Rahul Reddy Talatala

rahul.talatala@gmail.com | 716-939-5940 | linkedin.com/in/rahul-reddy-t | github.com/rahult18

Summary

GenAI Engineer specializing in LLM systems, agentic workflows, Graph RAG, and DevOps automation. Automated 60% of network triage at Verizon creating $2.46M annual cost avoidance, and reduced infrastructure debugging time 60% at Apple through fine-tuned LLM diagnostics and GPU observability systems.

Skills

Agentic & ML: LangGraph, LangChain, Autogen, CrewAI, PyTorch, TensorFlow, Google ADK, MCP, A2A
RAG & Knowledge Systems: Hybrid RAG, LlamaIndex, Neo4j, Elastic Vector DB, pgvector, Cohere Rerank, Semantic Search
MLOps & LLMOps: Ray, ONNX, MLflow, NVIDIA Triton, LangSmith, LangFuse, Kubeflow, W&B, Axolotl
Data Engineering: Spark, Airflow, Kafka, dbt, Snowflake, TimescaleDB, ETL/ELT, Data Warehousing, Data Modeling
Cloud & Deployment: AWS, GCP, Azure, Kubernetes, Terraform, Docker, Run:AI, ArgoCD, Github Actions, CI/CD
Programming: Python, SQL, Java, TypeScript/JavaScript, FastAPI, Spring Boot, Node.js, React

Experience

GenAI Engineer Aug 2025 – Present
Infinite Computer Solutions, Client: Verizon *Dallas, TX*

- Led automation of 60% of network triage as measured by 47,068 annual hours created and $2.46M cost avoidance, by designing a LangGraph multi-agent system with LlamaIndex and hybrid RAG across vendor logs and EMS data.
- Developed telecom ontology improving first-pass RCA accuracy as measured by 40% reduction in investigation time, by building a Neo4j knowledge graph and implementing hybrid Graph RAG using Cypher traversal and vector embeddings.
- Detected outage anomalies earlier as measured by 40% reduction in Mean Time To Detect, by building a predictive model on 5,000+ daily alarms using gradient boosted features and deploying via FastAPI inference service.
- Enhanced context precision by 35% as measured by alert accuracy logs, by implementing Elastic vector retrieval with Cohere reranking and enforcing strict JSON guardrails.
- Constructed a multimodal audit pipeline addressing $114M+ financial exposure as measured by audit baselines, by integrating Gemini extraction, pgvector embeddings, and 70+ automated guardrail rules.
- Established end-to-end agent observability as measured by 30% reduction in debugging cycles, by integrating LangFuse tracing and transaction-level confidence scoring across workflows.

Software Engineer – GenAI Infra (Contract) Apr 2025 – Aug 2025
Apple Inc., Data Platform Efficiency *Dallas, TX*

- Engineered a Kubernetes diagnostic system reducing infrastructure triage time by 60% as measured by incident resolution logs, by constructing an MCP-based debugger using LangGraph orchestration and gRPC streaming.
- Optimized LLM inference latency by 35% as measured by GPU benchmarks, by performing LoRA fine-tuning of Qwen 1.5B on synthetic telemetry data and deploying via NVIDIA Triton.
- Prepared structured fine-tuning datasets improving diagnostic output accuracy as measured by reduced manual review cycles, by generating JSONL training data for controlled LLM outputs.
- Delivered $1.5M annual cloud savings as measured by cost dashboards, by building Spark pipelines ingesting 5M+ daily metrics into TimescaleDB for GPU utilization analysis.
- Maximized GPU utilization by 60% as measured by idle compute reduction, by implementing Run AI fractional scheduling across shared Ray workloads.
- Produced $17M monthly cloud cost transparency as measured by executive reporting metrics, by generating SKU-level Spark dashboards and automated forecasting reports.

Software Engineer May 2024 – Apr 2025
Eminent Services Corporation *Frederick, MD*

- Revamped a legacy VB6 system improving scalability by 35% and reducing maintenance effort by 40% as measured by deployment stability metrics, by migrating to a modular MERN architecture with Azure DevOps CI/CD automation.

ML Research Assistant Jan 2024 – May 2024
University at Buffalo *Buffalo, NY*

- Reduced model energy usage by 20% as measured by inference benchmarks, by applying quantization and distillation techniques to GPT models.

Education

- **MS, Computer Science**, University at Buffalo GPA: 3.8/4
- **BTech, Computer Science**, Vellore Institute of Technology GPA: 3.9/4