Ritvik Singal

Urbana, IL | (217) 841-1652| ritvik.singal10@gmail.com | linkedin.com/in/ritvik-singal

EDUCATION

University of Illinois Urbana-Champaign — Urbana, IL
Bachelor of Science in Computer Engineering — Expected May 2028

- Relevant Courses: Computer Systems & Programming, Discrete Structures, Software Design, Data Structures
- Current Coursework (Fall 2026): Probability and Statistics, Algorithms & Computation

SKILLS

Languages & Tools: C, C++, Python, JavaScript, SQL, HTML/CSS, Linux, Ghidra, MATLAB/Simulink, Slurm
AI/ML & Agent Systems: LangGraph, LangChain, RAG, MCP, LLM Evaluation, OpenCV, YOLO, NumPy
Infrastructure: Kubernetes, Docker, Helm, ClickHouse, MySQL, OpenTelemetry, Distributed Systems, WebSockets

WORK EXPERIENCE

Artificial Intelligence Intern | OpenTelemetry, Distributed Systems, LLM Infrastructure — Gurgaon, India
Ixigo — Jun 2026 - Present

- Establishing **Heimdall**, a closed-loop **observability platform** for **agentic systems** using a **5-level tracing hierarchy** and **10+ span families**, achieving **99.9% session reconstruction** and **p95 queryability** under 10 seconds
- Formalizing deterministic **session replay** by persisting model calls, retrievals and tool I/O into **immutable S3 bundles**, enabling **100 identical reruns** of production failures and **CI validation of 1,000-scenario regression suites**
- Operationalizing a **self-healing QA pipeline** that clusters **recurring failures**, proposes fixes across **13 patch categories** and validates impact through cohort simulation and canary-based production rollout.

Artificial Intelligence Intern | Python, LangGraph, FastAPI, WebSockets — Uttar Pradesh, India
PwC — Jun 2025 - Aug 2025

- Built **Orion, an agentic deployment system** in Python, **LangGraph, and FastAPI** that parses **CI results**, gates pass or fail and triggers releases, streaming live status over **WebSockets** and cutting **development-to-release time by 60%**
- Converted the deployment agentic API to use **WebSockets** with **FastAPI**, enabling real-time communication and reducing **response latency** by over **45%**, improving interaction between users and the agent interface
- Built a dashboard and helpdesk animation in React.js to visualize agentic **AI workflows**, reducing manual documentation and unit test generation by **76%**

EXTRACURRICULARS

Artificial Intelligence Research Intern | Python, RAG, LLM APIs, MCP Servers, Slurm — Urbana, IL
Center for AI Innovation, NCSA, University of Illinois Urbana-Champaign — Sep 2025 – Present

- Devised a **deduplication and classification pipeline** for **Delta support tickets**, cutting **2,815 Q&A pairs to 1,878** across topic groups and **extended it to handle incremental batches** cleanly
- Designed and executed **5+ validation and stress tests** using **NVIDIA CUDA** samples and **MLCommons GPU benchmarks** across **8 nodes**, confirming zero node faults over **6-hour** sustained load **multi-node HPC environments**
- Evaluated **OpenCode** across real **HPC workflows** including **CUDA compilation**, Python environment setup and parallel job submission, documenting failure modes to inform AI tool integration decisions for the Delta support system

PROJECTS

Chergi AI | Python, React, CesiumJS

- Processed and normalized raw **NASA satellite data (AQI, humidity, temperature)** across **6** Chicago city regions to create a unified environmental dataset for **urban-planning analysis** and site-selection decisions
- Visualized environmental data as a dynamic Health Score on an **interactive 3D map** to enable real-time **comparative analysis** and data-driven decision support when proposing new buildings, built with **React and CesiumJS**
- Constructed a **Python-based simulation engine** that dynamically **recalculated environmental metrics** when users added **interventions (parks, water bodies, reflective surfaces)** to model **urban design impact** across regions

Prism AI | TypeScript, React, Supabase

- Architected an **intelligent LLM routing system** that dynamically classifies user prompts across **4 categories** (SIMPLE, STEM, NEWS, CREATIVE) and routes queries to optimal AI models from **6+ providers** (GPT-4o, Claude, Gemini, Llama, Perplexity), reducing **inference costs by up to 70%** while maintaining response quality
- Engineered a **real-time streaming chat interface** with adaptive theming, **cursor-tracking AI mascots** and optimistic UI updates using React and TypeScript, **delivering sub-100ms perceived response times** and seamless model switching with **context-aware background gradients** for enhanced user experience
- Implemented a production-grade gateway service with automatic failover logic, hosted prompt management via **Keywords AI dashboard** and **Supabase authentication** with Row-Level Security (RLS), enabling zero-downtime prompt tuning and secure multi-user deployment with fractional-penny cost tracking.