Avinandan Bose

PhD Candidate · CSE, University of Washington

avibose@uw.edu | avinandan22.github.io | Avinandan22 | Google Scholar

Research Interests

As LLMs get more capable, they are deployed across increasingly diverse domains, demographics, and user populations, and the bottleneck shifts from capability to *usability and safety*. What each user needs is sparse, latent, task-conditioned, and high-dimensional, yet rarely articulated upfront and sometimes adversarial. My research develops **post-training methods**, **large-scale evaluation benchmarks**, and **theoretical foundations** across this challenge, where understanding diverse users is both a usability and safety problem. My [**Research Overview**] describes how these themes connect across the following directions:

- **Few-Shot Personalization from Memory**: low-rank reward modeling (LoRe) [1], meta-learning transfer guarantees [2], democratic viewpoint representation via social choice theory [3]
- **Inference-Time Personalization via Multi-Turn Interactions**: evaluation of preference discovery in frontier models (PrefDisco) [4], Bayesian preference elicitation (PEP) [5], hybrid RLHF theory [7], cold-start initialization for diverse populations [8]
- **AI Safety & Robustness**: agent security evaluation (DoomArena) [9], certified defenses against adaptive data poisoning [10], backdoor injection via fine-tuning [11]
- **Mechanism Design for Multi-Agent Systems**: strategic optimization under bounded rationality [12],[13], network interdiction [14], fleet coordination [15]

Education

University of Washington *Seattle, WA*

PhD, Computer Science and Engineering *Sep. 2022 - Present*

- Advisors: Maryam Fazel, Simon S. Du.
- Dissertation: *Personalizing Intelligence: Learning For and From Diverse Users Through Active Interaction*.

Indian Institute of Technology Kanpur (IIT K) *Kanpur, India*

B.Tech, Computer Science and Engineering *Jul. 2018 - Dec. 2021*

- Advisors: Piyush Rai, Ketan Rajawat. GPA: 9.2/10.0.
- **Academic Excellence Award** (2019, 2020, 2021): top 10 percentile. **Class of 1990 Scholarship** (2018): awarded to top three freshmen.

Research Experience

Meta Superintelligence Labs FAIR *Advisor: Lin Xiao*

Visiting Research Scientist *Sep. 2024 - Present*

- Developed LoRe [1], exploiting intrinsic low-dimensionality of human preferences to learn personalized reward models from 5–10 comparisons without retraining; 12% improvement on the largest public preference datasets (1,500+ participants, 75 countries, 5 languages).
- Built PEP [5], decomposing preference elicitation into offline correlation learning and online Bayesian inference; ∼10K parameters outperform RL fine-tuning of an 8B-parameter LLM with 3–5× fewer interactions and 2× higher adaptivity.
- Formalizing democratic representation of diverse viewpoints in LLM responses via social choice theory [3], recovering the full spectrum of population preferences rather than collapsing to a majority view.

University of Washington *Advisors: Maryam Fazel, Simon S. Du*

Graduate Research Assistant *Sep. 2022 - Present*

- Built PrefDisco [4], the first large-scale evaluation of interactive preference discovery (21 models, 10 benchmarks, 10K user-task scenarios); found that 29% of elicitation attempts degrade alignment and frontier models fail to ask appropriate clarifying questions.
- Co-led DoomArena [9], a plug-in security evaluation for LLM agents (3 agentic benchmarks, ∼500 tasks); combined attacks reach 97% ASR and standard guardrails are largely ineffective. In Silent Sabotage [11], showed 5% poisoned fine-tuning traces embed stealthy backdoors while improving task performance.
- Certified robustness bounds against adaptive data poisoning via robust control theory [10]; first matching bounds on sample complexity for hybrid RLHF [7]; near-optimal cold-start guarantees for diverse populations [8]; multi-task transfer scaling in the low-rank dimension [2].

Singapore Management University *Advisors: Pradeep Varakantham, Arunesh Sinha*

Research Assistant *May 2021 - Sep. 2022*

- Scalable algorithms with guarantees for NP-hard strategic optimization with boundedly rational agents, applied to security games and public health [12], [13] (AAAI'23 Oral), Stackelberg network interdiction under bounded rationality [14], sustainable fleet optimization [15].

