Mayur Reddy

Senior Software Engineer, AI & Data
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*

Foundation Models | AI Agent Tooling (MCP, Context & Memory) | SDKs & APIs | Data Contracts

- Build versioned data contracts with Python/TypeScript bindings, automated validation, and compatibility checks, giving foundation-model pipelines and product integrations one shared interface.
- Design and develop shared translation components that convert generative-model outputs into validated, product-compatible design data, enabling reusable model-to-product workflows across teams.
- Build memory and checkpointing layers for production LLM agents: MCP tools that store and retrieve design data across prompts and iterations without overloading the context window.
- Define cross-team integration architecture for AI and agent workflows, specifying identity preservation, versioned artifact handoff, asynchronous updates, and recovery semantics across research, platform, and product systems.

Software Engineer – Data, Generative AI *May 2024 – Mar 2026*

Foundation Model Data | ML Infrastructure | Distributed Data Systems

- Built and operated petabyte-scale data pipelines on AWS, using Airflow orchestration and Ray Data to prepare heterogeneous datasets that feed foundation-model training, evaluation, and research.
- Developed a reusable Ray framework on Kubernetes (KubeRay) for distributed file processing with bounded concurrency, subprocess isolation, and checkpointing, so large extraction jobs recover from parser crashes and partial failures.
- 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 streaming and inspection infrastructure for large design 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

- Developed and optimized asynchronous Java/Vert.x backend services and APIs for cloud infrastructure and consumer-profile workflows, contributing to service design, testing, and performance debugging.

Projects

Murmur – LLM Inference Engine in C++/CUDA GitHub

LLM Inference | CUDA Kernels | Continuous Batching | Quantization

- Built an LLM inference engine from scratch in C++ and CUDA with hand-written attention, matrix and sampling kernels, a paged KV cache with prefix caching, and continuous batching, verified against Hugging Face Transformers to within 10^{-3} at every layer.
- Outperformed vLLM 0.28 on the same model, GPU and workload with FP8 weights (first token 9.9 vs. 18.0 ms, single-stream decode 329 vs. 166 tokens/s, 17% higher throughput at 64 requests) after Nsight profiling drove two attention-kernel rewrites.
- Added FP8/INT4 weights via a custom tensor-core GEMM (1.7×/2.2× faster decode, +2%/+9% perplexity) and grammar-constrained decoding (256/256 valid actions), serving 32 LLM-driven NPCs in a 60 fps game.

CyberRunner ArenaForge – Multiplayer FPS & Agentic Level Designer GitHub | Live Demo

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

- Built a tool-using AI level-design system that generates playable 3D arenas for a deployed multiplayer FPS, evaluates geometry and navigation with deterministic checks and seeded playtests, and iteratively revises maps before launching them in the live game.

Technical Skills

Languages: Python, C++, TypeScript/JavaScript, C#
Data & Distributed Systems: Ray / Ray Data, Kubernetes (KubeRay), Apache Spark, Apache Iceberg, Apache Airflow, AWS, Data Lakes, Data Lineage
AI/ML: LLM Applications, AI Agents, MCP, RAG, PyTorch, Hugging Face Transformers, vLLM
GPU & Performance: CUDA, Quantization (FP8/INT4), NVIDIA Nsight, Performance Profiling

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