

Savan Reddy Poduturi

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

EDUCATION

- **University of Texas at Arlington**, M.S. in Computer Science (Thesis) May 2025
 - 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
 - Software Engineer (Systems & ML) April 2024 – Present
 - Engineered a Deep Reinforcement Learning (DQN) algorithm in Python, using HPC GPU clusters to accelerate training by 40% and enable low-latency inference suitable for edge applications.
 - Optimized neural compute graphs via 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 VR headsets using Java (Android Studio) and ONNX Runtime to meet strict motion-to-photon latency requirements for real-time interactive apps.
 - 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 Python ETL pipeline handling 200K+ telemetry traces, automating feature validation and cleaning to ensure high data integrity for rigorous model testing.

PROJECTS

- **Tora: AI-Agentic Collaborative Workspace** — *Go, SvelteKit, ScyllaDB, Redis, LLM APIs*
 - Engineered a real-time, project-centric collaborative workspace featuring a shared code canvas, drawboard, and taskboard using a Go monolith and highly responsive SvelteKit frontend.
 - Integrated an autonomous AI agent capable of generating and editing project timelines and code, utilizing prompt caching and context management to optimize third-party API costs by 35%.
 - Architected a high-throughput persistence layer via ScyllaDB and Redis to manage global dashboards, persistent accounts, and real-time state synchronization across 75K+ concurrent clients.
- **Cross-Platform System Utility (macOS & Linux)** — *C++, Swift, System APIs*
 - Developed a native monitoring utility visualizing real-time CPU, memory, and network metrics as a highly customizable cross-platform task manager for UNIX-based environments.
 - Interfaced with low-level OS APIs to expose advanced customizations, enabling users to dynamically alter drag delays, key bindings, task schedulers, and sleep or wake behaviors to boost productivity.
 - Extended core operating system functionalities by implementing a persistent, media-rich clipboard history and custom configurations to automate complex user workflows and reduce manual input time by 25%.
- **Geo-Aware AI Marketplace Assistant** — *Python, FastAPI, RAG, LLM APIs, PostgreSQL*
 - Architected an intelligent conversational agent for a C2C marketplace, enabling users to discover optimal buyers and sellers through hyper-localized natural language queries.
 - Engineered an optimized RAG pipeline using LLM APIs, implementing advanced tokenization and semantic caching to reduce inference costs by 45% and achieve sub-second retrieval from 10K+ documents.

ACHIEVEMENTS AND PUBLICATIONS

- **Outstanding Master's Student:** Ranked 1/3420 for Research Excellence, Class of 2025.
- **CodeVita Coding Contest:** Ranked top 0.3% of 136,000 participants globally.
- **Master's Thesis:** Vioken: DQN-powered Adaptive Bitrate Control for VR.

PROFILES

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