

MAYUR REDDY

Senior Software Engineer | Full Stack | Distributed Systems & AI Systems

Austin, TX | +1 (352) 709-9161 | mayurjreddy@gmail.com
www.mayurreddyj.com | linkedin.com/in/mayur-reddy-j | github.com/raiden-02

Work Experience

Autodesk

May 2024 – Present

Sr Software Engineer – Generative AI

Apr 2026 – Present

3D Foundation Models | SDK & API Development | Platform Architecture | Product Integration

- Drive architecture and development of reusable SDKs, APIs, and versioned data contracts for generative 3D workflows, creating stable developer interfaces between foundation-model pipelines and downstream design products.
- Design and ship shared translation and integration components that convert generated 3D outputs into validated, product-compatible representations, enabling reusable model-to-product workflows across teams and applications.
- Define API and platform architecture across ML research, platform, and product teams, including schema evolution, compatibility, validation, versioned artifacts, identity/provenance, and reliable handoff patterns for shared developer workflows.

Software Engineer – Data, Generative AI

May 2024 – Mar 2026

3D Foundation Models | ML Infrastructure | Distributed Data Systems

- Built and operated petabyte-scale 3D data pipelines that transformed complex product data into training- and evaluation-ready datasets for foundation-model research, large-scale ML experimentation, and downstream processing on AWS.
- Developed a reusable Ray-based distributed processing framework for heterogeneous 3D workloads, isolating parser failures, bounding concurrency, checkpointing completed work, and producing durable outputs for recoverable large-scale processing.
- Built dataset cataloging, lineage, and provenance infrastructure tracing source assets through transformations and ML derivatives, improving reproducibility, dataset discovery, and experimental traceability.
- Built browser-based 3D streaming and inspection infrastructure for large models, enabling labeling, evaluation, and ML inspection workflows without requiring clients to download complete source datasets.

Autodesk

May 2023 – Aug 2023

Software Engineer Intern

SDK Development | Data Structures | Validation | Performance Engineering

- Developed a benchmark and validation system for a proprietary data structure library used across Autodesk SDKs supporting Civil 3D, Revit, Fusion 360, and Maya.
- Designed cross-language validation utilities for C++ and TypeScript implementations, enabling systematic comparison of library behavior, defect identification, and performance debugging with library developers.

Experian

December 2021 – May 2022

Software Engineer Intern

Backend Systems | API Design | Event-Driven Systems

- Designed and developed Java/Vert.x backend services and APIs for cloud infrastructure and consumer profile management using asynchronous, event-driven architectures.
- Tested, debugged, and optimized asynchronous application workflows while collaborating on service architecture and API design.

Projects

CyberRunner ArenaForge – Multiplayer FPS & Agentic Level Designer

[GitHub](#) | [Live Demo](#)

Real-Time Multiplayer | Agentic AI | Deterministic Evaluation | Game Systems

- Built and deployed a TypeScript/Node.js multiplayer FPS with 60 Hz server-authoritative simulation, client prediction/reconciliation, binary networking, and lag-compensated hit detection.
- Built a tool-using AI level-design system that generates playable 3D arenas, evaluates geometry and navigation with deterministic checks and seeded playtests, and iteratively revises maps before launching them in the live game.

ServeScope – Interactive LLM Inference System

[GitHub](#) | [Demo](#)

LLM Inference | GPU Performance | Workload Scheduling | Backpressure

- Built a GPU LLM serving workload-control and benchmarking system using vLLM and PyTorch, measuring first-token latency, end-to-end latency, queueing, throughput, and GPU utilization under mixed interactive and background traffic.
- Designed bounded background admission control that reduced interactive p95 time-to-first-token from 297 ms to 103 ms versus native vLLM priority in a controlled mixed-load benchmark, while explicitly measuring the resulting latency-throughput tradeoff.

Technical Skills

Languages: C++, Python, TypeScript, JavaScript, C#, Java

Backend, Cloud & Web: Full Stack Development, AWS, .NET, ASP.NET Core, Node.js, FastAPI, React, REST APIs, WebSockets, Terraform

Data & Distributed Systems: Ray, Apache Spark, Airflow, ETL, Distributed Systems, Data Modeling, Data Cataloging, Lineage, Metadata

AI & Graphics: Generative AI, RAG, AI Agents, Agentic Workflows, Foundation Model Pipelines, LLM Inference, vLLM, PyTorch, OpenGL, GLSL, Three.js, 3D Visualization

Education

University of Florida

August 2022 – May 2024

Master of Science in Computer Science

Jawaharlal Nehru Technological University Hyderabad (JNTUH)

June 2018 – July 2022

Bachelor of Technology in Computer Science and Engineering