Research Overview

Avinandan Bose

PhD Candidate, University of Washington (2022–)

Visiting Research Scientist, Meta Superintelligence Labs FAIR (2024–)

Website · CV · Google Scholar · Semantic Scholar · GitHub · X

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: personalized adaptation from limited feedback [1],[2],[3], inference-time multi-turn preference learning [4],[5],[6],[7],[8], robustness of agents and RLHF pipelines [9],[10],[11], and multi-agent coordination under misaligned objectives [12],[13],[14],[15]. Understanding diverse users is both a **usability and safety problem**: the same machinery that adapts to individual users without requiring them to resteer, reprompt, or drop out also (1) enables **robustness against manipulation**, (2) **surfaces diverse viewpoints** on safety-critical questions, and (3) **prevents harm from one-size-fits-all defaults** and generically helpful agents.

Few-Shot Personalization from Memory

Human preferences are **high-dimensional, diverse, and often conflicting**: the same response can be helpful for one user and actively harmful for another. Standard approaches that condition on demographics discard the variation that matters and risk encoding profiling biases; learning from scratch per-user hits a **data wall** of limited prior interactions. I developed **LoRe** [1] (paper | code | COLM 2025), built on the finding that human preferences are **intrinsically low-dimensional**: LoRe learns a **compact basis of reward functions** from **binary comparisons** and recovers any new user's reward model from **5–10 comparisons with no retraining**, enabling **personalized steering at inference time**. Evaluated on the **largest publicly available preference datasets**: PRISM (**1,500 participants, 75 countries**) and Community Alignment (**3,000+ annotators, 200K comparisons, 5 languages**), LoRe outperforms prior approaches by **12% in preference prediction accuracy**. A complementary theoretical result [2] (paper | AISTATS 2025) proves that **near-optimal adaptation** requires **sample complexity scaling only in the low-rank representation dimension, not raw state space**, providing meta-learning guarantees that explain why LoRe succeeds with few-shot examples. In ongoing work [3], I am adapting this machinery to guarantee **democratic representation of diverse viewpoints** on socially contentious questions, grounded in **social choice theory**. By recovering the **full spectrum of preferences** from a population rather than collapsing to a majority view, this directly addresses **sycophancy and echo-chamber risks**: the system can surface legitimate disagreement rather than reinforcing whatever the current user wants to hear.

Inference-Time Personalization via Multi-Turn Interactions

While LoRe addresses memory-based personalization, **new users have no historical data**, and even users with history routinely encounter tasks where relevant context never appeared in prior interactions. I formalized **interactive preference discovery** as a research problem: can models *elicit* what individual users need through **multi-turn conversation**? I built **PrefDisco** [4] (paper | data | blog | ICLR 2026) to test whether this capability emerges in frontier models, evaluating **21 models** spanning **GPT, Claude, and Gemini** across **10 benchmarks** in mathematical, logical, scientific, and social reasoning (**10,000 user-task scenarios**). We find that **it does not**: frontier models **fail to ask the appropriate clarifying questions** even when explicitly prompted to do so, and **29% of elicitation attempts degrade alignment** versus generic responses. This is a hard **routing problem**: each task admits **20–30 preference dimensions**, but individual users care about only **2–4**, and which subset varies per user. With a budget of **~5 questions**, non-adaptive questioning over the full space will miss the dimensions that matter. RL is the natural formulation, learning a questioning policy with reward based on final response alignment, but the reward is **sparse and terminal**: the agent cannot determine which questions were informative and in practice **collapses to fixed question sequences** regardless of the user. I developed **PEP** [5] (paper | under review, ICML 2026), which decomposes the problem: learn **preference correlations offline** from existing data, then run **Bayesian inference online** so each answer updates beliefs about all others, with

questions selected via **information gain** to maximally reduce uncertainty about the user's complete profile. With **~10K parameters**, PEP outperforms **RL finetuning a 8B-parameter LLM**, achieving **3–5× fewer interactions** and **2× higher adaptivity** across **4 reasoning domains**: the bottleneck is **inference structure, not model capacity**. This decomposition is not specific to preference elicitation: it provides a general blueprint for **active information gathering in agentic systems**, any setting where agents must efficiently determine what they need to know about a user from limited interaction, including **tutoring, clinical decision support, coding assistants, and personalized content generation**.

On the theoretical side, I proved the **first matching upper and lower bounds** on sample complexity for **hybrid RLHF**, showing **provably faster convergence** than pure offline or online training [7] (paper | arXiv), and **near-optimal cold-start guarantees** for diverse populations under **bandit feedback** [8] (paper | NeurIPS 2024). These methods assume preferences are stationary within a session, but in practice user needs shift both within and across sessions; earlier work on **changepoint detection** [6] (paper | arXiv 2021) provides the foundation for recognizing when accumulated beliefs should be partially or fully reset.

AI Safety & Robustness

The methods above assume honest user engagement. In practice, AI assistants are increasingly deployed as **autonomous agents** across customer service, web browsing, and system administration, operating **without human oversight** at each step while facing **adversarial users, corrupted data sources, and hostile environments**, making robustness critical. I built **DoomArena** [9] (paper | code | webpage | blog | COLM 2025), a **plug-in security evaluation** integrating into **3 major agentic benchmarks** (BrowserGym, τ-bench, OSWorld), testing **3 frontier models** across **~500 tasks** and **multiple threat configurations**: **every agent has exploitable blind spots**, combined attacks reach up to **97% ASR**, and standard frontier safety guardrails are **largely ineffective**, highlighting that **scaling model capability does not address these vulnerabilities**. On the theoretical side, prior robustness certificates all assume **static adversaries**, but real attackers **observe and adapt** to the learner. I established the **first certified bounds against dynamic data poisoning** [10] (paper | code | blog | AISTATS 2025), grounded in **robust control theory**, with direct applications to **RLHF safety**. In **Silent Sabotage** [11] (paper | ICML 2025 Workshop), we showed that poisoning **just 5% of fine-tuning traces** embeds **stealthy backdoors** while **improving task performance**. Combining **scalable threat evaluation** with **certified defenses against adaptive adversaries** is the path to solving a critical deployment bottleneck: agents that can be **trusted in adversarial environments** without requiring human oversight at every step.

Mechanism Design for Multi-Agent Systems

As LLM-based agents are deployed in **shared environments**, from collaborative workflows to competitive marketplaces, reasoning about **equilibria, incentive compatibility, and fairness** across agents with **misaligned objectives** becomes essential. My earlier work developed the mathematical toolkit for this: **scalable algorithms with provable guarantees** for **NP-hard strategic optimization** involving **boundedly rational agents** across **security games, public health, facility location, and fleet optimization** [12] (paper | AAAI 2023 Oral), [13] (paper | NeurIPS 2022), [14] (paper | IJCAI 2024), [15] (paper | ECAI 2023). The core challenges, designing mechanisms for agents that **don't behave optimally**, that have **private information**, and that operate under **computational constraints**, transfer directly to **multi-agent LLM coordination, reward hacking mitigation, and incentive-aware RLHF**.

References

[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)

[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, 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]

[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]

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