

Rahul Gupta

Senior Applied Science Manager

(213)587-1103
rahul.1987iit@gmail.com



HIGHLIGHTS

- Senior leader for Responsible AI in Amazon's AGI organization, reporting to VP-level leadership and driving safety, evaluation, red-teaming, fairness, privacy, and mitigation programs for frontier models.
- During my tenure at Amazon, led teams as large as 50+ across Amazon Nova safety, LMArena performance, and the Nova Trusted AI Challenge, connecting research strategy to external evaluation and product outcomes.
- Hands-on scientist with first-author work on [CBRN uplift evaluation for frontier language models](#) and [fairness in large-scale NLU systems](#); 25+ patented or patent-pending technologies and publications across NeurIPS, ICLR, ACL, EMNLP, FAccT, and ICASSP.

TECHNICAL EXPERIENCE

Senior Applied Science Manager June 2026 – Present
Amazon/ Frontier AI Assets, Responsible AI Cambridge, MA

- Overall head of Responsible AI in the newly architected AGI division, reporting directly to a Vice President.

Senior Applied Science Manager October 2022 – May 2026
Amazon AGI/ Nova Responsible AI Cambridge, MA

- Lead Nova models to be realized as leaders in safety by third party evaluators [\[1\]](#), [\[2\]](#), [\[3\]](#).
- Lead the Nova LMArena team, driving Nova models to top 15 on LMArena, and the Nova Trusted AI Challenge, generating 100+ applications and 10 participating university teams.

Applied Science Manager December 2018 – September 2022
Amazon Alexa/ Responsible AI Cambridge, MA

- Founded and led a team to work on Responsible AI research applied to Alexa, building foundational AI safety capabilities into Alexa's platform and products.
- Drove goals to enhance private and unbiased customer experiences, with outcomes reported to Amazon Senior Vice Presidents.

Applied Scientist August 2016 – November 2018
Amazon Alexa/ NLU at scale Cambridge, MA

- Served as a technical lead on a project to decouple training and testing of large scale ML models.
- Created significant business impact leading to an organizational restructuring.

EDUCATION

PhD, Department of Electrical Engineering, University of Southern California 2010 – 2016

Bachelor of Technology, Department of Electrical Engineering, Indian Institute of Technology, Kharagpur 2006 – 2010

RESEARCH LEADERSHIP & SERVICE

- Organizer of the [TrustNLP](#) workshop series and several other workshops [\[1\]](#), [\[2\]](#). Reviewer for top tier conferences including ACL, ICLR and ICASSP.
- Point of contact for several [Academic Research Awards](#) (ARA) collaborations at Amazon, collectively representing over \$20M in funded research.
- Keynote speaker at reputed venues such as [IEEE SLT](#) and other workshops [\[1\]](#), [\[2\]](#).

KEY PUBLICATIONS

Full list available at my [google scholar page](#).

- [BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation](#), ACM FACCT 2021. (700+ citations)
- [“I’m fully who I am”: Towards centering transgender and non-binary voices to measure biases in open language generation](#), ACM FACCT 2023.
- [FLIRT: Feedback Loop In-context Red Teaming](#), EMNLP 2024. Extensive coverage in the red teaming community [1], [2]; #3 paper of the day on Hugging Face.
- [Establishing Best Practices for Building Rigorous Agentic Benchmarks](#), NeurIPS 2025. Winner of the AgentX LLM Agents MOOC Competition, 2025.
- [A Threshold Exceedance Framework for CBRN Uplift Evaluation in Frontier Language Models](#), arXiv preprint, 2026. First-author, hands-on experimentation.

OTHER RECOGNITIONS

- Ming Hsieh department of Electrical Engineering best PhD pitch award, 2015.
- Two-time winner of Interspeech computational paralinguistics challenge (2013, 2015).
- Phi Beta Kappa alumni in Southern California scholarship, 2015.
- Ming Hsieh department of Electrical Engineering humanitarian award, 2014.
- Provost Ph.D. fellowship, 2010-2014.
- Undergraduate recognitions: InfoUSA exchange Scholar, OPJems scholarship, Rajendra memorial scholarship, 2009.