Shreyas Bachiraju

+1(602)880-1583 • shreyasbachiraju@gmail.com • linkedin.com/in/shreyasbachiraju • Google Scholar • GitHub

Education

Bachelor of Science in Informatics May 2026

Arizona State University | **GPA**: 3.95

Awards:

- 1st Place Winner HackHarvard 2024 (Sustainability Track), 1st Place Overall Winner of Ethical Spectacle Research Sustainability Hackathon,
- New American University Scholar (merit-based, $14.5k/yr), Dean's List (all semesters)

Leadership:

- Founding Co-President of CS + Social Good at ASU
- Undergraduate Teaching Assistant for Advanced Object-Oriented Programming in Java

Publications

Turnau, J., Da, L., Vo, K., Al Rafi, F., **Bachiraju, S.**, Chen, T., & Wei, H. (2025). *Joint-Local Grounded Action Transformation for Sim-to-Real Transfer in Multi-Agent Traffic Control*. Reinforcement Learning Journal, 6, 2271–2290. (Presented at RLC 2025)

Da, L., Chen, T., Li, Z., **Bachiraju, S.**, Yao, H., Li, L., Dong, Y., … Wei, H. (2025). *Generative AI in Transportation Planning: A Survey*. arXiv:2503.07158. (Under journal revision)

Research Experience

Research Intern September 2024 – Present

Data Mining and Reinforcement Learning Lab at Arizona State University

Dataset Distillation for Reinforcement Learning Environments for Resource-Constrained Setups

- Developed a Matching Training Trajectories (MTT) dataset-distillation method that compressed a 100K-step PPO training run into 50 optimized synthetic states capturing the expert's learning trajectory.
- Implemented a bi-level, gradient-through-gradient meta-learning pipeline (via the higher library) to optimize synthetic states through KL-based policy matching and parameter-trajectory alignment.
- Achieved 100x faster training, enabling a new agent trained for 1K supervised steps on the distilled dataset to retain 85%+ of expert performance. Currently extending the method to evaluate on more environments.

Generative AI in Transportation Planning: A Survey | *Preprint*

- Led a case study on optimizing Origin-Destination (OD) Matrices using LLMs, improving traffic flow prediction accuracy and convergence stability.
- Designed and implemented VanillaLLM and ExpertLLM pipelines, integrating domain-specific heuristics through contextualized Chain-of-Thought prompting with LLaMa and HuggingFace Transformers.
- Collaborated with civil engineers to develop context-aware heuristics that guided localized OD matrix updates.
- Contributed to research figures and experimental validation. Preprint now under revision for journal submission.

Compression of Deep Neural Networks for Edge Devices | *Poster Presentation: Fulton Forge Research Expo (2025)*

- Led a project through the Fulton Undergraduate Research Initiative on compressing and deploying depth-estimation models for real-time inference on resource-constrained Jetson Nano hardware.
- Built an end-to-end deployment pipeline using PyTorch, ONNX, and TensorRT, implementing mixed-precision quantization to improve inference efficiency.
- Diagnosed and resolved ONNX–TensorRT compatibility issues by redesigning the model's encoder-decoder architecture with supported convolutional operations while maintaining model accuracy.
- Achieved 33% lower latency, 42% smaller model size, and 50% higher throughput, retaining 99% of baseline accuracy during on-device inference.

JLGAT for Sim-to-Real Transfer in Multi-Agent Traffic Control | *Reinforcement Learning Conference 2025*

- Executed controlled experiments comparing centralized, decentralized, and JL-GAT policies under varying weather conditions and across 1x3 and 4x4 traffic networks.

Note: Text in blue is hyperlinked.

Research Intern June 2025 – August 2025

ASU Center for Semiconductor Microelectronics (ACME)

Energy-Efficient Inference in Mixture-of-Experts (MoE) Large Language Models

- Profiled Mixtral-8x7B on GPU–CPU systems using NVML and RAPL to quantify power, latency, and transfer overheads.
- Identified that expert-weight swapping between GPU and CPU accounted for ~60% of total energy usage and 62% of inference time, revealing a critical system-level bottleneck.
- Developed a cost lookup table (LUT) capturing expert-specific throughput, latency, residency, and energy costs to enable informed routing and scheduling decisions.
- Proposed an energy- and residency-aware routing policy that predicts subsequent expert activations and conditionally prefetches non-resident experts into GPU memory.

Experience

AI/ML Intern May 2024 – July 2024

Netradyne

- Benchmarked state-of-the-art (SOTA) depth estimation models for Time to Collision (TTC) estimation with forward vehicles for Advanced Driver Assistance Systems (ADAS) using PyTorch.
- Developed an end-to-end automated pipeline using Python and AWS S3 to extract and process 30,000+ dashcam videos and execute model inference, thereby reducing total processing time by 90%.
- Improved the model's per-sample inference time by 25% through batch inference and implemented TTC estimation pipelines.
- Reduced Absolute Relative Error by 15% through hyperparameter tuning and lab experiments; generated depth maps for 280,000+ images to supervise an adapted YOLOv5 model for depth prediction.

Software Engineer Intern June 2023 – July 2023

Clocr Inc.

- Integrated OpenAI's GPT API into My-Legacy.ai, an estate planning chatbot, to enhance conversational accuracy and domain-specific reliability.
- Decreased LLM hallucinations by 20% through fine-tuning and retrieval-augmented generation (RAG) pipelines built with custom legal knowledge bases.
- Led data-driven A/B testing iterations and user feedback analysis to enhance chatbot's personalized recommendations; projected a 40% increase in user satisfaction based on user experience trials.

Projects

U-Plan - HackHarvard 2024 Winner in Sustainability Track | Demo

Python, Segment Anything Model, LLMs, Pandas, Mapbox, Rasterio, Geopy, Folium

- Led a team of four to develop a city-scale sustainability platform analyzing urban heat, vegetation, and water coverage across 50+ Phoenix ZIP codes using SAM segmentation, GIS processing, and remote-sensing analytics.
- Built end-to-end Python pipelines and interactive 3D GIS visualizations to compute environmental indices (NDVI, NDWI, LST) from satellite data, generating localized heat-mitigation strategies and infrastructure recommendations using an LLM.

TransformersNotFound - Creating an Open-Source GPT

Python, PyTorch, HuggingFace, Pandas, Accelerate, vLLM, FastAPI

- Fine-tuned LLaMA-3.2B model using parameter-efficient LoRA adapters for reasoning tasks on a custom Chain-of-Thought dataset. Implemented multi-GPU distributed training and optimized model inference using HuggingFace Accelerate & vLLM.

Feal? Fake vs Real Image Binary Classifier

Python, PyTorch, Scikit-Learn, Matplotlib, Pandas

- Achieved 97.3% accuracy in binary classification of real vs. AI-generated images by fine-tuning a ResNet-18 model with PyTorch on an augmented dataset enriched with 10,000+ synthetic images generated by a custom Deep Convolutional GAN.

Skills

Languages & Frameworks/Tools*:* Python, PyTorch, Java, C, C++, R, JavaScript, LLMs, HuggingFace, Transformers, SQL, MATLAB, Scikit-Learn, AWS, Git, Docker, ONNX, OpenCV, Stable-Baselines3, Accelerate, vLLM, REST, TensorRT