

Venkata Ramana

USA | +1 (716) 970-9622 | ramanareddynani1098@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

SUMMARY

AI Engineer with 4 years of experience building reliable, client-facing AI systems in enterprise and life-science settings. Own the full path from requirements with stakeholders to production deployment and ongoing iteration with end users. Co-Founder of Notes9, AI Engineer at FedEx Logistics, hands-on across agents, RAG, evals, and AWS/Azure infrastructure.

WORK EXPERIENCE

Notes9 | Co-founder | United States

Nov 2025 - Present

- Built end-to-end a production multi-tenant, multi-agent system on AWS Bedrock and Anthropic API with a hand-built ReAct tool-use loop, powering Catalyst, an agent grounded in a team's own research graph (literature, protocols, experiments, samples, lab notes).
- Lead customer discovery and onboarding with research teams, turning user feedback directly into product and agent behavior changes.
- Shipped NLP-to-SQL agents over user data and a multi-stage hybrid-ranked RAG pipeline federating live search across PubMed, Europe PMC, and OpenAlex.
- Engineered per-claim, span-level citations (Anthropic Citations plus heuristic grounding) so every claim links to its source, and an LLM-as-Judge eval harness on retrieval precision/recall to catch regressions pre-release.

FedEx | AI Engineer | United States

Mar 2025 - Present

- Built an LLM tool-calling system on Azure AI Foundry that automates tariff refund processing end to end, using a controlled sequence of tool calls rather than open-ended agent reasoning so every step is auditable, cutting per-client turnaround from 18 days to 3 days.
- Cross-verify refund amounts, banking details, and supporting documents across three internal systems, and automatically flag any mismatch before a refund is submitted.
- Built a RAG layer over frequently changing trade regulations using Azure AI Search so each claim is checked against the current rules, not stale ones.
- Work directly with a 30-person operations team, each managing 100+ client accounts, as their technical point of contact: gather requirements, run review sessions, and adjust the system based on their feedback.
- Delivered a Copilot Studio agent that gives the team a conversational interface to review flagged issues before sign-off, and automated client-facing outreach through Power Automate to collect required compliance documents.
- Built an evaluation harness that scores every discrepancy flag against known refund outcomes, holding the system at 85% precision and 89% recall and catching regressions before each release; false positives and negatives feed directly back into prompt and rule refinement.

Fundae | AI/ML Engineer | United States

Aug 2024 - Feb 2025

- Delivered a production multi-tenant RAG platform (FastAPI, Azure OpenAI, Blob Storage, Azure AI Search) for enterprise clients including a top-tier pharma company, working directly with stakeholders on requirements, onboarding, and rollout.
- Automated CI/CD with GitHub Actions, Azure DevOps, Docker, and Terraform, cutting operational overhead 80%.
- Built human and model-based evals and telemetry on Application Insights, reducing hallucination rates 35% across tenants; optimized retrieval with chunking, embedding versioning, and semantic caching, improving latency 40% and cutting token usage 25%.

Groovy Web | Machine Learning Engineer | India

Jan 2022 - Aug 2023

- Designed and deployed end-to-end ML pipelines (Python, TensorFlow, Scikit-learn, XGBoost) for classification, regression, and recommendation with A/B testing and drift monitoring; built PySpark/Hive/SQL ELT pipelines across distributed data lakes for ML-ready features.

TECHNICAL SKILLS

Languages: Python, SQL, Javascript, Typescript.

Gen AI: Claude, GPT-4, Azure OpenAI, AWS Bedrock, RAG, Agentic Workflows, Tool Use/Function Calling, MCP, LangChain, LangGraph, CrewAI, Pydantic AI, Prompt Engineering, Fine-tuning, Evals & Guardrails, LLMops.

ML: PyTorch, TensorFlow, Scikit-learn, HuggingFace, BERT, PySpark, Pandas, Feature Engineering, A/B Testing, Model Monitoring.

Backend & Data: FastAPI, Flask, Next.js, REST APIs, Microservices, Postgres, pgvector, ChromaDB, Azure AI Search, Elasticsearch, MongoDB.

Cloud & Infra: AWS, Azure, Terraform, Docker, Kubernetes, GitHub Actions, CI/CD, MLOps, System Design.

Observability: Distributed Tracing, Token/Cost & Latency Telemetry, LLM-as-Judge, LangSmith, Application Insights, CloudWatch.

Client & Domains: Requirements Gathering, Stakeholder Management, | Life Sciences, Finance & Compliance, Enterprise Saas.

CERTIFICATIONS & ACHIEVEMENTS

- AWS Certified Generative AI Developer - Professional - AWS
- Building with Claude API - Anthropic

Jul 2026 - Jul 2029

Jul 2026

PROJECTS

Mercor | Machine Learning Specialist

Nov 2025 - Feb 2026

- Kaggle strategy on AWS GPUs, 15-20% model gains via tuning and behavior analysis, LLM fine-tuning and NLP benchmarking across predictive, image-to-image, text-to-image, and CV tasks.

AI fellow | Handshake

Sep 2025 - Nov 2025

- Evaluated 500+ Kaggle models for deployment readiness; built modular RLHF pipelines with researchers, cutting experiment cycle time 40% and architectural support for LLM workflows

EDUCATION

University at Buffalo, State University of New York

Aug 2023 - Jan 2025

Masters, Artificial Intelligence