EMILY SHENG

ewsheng at gmail · website · linkedin

SUMMARY

Applied researcher specializing in evaluation of generative AI systems, with experience spanning LLM evaluation, responsible AI, NLP, fairness, and machine learning. 10+ years of industry and research experience, with a track record of translating research questions into empirical methods, evaluation frameworks, and practical tooling.

EXPERIENCE

SR APPLIED SCIENTIST May 2022 – Present
Sociotechnical Alignment Center at Microsoft Research *Seattle, WA*

- Developed methods for measuring and evaluating safety and capabilities of generative AI systems, translating emerging research questions into practical evaluation frameworks for AI practitioners
- Led research side of collaboration across research, science, policy, and engineering to turn evaluation research into open-source evaluation framework (ASSERT) for improving generative AI systems
- Designed and implemented empirical studies to test rigor of evaluation framework across different target concepts and against other existing approaches
- Mentored several interns on LLM evaluation and responsible AI, leading to academic and open-source impact

MACHINE LEARNING ENGINEER July 2021 – May 2022
Twitter *Los Angeles, CA*
Improved content health & fairness of the platform

RESEARCH INTERN May 2019 – April 2020
Google *Mountain View, CA*
Evaluated and mitigated biases towards different demographics in a poetry composition system

SOFTWARE ENGINEERING INTERN May 2018 – Aug 2018
Google *Mountain View, CA*
Evaluated textual similarity across text lengths for bag-of-words and convolutional neural network model variants

SOFTWARE ENGINEER July 2014 – July 2015
Expect Labs *San Francisco, CA*
Prototyped classifier for domain-specific named entity recognition to improve a natural language understanding system, and full-stack development of developer platform tools

SKILLS

- **AI/ML:** generative AI (GenAI) evaluation, NLP, safety, responsible AI, fairness, automated evaluation
- **Methods:** experimental design, human/LLM-as-a-judge annotation, measurement validity, disagreement/error analysis, dataset construction
- **Programming:** Python

SELECTED PUBLICATIONS

Fogliato, R., Palia, A., Wang, X., **Sheng, E**., Atalla, C., Garcia-Gathright, J., Pangakis, N., Tan, S., Vann, D., Washington, H., Dow, P. A., Elfardy, H., Wallach, H., & Atluri, S. (2026). ASSERT: A measurement pipeline for GenAI audits. In *EMNLP Industry Track 2026.*

Agarwal, D.*, **Sheng, E.***, Atalla, C., Garcia-Gathright, J., Mozannar, H., Washington, H., Chouldechova, A., Barocas, S., & Wallach, H. (2026). AI-assisted systematization for evaluating GenAI systems.

Harvey, E., **Sheng, E.**, Blodgett, S. L., Chouldechova, A., Garcia-Gathright, J., Olteanu, A., & Wallach, H. (2025). Understanding and meeting practitioner needs when measuring representational harms caused by LLM-based systems. In *Findings of the ACL 2025.*

Wallach, H., Desai, M., Cooper, A.F., Wang, A., Atalla, C., Barocas, S., Blodgett, S.L., Chouldechova, A., Corvi, E., Dow, P.A., Garcia-Gathright, J., Olteanu, A., Pangakis, N.J., Reed, S., **Sheng, E.**, Vann, D., Vaughan, J.W., Vogel, M., Washington, H., & Jacobs, A.Z. (2025). Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In *Proceedings of the 42nd ICML.*

Chouldechova, A., Atalla, C., Barocas, S., Cooper, A. F., Corvi, E., Dow, P. A., Garcia-Gathright, J., Pangakis, N., Reed, S., **Sheng, E.**, Vann, D., Vogel, M., Washington, H., & Wallach, H. (2024). A shared standard for valid measurement of generative AI systems' capabilities, risks, and impacts. In *NeurIPS 2024 Workshop on Statistical Foundations of LLMs and Foundation Models (SFLLM).*

Magooda, A., Helyar, A., Jackson, K., Sullivan, D., Atalla, C., **Sheng, E.**, Vann, D., Edgar, R., Palangi, H., Lutz, R., Kong, H., Yun, V., Kamal, E., Zarfati, F., Wallach, H., Bird, S., & Chen, M. (2023). A framework for automated measurement of responsible AI harms in generative AI applications.

Dev, S.*, **Sheng, E.***, Zhao, J.*, Amstutz, A.*, Sun, J., Hou, Y., Sanseverino, M., Kim, J., Peng, N., Chang, K.-W. (2022). On Measures of Biases and Harms in NLP. In *Proceedings of AACL-IJCNLP-Findings 2022*.

Sheng, E. (Aug 2021). Fairness in Natural Language Generation*. PhD Thesis.*

EDUCATION

PH.D. IN COMPUTER SCIENCE
University of Southern California
Advisors: Dr. Prem Natarajan and Dr. Nanyun Peng

M.S. IN COMPUTER SCIENCE
University of Southern California

B.A. IN COMPUTER SCIENCE, COGNITIVE SCIENCE
University of California, Berkeley

PROFESSIONAL SERVICE

Reviewer for *ACL, EMNLP, FAccT, COLM, AAAI; Publicity Co-Chair (NAACL 2022)