Rahul Reddy Talatala

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

Summary

ML Engineer with 3 years of experience designing agentic AI systems, fine-tuned LLM pipelines, and MLOps infrastructure across telecom, cloud, and enterprise platforms. Specializes in multi-agent orchestration, privacy-preserving data architectures, and production observability, consistently translating technical design into measurable cost savings and operational efficiency gains.

Skills

Agentic & ML: LangGraph, LangChain, Autogen, CrewAI, PyTorch, TensorFlow, HuggingFace, Google ADK, MCP, A2A
Foundation Models: GPT-4, Claude, Gemini, Llama, Mistral, Qwen
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, numpy, pandas, scikit-learn, FastAPI, Node.js, React

Experience

GenAI Engineer Aug 2025 – Present
Infinite Computer Solutions | Client: Verizon *Tampa, FL*

- Built a five-stage document pipeline that parses invoices and workbooks with Gemini 2.5 Flash, matches line items to purchase orders via hybrid vector and keyword search, and routes flagged items through a 70-rule audit engine, surfacing $114–140M in annual overcharge exposure across 500–2,000 capital projects.
- Preserved on-prem data privacy without sacrificing analysis quality by fine-tuning Qwen2.5-3B with LoRA to mask sensitive financial fields before any document reaches a frontier LLM, then using DSPy to optimize extraction prompts, lifting structured-field accuracy from 74% to 91%.
- Calibrated the document-extraction pipeline's per-field confidence scoring by validating outputs against a hand-checked 100-project baseline sample, measuring a roughly 16% average overcharge rate to set the audit-routing thresholds the 70-rule engine runs on in production.
- Solved the cold-start problem for a 346-table natural-language-to-SQL agent by building an offline pipeline that samples column values, infers table descriptions via LLM-based prompting, and constructs a 1,536-dimension ChromaDB vector store alongside a 17,772-edge NetworkX join graph, enabling accurate schema and join-path retrieval from day one across 11.8M+ rows with no manual annotation.
- Partnered with network engineering leads to architect a 5-agent LangGraph system, a supervisor coordinating four specialists on Claude Sonnet 4.6 that classifies engineer intent, routes work to the right specialist, and logs every decision through Langfuse, automating circuit decommissioning across 7 backend systems.
- Designed a multi-agent FiOS planning system that clusters copper-circuit addresses with DBSCAN and coordinates six specialist agents into prioritized build plans with cost estimates, hub assignments, and generated maps and reports, cutting manual fiber build planning effort by 65%.
- Maintained 100% reporting accuracy across 14 revenue dashboards serving 250+ engineers by building an agentic RAG assistant that retrieves from Airflow logs and a custom data-lineage engine, diagnoses pipeline failures, and streams root-cause answers with automated retry dispatch.
- Worked with network operations to build a LangGraph routing agent that runs vendor-specific LLM prompts grounded in a telecom glossary, triaging 278,350 tickets at 90% accuracy and eliminating roughly 47,000 hours of annual L1/L2 triage for a projected $2.46M in cost avoidance.
- Established an agent evaluation harness measuring hallucination rate, tool-call accuracy, and answer quality across 500+ synthetic test cases per release cycle, cutting regression-causing deployments by 50%.
- Collaborated with the platform team to layer vector search with reranking and schema-enforced guardrails on retrieval, and wire tracing and confidence scoring into regression and observability dashboards, boosting RAG answer precision by 35% and cutting debugging cycles by 30% across deployed agents.

MLOps Engineer (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 for real-time pod telemetry.
- Optimized LLM inference latency by 35% as measured by GPU benchmarks, by performing LoRA fine-tuning of Qwen 1.5B on 12K synthetic telemetry samples using Axolotl and deploying via NVIDIA Triton with dynamic batching.
- Improved fine-tuning data quality by generating 12K JSONL training examples via structured prompting, reducing manual annotation cycles by 50% and improving diagnostic output specificity.

