Sundar Sripada V. S.

sundarsripada26@gmail.com github.com/ss26 ss26.github.io linkedin.com/in/sundarsripada26

Education

University of Texas at Austin — Austin, TX, USA
Master of Science in Engineering in Electrical and Computer Engineering — Aug. 2022 – Dec. 2024
Graduate Teaching Assistant at Department of Physics — Jan. 2024 – Dec. 2024

Experience

ML Research Engineer — Jan. 2025 – Present
AI Safety Camp — Remote

- Conducting empirical research on adversarial behaviors in LLMs trained with PPO and DRRN policies, aiming to classify and counteract adversarial reward optimization strategies
- Developing an actor-critic Python PPO agent with Causal LM tokenizers for the Machiavelli dataset (merged PR)
- Evaluating agents' propensity to optimize for adversarial rewards on the Machiavelli and AgentHarm benchmarks by running PPO and PPO-LLM experiments with Center for AI Safety's compute clusters

Software Engineer Intern — Jun. 2023 – Aug. 2023
Hewlett Packard Enterprise — San Jose, CA, USA

- Enhanced scale characterization of HPE Aruba's network management tool by developing Virtual Machine (VM) simulation support in Python, doubling the tool's device handling capacity by up to 2,000,000 simulated devices
- Developed CLI-based VM simulation functionality that enables users to simulate VMs within Docker containers deployed with Kubernetes, achieving fast on-boarding (¡1 minute per VM) to enhance scalability
- Collaborated in a cross-functional team writing SQL code for extracting VM system information, integrating it into the simulation framework to enable scale simulation and support data-driven decisions for cloud infrastructure

Graduate Research Assistant — Aug. 2022 – Jun. 2023
Swarm Lab, University of Texas at Austin — Austin, TX, USA

- Built TensorFlow-based activity recognition model with 87% classification accuracy across 4 tool-use behaviors, enhancing ergonomic study of user-tool interaction. Paper
- Led a team of 3 students in a data collection project using a 9-DOF IMU on a robot arm that emulates human performance.
- Engineered a data processing pipeline with Pandas to analyze a dataset of over 10,000 entries for activity classification.
- Implemented an ML pipeline in TensorFlow Lite (TFLite) and MLFlow for logging, utilizing convolutional neural networks (CNN) for real-time edge inference on a Raspberry Pi, achieving 162 ms per prediction for activity classification.

Research Engineer — Oct. 2020 – Jul. 2022
Robotics Research Center, International Institute of Information Technology — Hyderabad, India

- Managed a large-scale data mining and collection pipeline using Python scripts with the CARLA simulator's API (autonomous systems) to gather contrived scenes of environments containing over 100,000 LIDAR data points for SLAM
- Ported LOAM's C++11 source code to C++14, enabling usage in the new ROS version (GitHub).
- Designed classification and regression models in scikit-learn to predict the presence and magnitude of drift in self-driving vehicles, achieving prediction accuracy of up to 92% with PyTorch
- Devised a robust data processing pipeline by developing API functions and automation scripts in Python to transform raw LIDAR data into custom datasets for PyTorch's DataLoader, significantly enhancing the efficiency of model training
- Formulated a reinforcement learning model using Proximal Policy Optimization (PPO) to reduce drift in highly dynamic autonomous driving scenes. Paper, GitHub
- Built a custom learning-to-rank triplet loss function to train a multi-modal CNN in PyTorch, effectively minimizing drift in autonomous driving and improving performance by up to 76.76% over previous SOTA. Paper, GitHub

Skills

Programming Languages	Python, C++, C, Bash, R, MySQL, MATLAB, Java, Julia, HTML, CSS, JavaScript, LaTeX
Frameworks	PyTorch, TensorFlow, TensorFlow Lite, InspectAI (AISI), LangChain, Gymnasium, gym, ROS, MLFlow, Docker, Apache Spark
Libraries	numpy, pandas, polars, matplotlib, seaborn, OpenCV, HuggingFace Transformers, HuggingFace TRL, CleanRL, RLlib, Pillow, scikit-learn, MLlib
Miscellaneous	Linux, Windows 10/11, git, GitHub, AI Safety Fundamentals Certificate