Publications

Few-Shot Personalization from Memory

[1] **Avinandan Bose**, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, Maryam Fazel. "LoRe: Personalizing LLMs via Low-Rank Reward Modeling." **COLM 2025**. [paper] [code]

[2] **Avinandan Bose**, Simon Shaolei Du, Maryam Fazel. "Offline Multi-task Transfer RL with Representational Penalization." **AISTATS**

2025. [paper]

[3] Brandon Amos, Ratip Emin Berker, Himaghna Bhattacharjee, **Avinandan Bose**, Edith Elkind, Sonja Kraiczy, Smitha Milli, Max Nickel, Ariel Procaccia, Jamelle Watson-Daniels. "Inference-Time Social Choice for Democratic Representation of Viewpoints in Large Language Models." **In Preparation** *(alphabetical order)*

Inference-Time Personalization via Multi-Turn Interactions

[4] Shuyue Stella Li*, **Avinandan Bose***, Faeze Brahman, Simon Shaolei Du, Pang Wei Koh, Maryam Fazel, Yulia Tsvetkov. "Personalized Reasoning: Just-In-Time Personalization and Why LLMs Fail At It." **ICLR 2026**. [paper] [data] [blog]

[5] **Avinandan Bose**, Shuyue Stella Li, Faeze Brahman, Pang Wei Koh, Simon Shaolei Du, Yulia Tsvetkov, Maryam Fazel, Lin Xiao, Asli Celikyilmaz. "Cold-Start Personalization via Training-Free Priors from Structured World Models." *Under review at* **ICML 2026**. [paper]

[6] **Avinandan Bose**, Soumendu Sundar Mukherjee. "Changepoint Analysis of Topic Proportions in Temporal Text Data." *arXiv 2021*. [paper]

[7] **Avinandan Bose**, Zhihan Xiong, Aadirupa Saha, Simon Shaolei Du, Maryam Fazel. "Hybrid Preference Optimization for Alignment: Provably Faster Convergence Rates by Combining Offline Preferences with Online Exploration." *arXiv*. [paper]

[8] **Avinandan Bose**, Mihaela Curmei, Daniel L. Jiang, Jamie Morgenstern, Sarah Dean, Lillian J. Ratliff, Maryam Fazel. "Initializing Services in Interactive ML Systems for Diverse Users." **NeurIPS 2024**. [paper]

AI Safety & Robustness

[9] Léo Boisvert*, Mihir Bansal*, Chandra Kiran Reddy Evuru*, Gabriel Huang*, Abhay Puri*, **Avinandan Bose***, Maryam Fazel, Quentin Cappart, Jason Stanley, Alexandre Lacoste, Alexandre Drouin, Krishnamurthy Dvijotham. "DoomArena: A Framework for Testing AI Agents Against Evolving Security Threats." **COLM 2025**. [paper] [code] [webpage] [blog]

[10] **Avinandan Bose**, Laurent Lessard, Maryam Fazel, Krishnamurthy Dj Dvijotham. "Keeping Up with Dynamic Attackers: Certifying Robustness to Adaptive Online Data Poisoning." **AISTATS 2025**. [paper] [code] [blog]

[11] Léo Boisvert*, Abhay Puri*, Chandra Kiran Reddy Evuru*, Joshua Kazdan, **Avinandan Bose**, Quentin Cappart, Maryam Fazel, Sai Rajeswar, Jason Stanley, Nicolas Chapados, Alexandre Drouin, Krishnamurthy Dj Dvijotham. "Silent Sabotage: Injecting Backdoors into AI Agents Through Fine-Tuning." **ICML 2025 Workshop**. [paper]

Mechanism Design for Multi-Agent Systems

[12] **Avinandan Bose**, Tracey Li, Arunesh Sinha, Tien Mai. "A Fair Incentive Scheme for Community Health Workers." **AAAI 2023**, *Oral Presentation*. [paper]

[13] **Avinandan Bose**, Arunesh Sinha, Tien Mai. "Scalable Distributional Robustness in a Class of Non-Convex Optimization with Guarantees." **NeurIPS 2022**. [paper]

[14] Tien Mai, **Avinandan Bose**, Arunesh Sinha, Thanh Hong Nguyen, Ayushman Kumar Singh. "Tackling Stackelberg Network Interdiction against a Boundedly Rational Adversary." **IJCAI 2024**. [paper]

[15] **Avinandan Bose**, Hao Jiang, Pradeep Varakantham, Zichang Ge. "On Sustainable Ride Pooling Through Conditional Expected Value Decomposition." **ECAI 2023**. [paper]

Honors & Awards

2018	**Gold Medal**, Indian National Physics Olympiad (Top 35 nationally)	*India*
2015	**Gold Medal**, Indian National Junior Science Olympiad (Top 35 nationally)	*India*
2018	**All India Rank 104**, JEE Advanced (200,000 candidates)	*India*
2017	**All India Rank 68**, KVPY Scholarship, Indian Institute of Science	*India*
2018	**All India Rank 1**, West Bengal Joint Entrance Examination	*India*
2016-17	**National Top 1%**, NSE Physics, Chemistry, Astronomy (multiple years)	*India*

Service & Skills

Peer Review: ICLR 2025, AAAI 2025, NeurIPS 20-22,23,24,25 ICML 2024, L4DC 2024

Grad Application Review: Reviewed 30+ PhD applications for University of Washington CSE (2023, 2024)

Seminar Organizer: ML-OPT / IFDS seminar at UW (2022–24)

Skills: Python, C++, MATLAB, Shell; PyTorch