Abhi Vetukuri

ahv37@cornell.edu | 508-818-6778 | LinkedIn | Github

EDUCATION

CORNELL UNIVERSITY - GPA: 4.0/4.0 Ithaca, NY
BSc/M.Eng - Computer Science: Artificial Intelligence Minor *May 2026*

Course Work: *Data Structures & Algorithms, Operating Systems, Computer Networks, Deep/Machine Learning, Statistics, Reinforcement Learning: Robotics, Linear Algebra, Optimization, Multi–Var Calculus, NLP, Physics*

EXPERIENCE

PALANTIR TECHNOLOGIES Palo Alto, CA
Software Engineer Intern *May 2024 - Aug 2024*

- Built high-perf front-end features for Palantir GeoViewer using Typescript, React/Redux, Java, and Go
- Utilized typescript to develop several new front-end features and optimizations to the platform
- Added performance optimization features like clustering, resulting in 30% rendering efficiency and optimized file system/organization reducing data retrieval times by 50% for large workflows
- Developed an LLM-powered tool in Java that dynamically generates and executes cURL requests

CORNELL CENTER OF ADVANCED COMPUTING Ithaca, NY
Cloud Engineer Intern *Aug 2023 – Jul 2024*

- Architected an AWS ParallelCluster-based HPC solution for WRF climate simulations for the NSF
- Utilized Terraform for infra management and Singularity containers with MPI for parallel processing
- Implemented automated data pipeline using S3 for storage and Apache Airflow for workflow orchestration, reducing data processing time by 35% and cutting overall computation costs by 25%

FIDUCIA AI San Ramon, CA
Software Development Intern *May 2023 - Aug 2023*

- Implemented a sentiment analysis application using MongoDB, Express, Angular, and Node, increasing engagement by 40% for 1000+ users by providing real-time NLP insights
- Leveraged BERT and OpenAI API for NLP analysis reducing costs by 60% and improving accuracy by 10% over previous solutions, and then containerized app on AWS EKS, reducing infra costs by 20%

CORNELL AUTONOMOUS SAILBOAT TEAM Ithaca, NY
Software Developer *Oct 2022 - Dec 2023*

- Engineered autonomous navigation using Deep Deterministic Policy Gradient algo, deploying on Nvidia Jetson with Python, PyTorch, and OpenAI Gym for waypoint navigation and object avoidance
- Developed Reinforcement Learning training simulator with ROS and Gazebo, ran over 1000 scenarios
- Built real-time control systems using Arduino for hardware interfacing and Linux for algo execution

PROJECTS

AI INFERENCE | *Performance Optimization, PyTorch, CUDA, Memory Management* | 🔗 Feb 2025 - Mar 2025

- Engineered high-performance PyTorch inference engine achieving 25+ tokens/sec on 1B param models
- Implemented 4/8-bit quantization reducing memory footprint 2GB to 0.5GB per billion parameters and developed optimized CUDA kernels and tensor operations for efficient large model inference
- Built dynamic batching system with Flash Attention 2 improving throughput by 15% over baseline

DISTLM | *Distributed Systems, Python, PyTorch, Ray, FastAPI* | 🔗 *Feb 2024 - Jul 2024*

- Built distributed LLM training platform with Ray for tolerant scaling across heterogeneous compute nodes
- Implemented token-based compute sharing system with JWT authentication and role-based access control, and developed real-time monitoring and metrics collection for node health, resource usage, etc

SKILLS

Languages: Python, Java, C++, SQL, JavaScript, TypeScript, HTML/CSS, Go, Haskell
Libraries: TensorFlow, PyTorch, Ray, Pandas, Numpy, Scikit-Learn, React, Node.js, Django, Spark, Kafka
Technologies: AWS, Docker, Kubernetes, Terraform, Pulumi, Singularity, MongoDB, Git, Linux, AirFlow