

Savan Reddy Poduturi

+1 469-751-7874 | savanreddypoduturi@gmail.com | Arlington, TX | Portfolio | LinkedIn | GitHub

Software Engineer — Systems & ML · 3+ years building production data, ML, and full-stack systems

TECHNICAL SKILLS

Languages: Python, C++, Java, Go, TypeScript, SQL
Backend: FastAPI, Node.js, REST APIs, Kafka, Redis

ML & Data: PySpark, PyTorch, CUDA, Scikit-learn, RAG
Cloud: AWS, Databricks, Docker, CI/CD, Kubernetes

EXPERIENCE

Software Engineer
UnitedHealth Group

Aug 2025 – Present
Remote, United States

- Designed Python and PySpark ML pipelines processing 1.5M+ monthly claims, eligibility, and provider records to detect anomalies, care gaps, and utilization shifts.
- Launched a React, TypeScript, and FastAPI healthcare assistant using LLM APIs and RAG, reducing case-summary and policy-review time by 30%.
- Developed LangGraph agentic workflows coordinating retrieval, policy validation, case analysis, tool execution, and human-in-the-loop escalation for complex cases.

Research Software Engineer (ML Systems)
University of Texas at Arlington

Apr 2024 – May 2025
Arlington, TX

- 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.

Software Engineer
Capital One

Jun 2022 – Jul 2023
Remote, India

- Automated Python and PySpark pipelines processing 5M+ merchant and card-transaction records into validated Amazon S3 datasets for downstream model training and analytics.
- Trained and evaluated XGBoost and Scikit-learn models to classify merchants, identify recurring payments and subscriptions, and detect billing-pattern changes across customer accounts.
- Delivered TypeScript and Node.js microservices handling 100K+ daily requests for transaction classifications and ML insights across card-servicing and digital banking applications.

PROJECTS

TORA — AI-Agentic Collaborative Workspace

Go · SvelteKit · ScyllaDB · Redis · LLM APIs

- Engineered shared code, drawing, task, and chat surfaces with a Go monolith and responsive SvelteKit frontend.
- Used prompt caching and project context management to reduce third-party LLM API costs by 35%.

Edge-Ready Neural Engine — ML Systems Optimization

PyTorch · AWS · CUDA · Docker

- Reduced model footprint by 4x for constrained edge environments.

EDUCATION

University of Texas at Arlington — M.S. in Computer Science (Thesis)

May 2025

GPA: 4.0 · Outstanding Master's Student (Class of 2025) · Thesis: Vioken: DQN-powered Adaptive Bitrate Control for VR

ACHIEVEMENTS

CodeVita Coding Contest: ranked in the top 0.3% of 136,000 participants.