- Delivered $1.5M annual cloud savings as measured by cost dashboards, by building Spark pipelines ingesting 5M+ daily GPU metrics into TimescaleDB and surfacing anomalies through automated threshold alerting.
- Maximized GPU utilization by 60% as measured by idle compute reduction, by implementing Run:AI fractional GPU scheduling across shared Ray workloads with priority-based preemption policies.
- Produced $17M monthly cloud cost transparency by generating SKU-level Spark dashboards with automated week-over-week variance forecasting for infrastructure leadership.

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

- Improved system scalability by 35% and cut maintenance effort by 40%, by migrating a legacy VB6 monolith to a modular MERN stack with Azure DevOps CI/CD and automated regression testing.
- Reduced average API response time by 45% as measured by load test benchmarks, by redesigning core REST endpoints with Node.js, Redis caching, and query optimization across a MongoDB backend.
- Cut manual QA effort by 55% as measured by test cycle duration, by integrating an LLM-assisted test generation pipeline using GPT-4o that produced Playwright regression suites directly from user story descriptions.

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

- Reduced GPT model energy usage by 20% as measured by inference benchmarks, by applying structured pruning, INT8 quantization, and knowledge distillation to compress transformer architectures for edge deployment.
- Improved cross-domain transfer performance by 12% as measured by zero-shot evaluation benchmarks, by implementing LoRA parameter-efficient fine-tuning on domain-specific scientific corpora using HuggingFace Transformers.

Solutions Engineer Aug 2022 – Aug 2023
Swym Corporation *Bangalore, India*

- Improved wishlist-driven purchase conversion by 28% as measured by A/B test lift across 50+ merchant cohorts, by engineering a neural collaborative filtering recommendation engine trained on 10M+ behavioral events using PyTorch, served via FastAPI on GCP.
- Reduced customer churn by 22% as measured by 90-day retention metrics, by developing an XGBoost propensity model on RFM features with real-time scoring integrated into the merchant analytics dashboard.
- Accelerated merchant analytics reporting 3x as measured by pipeline SLAs, by building Airflow-orchestrated ETL pipelines ingesting Shopify and BigCommerce event streams into a Snowflake warehouse with dbt transformations.
- Improved personalization experiment throughput by 35% as measured by experiment coverage logs, by designing a multi-armed bandit A/B testing framework using Thompson Sampling for real-time offer and recommendation selection.
- Reduced model retraining overhead by 30% as measured by MLflow experiment tracking logs, by building an automated retraining pipeline triggered on data drift signals with versioned model registry and staged rollout via canary deployments.

Projects

Prism-mem | *Python, SQLite, sentence-transformers, MCP, LLM, PyPI* Live | GitHub

- Built an open-source PyPI CLI that extracts project context from Claude Code transcripts and git diffs into semantic triples stored in SQLite with local embeddings, achieving 35x compression (411K bytes condensed to 11.7K characters of queryable knowledge).
- Auto-detects 45.7% of facts as outdated across 416 cross-session connections and regenerates CLAUDE.md, .cursorrules, and AGENTS.md from top-scoring triples; exposes the knowledge base via MCP servers for Claude Code, Cursor, and Codex.

ApplyAI | *LangGraph, Gemini, FastAPI, Next.js, Chrome Extension MV3, Supabase* Live | GitHub

- Built a full-stack job application assistant solo: a Chrome extension with a LangGraph DAG agent that autofills entire application forms (dropdowns, React Select, file uploads), resume-to-job match scoring with keyword breakdown, and an application tracking dashboard. Live on the Chrome Web Store.
- Architected a FastAPI backend with Gemini-powered resume parsing, Supabase vector storage for job-resume similarity scoring, and a Next.js dashboard with real-time application status tracking and KPI analytics.

SecureStack | *AWS EKS, ArgoCD, GitOps, Terraform, DevSecOps, Kubernetes* GitHub

- Designed an end-to-end DevSecOps pipeline with a three-tier architecture on AWS EKS, GitOps-driven deployments via ArgoCD, Terraform-managed infrastructure, and automated security scanning integrated into CI/CD.
- Configured Trivy container scanning, OWASP dependency checks, and OPA policy enforcement as GitHub Actions gates, blocking vulnerable builds from reaching production across all environments.

Education

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

Certifications

- **AWS Certified Developer – Associate**, Amazon Web Services Issued Dec 2024 – Expires Dec 2027