: Explore opportunities and future research agenda.
- **Closing remarks** *(3:30-4:00pm)*.

Overall, workshop participants will spend around an hour listening to talks and three hours on interactive activities. In particular, the interactive Q/A and group activity facilitated by the organizers will provide participants with abundant opportunities to discuss and exchange ideas. To better engage participants in identifying current challenges and exploring future research agendas, we designated 2 hours for group activities in the afternoon sessions. We will share the concrete focus group activity based on the accepted work and the pre-workshop survey we will send to our participants in advance of the workshop. Please note that the concrete schedule and topics of the four lightning talks are tentative.

Recruiting and Dissemination Plan

Submissions to our workshop will be single-blind and reviewed and curated by the organizers. We will ensure that each submission is reviewed by at least two organizers who have no conflict of interest with the authors.

The submission deadline will be August 21, 2024. We will notify participants of their acceptance by September 6.

All workshop information and the Call for Participation will be publicly announced on social media platforms such as X, LinkedIn, and our workshop website. We also plan to disseminate the Call for Participation through our personal networks and email lists. We believe that we can reach a diverse group of potential participants, given that the organizers are from six institutions with diverse backgrounds and regions. We elaborate on this in the next section.

Diversity

Organizers and lightning speakers of our workshop represent **demographic diversity** (e.g., by gender, ethnic/racial background), **disciplinary backgrounds** (e.g., Human-Computer Interaction, Computer Science, Sociology, Public Policy, Communication), and **institutional diversity** (e.g., academic universities, for-profit industry organizations, and non-profit research institutes). Our workshop also encourages diverse viewpoints from interdisciplinary research studies. In our call for paper, we will specifically aim to engage systematically marginalized or vulnerable communities such as BIPOC, LGBTQIA+, or in non-Western geographical or cultural contexts in auditing and evaluating generative AI systems.

References

Anastasiou, D.; and Gupta, R. 2011. Comparison of crowdsourcing translation with Machine Translation. *Journal of Information Science*, 37(6): 637–659.

Anthropic. 2023. Frontier threats red teaming for AI Safety.

Bigham, J. P.; Bernstein, M. S.; and Adar, E. 2014. Human-Computer Interaction and Collective Intelligence.

Deng, W. H.; Guo, B.; Devrio, A.; Shen, H.; Eslami, M.; and Holstein, K. 2023a. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

Deng, W. H.; Lam, M. S.; Cabrera, Á. A.; Metaxa, D.; Eslami, M.; and Holstein, K. 2023b. Supporting User Engagement in Testing, Auditing, and Contesting AI. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, 556–559.

Deng, W. H.; Nagireddy, M.; Lee, M. S. A.; Singh, J.; Wu, Z. S.; Holstein, K.; and Zhu, H. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 473–484. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.

DeVos, A.; Dhabalia, A.; Shen, H.; Holstein, K.; and Eslami, M. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. *CHI Conference on Human Factors in Computing Systems*.

Feffer, M.; Sinha, A.; Lipton, Z. C.; and Heidari, H. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? *arXiv preprint arXiv:2401.15897*.

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

Kenthapadi, K.; Lakkaraju, H.; and Rajani, N. 2023. Generative ai meets responsible ai: Practical challenges and opportunities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5805–5806.

Lam, M. S.; Gordon, M. L.; Metaxa, D.; Hancock, J. T.; Landay, J. A.; and Bernstein, M. S. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Microsoft. 2023. Planning red teaming for large language models (LLMs) and their applications - Azure OpenAI Service.

Pendse, S. R.; Massachi, T.; Mahdavimoghaddam, J.; Butler, J.; Suh, J.; and Czerwinski, M. 2024. Towards Inclusive Futures for Worker Wellbeing. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–32.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

Shen, H.; DeVos, A.; Eslami, M.; and Holstein, K. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–29.

The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Vaughan, J. W. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.*, 18(1): 7026–7071.

Wang, Q.; Madaio, M. A.; Kapania, S.; Kane, S.; Terry, M.; Wilcox, L.; et al. 2023. Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges.

Xiao, Z.; Deng, W. H.; Lam, M. S.; Eslami, M.; Kim, J.; Lee, M.; and Liao, Q. V. 2024. Human-Centered Evaluation and Auditing of Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–6.

Yildirim, N.; Pushkarna, M.; Goyal, N.; Wattenberg, M.; and Viégas, F. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People+ AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–13.

Zaidan, O.; and Callison-Burch, C. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 1220–1229.

Zhang, A. Q.; Shaw, R.; Anthis, J. R.; Milton, A.; Tseng, E.; Suh, J.; Ahmad, L.; Kumar, R. S. S.; Posada, J.; Shestakofsky, B.; et al. 2024. The Human Factor in AI Red Teaming: Perspectives from Social and Collaborative Computing. *arXiv preprint arXiv:2407.07786*.

Zhou, S.; Gordon, M.; Krishna, R.; Narcomey, A.; Fei-Fei, L. F.; and Bernstein, M. 2019. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems*, 32.