sujai1.github.io
sh2583@cornell.edu
Google Scholar
LinkedIn

Sujai Hiremath

TL;DR

- Published 4 first-author papers (NeurIPS 24', UAI 25', NeurIPS 25', AISTATS 26') within 2 years of starting research. ORIE PhD at Cornell Tech.
- Was an Amazon Science Tübingen RS intern (2025), now interning @ Amazon Science San Francisco on the LLM Search, Security, and Observability Team until Aug 2026.
- Thinking about LLM post-training, reward modelling, and agentic security/AI safety.

Education

Cornell Tech | New York, NY 2023 - Now
PhD in Operations Research and Information Engineering | GPA: 3.9
- Areas: Causal Inference, Reinforcement Learning, LLMs

California Institute of Technology | Pasadena, CA 2019 - 2023
BS in Applied and Computational Mathematics | GPA: 4.0
- Areas: Machine Learning, Mathematical Modelling, Deep Learning

Work Experience

Research Scientist Intern | Amazon Research Bay Area, California. 01.2026 - 08.2026
- Manager: Dr. Shiva Kasiviswanathan.
- Working on 1) a regularization method for improving reward models and Q-functions, and 2) a RCA method for detecting the source of prompt injections.

Research Scientist Intern | Amazon Research Tübingen, Germany. 06.2025 - 11.2025
- Managers: Dr. Dominik Janzing, Dr. Elke Kirschbaum.
- Developed a method leveraging LLMs as unreliable experts to improve causal learning in finite samples. Published as first-author at AISTATS 2026.

PhD Student Researcher | Cornell Tech, NYC. 11.2023 - Present
- PI: Dr. Kyra Gan.
- Leveraged diffusion models, independence tests, and nonparametric regression for causal inference. Published 3 first-author papers at NeurIPS (2024, 2025), UAI 2025.

Publications and Preprints

1. **Hiremath, S.***, et al. From Detection to Attribution: Identifying Malicious Documents in Prompt Injection Attacks on LLM Agents via Root Cause Analysis. *preprint*, 2026.
2. **Hiremath, S.***, et al. Towards Efficient Representations for Reward Modeling and Reinforcement Learning via Rank Regularization *preprint*, 2026.
3. **Hiremath, S.***, et al. From Guess2Graph: When and How Can Unreliable Experts Safely Boost Causal Discovery in Finite Samples? *AISTATS*, 2026.
4. Meier, D.* and **Hiremath, S.***, et al. When Additive Noise Meets Unobserved Mediators: Bivariate Denoising Diffusion for Causal Discovery. *NeurIPS*, 2025.
5. **Hiremath, S.***, et al. LoSAM: Local Search in Additive Noise Models with Mixed Mechanisms and General Noise for Global Causal Discovery. *UAI*, 2025.
6. **Hiremath, S.***, et al. Hybrid Top-Down Global Causal Discovery with Local Search for Linear and Nonlinear Additive Noise Models. *NeurIPS*, 2024.

Service & Awards

Service: Reviewer for NeurIPS 25', ICLR 25', AISTATS 25', CLeaR 26', UAI 26'.
Awards: NeurIPS Top Reviewer 25' | Cornell Fellowship 23'.