Shreyas Bachiraju

+1(602)880-1583 | shreyasbachiraju@gmail.com | Portfolio | LinkedIn | Google Scholar | GitHub

Education

Carnegie Mellon University | *M.S. in Machine Learning* Aug. 2026 – Dec. 2027

Arizona State University | *B.S. in Informatics, GPA: 3.98* May 2026

- **Honors:** Fulton Engineering Outstanding Graduate; 2026 CRA Outstanding Undergraduate Researcher Award – Honorable Mention; 1st Place – HackHarvard 2024 (Sustainability); New American University Scholar ($14.5k/yr); Dean's List

Experience

AI/ML Intern May 2024 – July 2024

Netradyne

- Built an end-to-end Python and Amazon S3 pipeline to extract frames from 30,000+ dashcam videos and run depth-model inference in parallel, reducing total processing time by 90%.
- Improved per-sample inference time by 25% via batch inference while benchmarking state-of-the-art monocular depth-estimation models in PyTorch for time-to-collision estimation in ADAS.
- Reduced Absolute Relative Error by 15% through hyperparameter tuning; generated depth maps for 280,000+ images as synthetic training data for a YOLOv5-based model.

Software Engineer Intern June 2023 – July 2023

Clocr Inc.

- Integrated OpenAI GPT API into an estate-planning chatbot and reduced hallucinated responses by 20% via RAG pipelines over custom legal knowledge bases and domain-specific prompt tuning.
- Improved chatbot helpfulness ratings by 40% in UX trials via A/B tests on prompt templates, retrieval relevance, and personalized recommendation logic.

Research Experience

Research Intern Sept. 2024 – Dec. 2025

Data Mining & Reinforcement Learning Lab – Prof. Hua Wei, ASU

- Built a PyTorch–ONNX–TensorRT mixed-precision pipeline for real-time depth estimation on Jetson Nano, reducing latency by 33%, shrinking model size by 42%, and increasing throughput by 50% while retaining 99% baseline accuracy. **Poster Link**.
- Designed a dataset-distillation method for RL using a bi-level meta-learning pipeline in PyTorch, compressing a 100K-step PPO trajectory into 50 synthetic states and retaining 96% of expert policy performance.

Research Intern June 2025 – Aug. 2025

ASU Center for Semiconductor Microelectronics – Prof. Krishnendu Chakrabarty

- Profiled Mixtral-8×7B MoE inference on GPU–CPU systems using NVML, RAPL, and CUDA; built a thread-safe expert-swapping pipeline with explicit GPU memory management, pinned-host transfers, and CUDA stream scheduling.
- Identified memory swapping as the primary bottleneck, contributing ∼60% of energy and latency overhead; proposed an energy- and residency-aware routing policy to prefetch non-resident experts into GPU memory.

Publications

Da, L., Chen, T., Li, Z., **Bachiraju, S.**, Yao, H., Li, L., et al. "Generative AI in Transportation Planning: A Survey." *arXiv:2503.07158*, 2025. Preprint. **Link**.

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

Projects

Circle.ooo | *Capstone Project, AI Event and Sponsorship Marketplace*

- Developed a LangGraph-based multi-agent event discovery engine that converts natural-language queries into ranked, registration-ready events aggregated from web search, Eventbrite, and Luma.
- Built a sponsor–event matching system with an LLM-based voice assistant for discovery, valuation, and outreach.

TransformersNotFound | *PyTorch, HuggingFace, Accelerate, vLLM*

- Fine-tuned LLaMA-3.2 3B with parameter-efficient LoRA adapters on a custom Chain-of-Thought dataset; implemented multi-GPU distributed training and optimized inference with HuggingFace Accelerate and vLLM.

U-Plan – HackHarvard 2024 Winner, Sustainability | *SAM, LLMs, GIS, Mapbox, Rasterio*

- Led a team of four to build a city-scale sustainability platform over 50+ Phoenix ZIP codes using SAM segmentation and remote-sensing analytics including NDVI, NDWI, and LST, generating LLM-driven heat-mitigation strategies.

Skills & Coursework

Languages: Python, Java, C/C++, SQL, JavaScript, R, MATLAB

ML / Tools: PyTorch, HuggingFace, Transformers, scikit-learn, vLLM, Accelerate, TensorRT, ONNX, Stable-Baselines3, OpenCV, NumPy/SciPy, AWS, Docker, REST

Relevant Coursework: Machine Learning, Deep Learning, Natural Language Processing, Artificial Intelligence, Data Structures and Algorithms, Advanced Object-Oriented Programming, Software Engineering, Statistics, Linear Algebra