Savan Reddy Poduturi

+1 469-751-7874 | savanreddypoduturi@gmail.com | LinkedIn | Github

Education

- **University of Texas at Arlington**, M.S. in Computer Science (Thesis) May 2025
  Specialized in AI/ML Systems, Distributed Computing • GPA: 4.0 • Outstanding Master's Student Arlington, TX

Technical Skills

- **Languages**: Python, C++, Java, Go, TypeScript, JavaScript, Swift, Kotlin, C#, SQL
- **AI / ML Systems**: PyTorch, TensorFlow, Reinforcement Learning (RL), Model Quantization, ONNX, Pandas
- **Backend & Microservices**: Node.js, Spring Boot, Kafka, gRPC, WebSockets, RESTful APIs, GraphQL
- **Databases & Caching**: PostgreSQL, Redis, ScyllaDB, MongoDB, MySQL, DynamoDB, Firebase
- **Frontend & Mobile**: React, Svelte, Next.js, HTML5/CSS3, Tailwind, Redux, Android Studio, SwiftUI
- **Cloud & DevOps**: AWS (Lambda, S3, EC2), Docker, Kubernetes, CI/CD, Git, GitHub Actions, Terraform

Experience

- **University of Texas at Arlington** Arlington, TX
  Research Software Engineer (ML Systems) April 2024 – Present
  - Engineered a Deep Reinforcement Learning (DQN) algorithm in Python, utilizing HPC GPU clusters to accelerate training by 40% and enabling low-latency inference suitable for edge applications.
  - Optimized neural compute graphs using C++ and CUDA to address resource constraints, reducing memory overhead by 54% while maintaining model accuracy for on-device deployment.
  - Integrated quantized models into Android-based VR headsets using Java (Android Studio) and ONNX Runtime, meeting strict motion-to-photon latency requirements for real-time interactive applications.
  - Developed a full-stack analytics dashboard using React and Node.js to visualize network throughput and VR cybersickness metrics, accelerating data-driven decision-making by 30%.
  - Implemented a Redis caching layer for the analytics API to reduce redundant database queries, lowering average response latency by 200ms during high-load experiment runs.
  - Architected a fault-tolerant ETL pipeline in Python handling 200K+ telemetry traces, implementing automated feature validation and cleaning to ensure high data integrity for rigorous model testing.

Projects

- **CONVERSE: High-Throughput Distributed System** — *Go, Kubernetes, Kafka, ScyllaDB, Redis*
  - Architected a scalable microservices platform on Kubernetes (K8s), handling 75K+ concurrent WebSocket connections with non-blocking I/O patterns similar to embedded resource management.
  - Optimized data ingestion via Kafka and ScyllaDB to achieve <80ms p99 latency for real-time messaging, implementing E2E encryption and Chaos Mesh testing to ensure zero message loss.
- **Edge-Ready Neural Recommendation Engine** — *PyTorch, AWS, CUDA, Docker*
  - Engineered a Neural Collaborative Filtering (NCF) network in PyTorch, applying Post-Training Quantization (INT8) to reduce model footprint by 4x for constrained edge environments.
  - Accelerated inference throughput for 5M+ interactions using Cloud GPU parallelism (CUDA) and deployed a scalable training pipeline on AWS Spot Instances to benchmark accuracy vs. cost trade-offs.
- **CSocial: Full-Stack Recommendation Platform** — *Python, Flask, React, PostgreSQL*
  - Built a recommendation system improving user engagement by 130% via implicit NCF modeling, served via a Flask API with a responsive React frontend.
  - Developed a background ETL pipeline ingesting 50K+ events/hr into PostgreSQL, utilizing Pandas for efficient data transformation and storage.

Achievements and Publications

- **Outstanding Master's Student**: Research Excellence, Class of 2025.
- **CodeVita Coding Contest**: Ranked top 0.3% of 136,000 participants.
- **Master's Thesis**: Vioken: DQN-powered Adaptive Bitrate Control for VR.

Profiles

LeetCode: savanpoduturi | **Codeforces**: noobiest-coder | **CodeChef**: noobiest_coder