Som Sagar

Email: ssagar6@asu.edu | +1 (623)-283-8669 | LinkedIn | Github | Scholar

EDUCATION

Arizona State University Tempe, Arizona
*PhD Candidate, Computer Science – **GPA:** 4.16/4* August 2023 – May 2028 (Expected)
Relevant coursework: Natural Language Processing, Data Structures & Algorithms, Machine Learning, Data Mining, Deep Learning, Artificial Intelligence, Object-Oriented Programming, Statistics , Big Data Processing.

RESEARCH PUBLICATION

- S. Sagar, A. Taparia, R. 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**. [paper] *(Spotlight)*
- S. Sagar, SS. Didde, CC. Kilillor, **"A Sentiment Word2Vec Approach for Simplification of Legal Terms,"** International Conference on Computing Science, Communication and Security **(COMS2), 2023**. [paper]

RESEARCH EXPERIENCE

LENS Lab, Arizona State University August 2023 – Present
Graduate Research Assistant

- Executed **Deep Q-Networks** (DQN) to identify **out-of-distribution** (OOD) instances in classifiers and language models (T5, BART), **fine-tuning** reduced OOD errors by 30%, boosted prediction accuracy by 13%.
- Applied **LoRA** techniques to decrease bias in Stable Diffusion v1-5 model by 37% and improve output quality.
- Engineered a DQN-based concept generation framework using **TCAV** (Testing with Concept Activation Vectors) rewards and a **preference optimizing** step function for concept-based explanations. (*ICLR25 Under Review*)
- Designed and deployed a **Bayesian TCAV** framework with **uncertainty estimations**, enhancing interpretability of robotic actions by 22% across diverse simulation platforms including MuJoCo, DonkeyGym, and JetBot. (*ICRA25 Under Review*)
- Developed ExpressiveArena to evaluate implicit communication across multiple **Large Language Models** (LLMs), achieving up to 72% accuracy in most tasks, with GPT-4 showing best performance.

RELEVANT PROJECTS

Preference Diffusion, *Stable Diffusion, Clustering* February 2024 – May 2024

- Implemented **Direct Preference Optimization** (DPO) on stable diffusion models to enhance clarity of generated images from fuzzy prompts, reducing output uncertainty by 15%.
- Developed an innovative feedback mechanism leveraging real-time minimal human feedback with **Gaussian Mixture Modeling** (GMM) on **CLIP embedding** of input to refine DPO, **aligning** generated images more closely with **user preferences**, resulting in a 20% improvement in user satisfaction scores.

Automated Stock Trading, *Reinforcement learning, Reward Modeling* January 2023 – April 2023

- Crafted a robust RL trading system utilizing DQN, **Proximal Policy Optimization** (PPO), and **Advantage Actor-Critic** (A2C) algorithms to refine trading strategies.
- Formulated a multiple reward function incorporating various technical indicators using **Multi-Layer Perceptron** (MLP) and **Long Short-Term Memory** (LSTM), achieving a 23% increase in decision-making accuracy.
- Implemented a **LLM reward function** with **on-the-fly human feedback** for continuous learning, showing a preliminary improvement of 21%.

TECHNICAL SKILLS

- **Programming Languages:** Python, C, C++, Dart, JavaScript, HTML/CSS
- **Technologies/Frameworks:** PyTorch, TensorFlow, Scikit-learn, Keras, NumPy, Pandas, Captum, Stable Baselines, Diffusers, Transformers, NLTK, Gymnasium, Gradio, Flutter
- **Simulation and Environment Tools:** MuJoCo, CARLA, OpenAI Gym, RLBench
- **Databases and Cloud Services:** MySQL, Firebase, AWS
- **Development Tools:** Visual Studio Code, Spyder, Jupyter Notebook, Android Studio, Git, Docker

AWARDS AND ACHIEVEMENTS

- Secured Spotlight Paper (Top 3.5% of submissions) at ICML 2024. July 2024
- Awarded Graduate College Travel Award, Arizona State University. June 2024
- Received SCAI Conference Award, School of Computing and Augmented Intelligence. May 2024