

CHARU SNEHA LAGUDUVA RAVI

Tempe, AZ | +1 (623) 271-5755 | charusneha266@gmail.com | Portfolio | LinkedIn | Github

SUMMARY

Software engineer with 2+ years of industry experience focused on building intelligent systems with real-world impact - from autonomous workflows to scalable AI-driven products people actually use. Experienced in fast-paced agile environments, turning evolving ideas into production-ready solutions across backend systems and distributed agentic infrastructure.

PROFESSIONAL EXPERIENCE

- Software Engineer, Backend & AI Systems – RoundTechSquare** Tempe, AZ | January 2026 – Present
- Designed and shipped a agentic chatbot using ReAct-style planner with 8+ typed tool functions for task CRUD, querying, and analytics – with validation guardrails and observability tracing – achieving 92% tool-call accuracy for 100 labeled evaluation prompts.
 - Built a Python MCP server with JSON-schema-validated agent for planning, retrieval, and memory, backed by Redis short-term and PostgreSQL long-term memory layers – reducing hallucinated tool arguments from 45 to 20 per 200 conversations (55% drop).
 - Integrated agent stack with a Next.js backend via 12+ JWT-secured REST endpoints, containerized LLM inference using multi-stage Docker on GCP, and instrumented Prometheus monitoring – sustaining 1.3s p95 latency under concurrent production load.
- Software Engineer I, ML Systems – Rocket Mortgage (prev. Mr. Cooper)** Dallas, TX (Remote) | December 2022 – July 2024
- Engineered PyTorch/BERT/CRF document classification, field extraction, and cross-document validation pipelines over Kafka streaming across 2M+ monthly mortgage transactions – reducing manual annotation and verification effort by 70%.
 - Optimized batch and streaming GPU inference with async I/O, Redis caching, and fault-tolerant Kafka consumers – dropping p95 latency from 520ms to 390ms and sustaining 99.9% delivery reliability under peak production load.
 - Developed semantic document onboarding using sentence-transformer embeddings, FAISS ANN search, and quality detection – cutting deployment cycles to under 10 minutes and raising underwriter throughput from 18 to 25 docs/hour.
- Data Analyst Intern – JK Fenner (Siemens Group)** Chennai, India | January 2021 – March 2021
- Processed 10,000+ unstructured machine telemetry logs using Python and Pandas to surface downtime and bottleneck patterns – built Power BI dashboards adopted by managers, contributing 15% increase in machine utilization over the following 3 months.
 - Built statistical anomaly-detection pipelines using isolation forest models over 10,000+ sensor readings to identify recurring failure patterns – flagging 340+ actionable anomalies with a false positive rate under 8% for predictive maintenance decision-making.

PROJECTS (SELECTED)

- SearchLens – RAG Observatory** *GitHub* November 2025 – Present
- Built an end-to-end RAG search platform on a 5K-document corpus using React + TypeScript, FastAPI with 11 REST APIs, GCP deployment with Docker and CI/CD, and Gemini APIs for grounded answer generation, query rewriting, and relevance reranking.
 - Implemented BM25, dense bi-encoder, and hybrid retrieval behind pluggable strategy interface with Redis caching and two-stage cross-encoder reranker – cutting repeat-query latency from 480ms to 190ms (60%) and lifting nDCG@10 from 0.61 to 0.74 (+22%).
- NarrativeRift – Multi-Agent Contradiction Detection** *GitHub* December 2024 – April 2025
- Built a multi-agent NLP pipeline to detect semantic contradictions across 6 news sources (RT, Al Jazeera, SCMP, Reuters, BBC, AP) – extracting and embedding 1,611 propositions via Gemini APIs, BGE-M3 embeddings, and ChromaDB HNSW vector search.
 - Developed a two-stage NLI system using DeBERTa-v3 and Llama-3 verified via Groq, lifting contradiction detection precision from 52% to 89% and identifying 37 confirmed pairs – visualized as a 55-node/37-edge graph with 19 in a Streamlit dashboard
- Karigai – Condition-Aware Wellness Intelligence** *Live* March 2025 – Present
- Built a full-stack agentic wellness platform using Next.js 14, Supabase (PostgreSQL + RLS), and a context graph linking menstrual cycle phase, 18 health conditions (PCOS, thyroid, ED history), nutrition state, and fitness into unified LLM prompts.
 - Designed a condition-aware rules engine with RAG-style prompt injection (Anthropic + Groq) for meal, workout, and chat generation – with hard safety overrides for high-risk states (pain ≥ 8 , pregnancy + fasting), deployed on Vercel with 24+ releases.

SKILLS

Programming Languages: Python, TypeScript, Go, Java, C++, SQL, Bash

AI / ML & Generative AI: PyTorch, Hugging Face Transformers, Scikit-learn, Fine-Tuning, RAG, Embeddings, NLP, OCR, Distributed/Multi-GPU Training, DeepSpeed

Agentic Systems & Orchestration: Agentic Architectures (ReAct, Plan-Execute, Self-Reflection), Tool/Function Calling, Multi-Agent Systems, MCP, Memory-Augmented Agents, LangChain, LangGraph, Prompt Engineering

Backend & APIs: FastAPI, Node.js, Next.js, REST APIs, GraphQL, WebSockets, JWT/OAuth2, Microservices, Distributed Systems

Databases & Streaming: PostgreSQL, MySQL, Redis, Apache Kafka, ChromaDB, Idempotent Processing, Dead-Letter Queues

MLOps & DevOps: Docker, Kubernetes, GitHub Actions, CI/CD, Prometheus, Grafana, A/B Testing, Model Versioning

Cloud: GCP (Vertex AI, Cloud Run, Cloud Functions, BigQuery), AWS (EC2, S3, Lambda, ECS), Linux

EDUCATION

- Master of Science, Software Engineering** Aug 2024 – May 2026
Arizona State University, Tempe, AZ (Full-time) **GPA – 4.0/4.0**
- Bachelor of Engineering, Computer Science** Aug 2019 – July 2023
Anna University (TCE), Chennai, India **GPA – 3.82/4.0**