Som Sagar

ssagar6@asu.edu
https://somsagar07.github.io

Education

Arizona State University — Tempe, Arizona
Ph.D. in Computer Science — Aug. 2023 - Present

- Advisor: Ransalu Senanayake
- GPA : 3.9/4.0
- Relevant Coursework : Natural Language Processing, Data Mining, Planning Learning Methods in AI, Statistical Machine Learning, Knowledge Representation, Reinforcement Learning

Indian Institute of Information Technology (IIIT) — Kottayam, Kerala
B.Tech (Honors) in Computer Science — Aug. 2019 - May 2023

- CGPA : 8.72/10.0
- Relevant Coursework : Machine Learning, Deep Learning, Python, Object Oriented Programming, Linear Algebra, Big Data, Data Structures and Algorithm, Data Warehousing and Data Mining, Applied Predictive Analysis, Probability and Statistics, Calculus I, II

Research Interest

Agentic AI, Deep Reinforcement Learning, Foundation Models, Failure Detection and Mitigation, Generative AI, Explainability

Preprints & Publications

*denotes equal contribution

1. **Som Sagar***, Aditya Taparia*, Harsh Mankodiya, Pranav Bidare, Yifan Zhou, Ransalu Senanayake. Trustworthy Explanations for Robot Behaviors. *International Conference on Intelligent Robots and Systems (IROS)*, 2025.
2. Aditya Taparia, **Som Sagar**, Ransalu Senanayake. Explainable Concept Generation through Vision-Language Preference Learning for Understanding Neural Networks' Internal Representations *International Conference on Machine Learning (ICML)*, 2024.
3. **Som Sagar**, Jiafei Duan, Sreevisakh Vasudevan, Heni Ben'Amor, Dieter Fox, Ransalu Senanayake. From Mystery to Mastery: Failure Diagnosis for Improving Manipulation Policies. RSS Workshop on Out-of-Distribution Generalization in Robotics, 2025.
4. **Som Sagar**, Aditya Taparia, Ransalu Senanayake. Failures Are Fated, But Can Be Faded: Characterizing and Mitigating Unwanted Behaviors in Large-Scale Vision and Language Models. *International Conference on Machine Learning (ICML)*, 2024. (**Spotlight**)
5. **Som Sagar***, Aditya Taparia*, Harsh Mankodiya, Pranav Bidare, Yifan Zhou, Ransalu Senanayake. Trustworthy Conceptual Explanations for Neural Networks in Robot Decision-Making. *NeurIPS Workshop on Safe and Trustworthy Agents*, 2024.
6. **Som Sagar**, Aditya Taparia, Ransalu Senanayake. LLM-Assisted Red Teaming of Diffusion Models through "Failures Are Fated, But Can Be Faded" *NeurIPS Workshop on Red Teaming GenAI: What Can We Learn from Adversaries?*, 2024.
7. Aditya Taparia, **Som Sagar**, Ransalu Senanayake. Explainable Concept Generation through Vision-Language Preference Learning. *NeurIPS Workshop on Interpretable AI: Past, Present and Future*, 2024.
8. Joshua Tint, **Som Sagar**, Aditya Taparia, Caleb Liu, Kelly Raines, Bimsara Pathiraja, Ransalu Senanayake. ExpressivityArena: Can LLMs Express Information Implicitly?. *NeurIPS Workshop on Behavioral Machine Learning*, 2024.
9. **Som Sagar**, Swani Sundara Didde, Cinu S Killilor. Embedding Based Analysis for Simplification of Legal Terms. *International Conference on Computing Science, Communication and Security*, 2023.

Experience

Machine Learning Research Intern May. 2025 - Present

LinkedIn Corporation, Agents Platform Team

- Building intelligent agent systems by combining reinforcement learning (RL), preference optimization, and supervised fine-tuning (SFT) to train domain-specialized agents for real-world tasks.
- Improving multi-agent communication by designing faster, more effective agent-to-agent interaction protocols and scalable training pipelines.
- Enhancing agent alignment through structured reward signals and human feedback, enabling generative agents to collaborate and adapt across diverse environments.

Research Assistant Aug. 2023 - Present

Laboratory for Learning Evaluation of autoNomous Systems (LENS Lab), ASU

- Conducting research at the intersection of reinforcement learning, foundation models, and robotics, with a focus on improving model adaptability and robustness in real-world applications.
- Working on developing a framework that enhance the interpretability and trustworthiness of AI systems in dynamic environments.
- Collaborating with interdisciplinary teams to address key challenges in explainability, preference learning, and failure detection in machine learning models.

Awards and Honors

- **Spotlight (Top 3%)**, International Conference on Machine Learning 2024
- **ASU Graduate & Professional Student Association Travel Award** , Arizona State Univeristy 2024
- **Interpretable AI NeurIPS Travel Grant**, NeurIPS 2024
- **Graduate College Travel Award**, Arizona State Univeristy 2024
- **SCAI Conference Award**, School of Computing and Augmented Intelligence 2024
- **Prime Minister Scholarship**, Government of India 2019-23
- **Inter IIIT Hackthon Winner**, Indian Institute of Information Technology 2022

Service

Reviewer: *Conference on Robot Learning (CoRL) 2025*,
International Conference on Intelligent Robots and Systems (IROS) 2024,
International Conference on Learning Representations (ICLR) 2025,
Conference on Neural Information Processing Systems (NeurIPS) 2024,
International Conference on Intelligent Robots and Systems (IROS) 2025

Teaching

- **Instructor**, FSE 100 : Introduction to Engineering, ASU Fall 2023, 2024, 2025
- **Teaching Assistant**, CSE 598 : Operational Deep Learning, ASU Spring 2024
- **Teaching Assistant**, CSE 100 : Introduction to C++, ASU Spring 2024

Skills

Programming Languages: Python, C, C++, Dart, JavaScript

Frameworks: PyTorch, NumPy, Pandas, Captum, Stable Baselines, Diffusers, NLTK, Gymnasium, Gradio, TensorFlow, Sckit-learn, Keras, TRL

Simulation and Environment Tools: MuJoCo, Issac Gym/Sim/Lab, CARLA, OpenAI Gym, RLBench

Databases and Cloud Services: MySQL, AWS, Firebase

Development Tools: Visual Studio Code, Spyder, Andriod Studio, Git, Docker.