Ramesh Polisetti

rameshnaidux@gmail.com | +1 (972) 282-0802 | Austin, TX | LinkedIn | Portfolio

Professional Summary

AI-Driven Software Engineer with 5+ years of experience delivering production-grade applications, ML pipelines, cloud-native services, and scalable APIs across enterprise and healthcare systems. Skilled in Python, Flask/FastAPI, React, SQL, PySpark, ML/LLM model integration, and AWS-based deployments. Known for building reliable, efficient, and automation-focused solutions that enhance analytics, forecasting, personalization, and business intelligence workflows. Values clean engineering, predictable performance, and team collaboration.

Technical Skills

Generative AI & LLMs: LLM APIs, Llama, GPT-4/4o, RAG, Vector Databases (FAISS, Pinecone), LangChain, LoRA, PEFT, Tokenizers, Prompt Engineering, Hugging Face
Machine Learning: TensorFlow, PyTorch, Scikit-Learn, XGBoost, Autoencoders, CNN/RNN/LSTM, Transformers, NLP
Data Engineering: PySpark, Pandas, NumPy, SQL, ETL Pipelines, Data Warehousing (Redshift, Snowflake), Airflow
Backend & Cloud: Python (Flask/FastAPI), JavaScript, SQL, AWS (SageMaker, Lambda, S3, Redshift, CloudWatch), Docker, Kubernetes, Terraform
MLOps & DevOps: CI/CD (GitHub Actions, Jenkins), Model Deployment, Feature Stores, Monitoring, API Optimization
Other: Tableau, Power BI, Git, Jira, Metrics-Driven Development, Visual Studio Code, PyCharm, Colab, A/B Testing

Professional Experience

Software Engineer – Artificial Intelligence, *Adobe* Aug 2024 – Present | Texas, USA

- Built and deployed end-to-end ML pipelines — covering data ingestion, preprocessing, feature engineering, model training, evaluation, and production deployment in cloud environments.
- Integrated LLMs and AI agents into applications by implementing retrieval-augmented generation (RAG), vector search, and embedding-based knowledge retrieval to deliver context-aware, enterprise-ready intelligence.
- Designed and implemented scalable agentic workflows supporting real-time inference, task orchestration, and stateful multi-agent interactions for automation and adaptive user experiences.
- Optimized model performance and cost using techniques like quantization, pruning, and performance tuning, improving inference latency and resource efficiency across cloud deployments.
- Automated CI/CD and deployment workflows using containerization, safe rollout strategies, monitoring, and telemetry to ensure reliable, secure, and maintainable production ML services.
- Added robust observability and monitoring — including logging, tracing, metrics, model-drift detection, and incident response — improving live-site reliability and troubleshooting.
- Collaborated cross-functionally with product managers, data engineers, and platform teams to translate business needs into AI-driven features while ensuring compliance, privacy, and responsible AI principles.
- Converted ML prototypes into production-ready microservices using Flask/FastAPI, enabling stable APIs and seamless integration across multiple engineering teams.
- Improved data validation and preprocessing workflows by adding schema checks, automated anomaly detection, and structured pipelines — reducing data-related model failures and improving overall model reliability.
- Documented model behavior, API usage guidelines, and integration steps to streamline onboarding, increase clarity for internal teams, and standardize AI service adoption across the organization.

Projects:

Content Intelligence & RAG-Powered AI Services

- Built production-ready AI microservices using FastAPI and deployed scalable LLM-powered RAG pipelines (embeddings + vector search) to deliver context-aware content intelligence across Adobe products.
- Implemented schema validation and anomaly detection in data pipelines and added full observability (logging, tracing, telemetry), reducing data issues and improving reliability of AI-driven content services.

Agentic AI Workflows & Scalable Model Deployment

- Developed end-to-end ML pipelines and engineered agentic AI workflows supporting real-time inference and intelligent automation features used across Adobe's creative and marketing tools. Designed preprocessing logic, prompt templates, and output formatting for business users.
- Optimized model performance using quantization/pruning and collaborated cross-functionally to deploy secure, compliant, high-throughput AI services with clear API documentation and integration guidelines.

Machine Learning Engineer, *Anblicks* Apr 2022 – Jul 2023 | Hyderabad, India

- Developed end-to-end forecasting pipelines using ARIMA, RF, XGBoost, and LSTMs, improving medical supply prediction accuracy by 38% across 15+ hospital departments.
- Extracted, cleaned, and transformed over 500,000 EHR records using PySpark, SQL, Pandas, and NumPy to build reliable datasets.
- Experimented with early LLM fine-tuning approaches (LoRA/PEFT) to reduce inference cost and improve personalization for healthcare NLP tasks.
- Built APIs integrating ML predictions with hospital ERP systems, achieving 97% uptime and automating procurement workflows (70% less manual work).
- Deployed models to AWS Lambda and S3 with minimized cold-start latency, enabling near-real-time updates.

PROJECTS:

Medical Supply Forecasting Engine Enhancement

- Developed forecasting features such as rolling averages, lag variables, and seasonal indicators to enrich model inputs for ARIMA, XGBoost, and LSTM models.
- Improved forecast stability and accuracy across multiple hospital departments.

Senior Software Engineer, *Genpact* Jun 2019 – Apr 2022 | Hyderabad, India

- Engineered scalable web applications using Python (Flask), React.js, SQL, and Docker, improving system responsiveness and ensuring reliable performance for high-traffic internal tools.
- Designed and built RESTful APIs for analytics platforms, implementing efficient request handling, caching, and role-based authentication to reduce response latency by 32%.
- Developed robust integration layers that connected backend services with data-processing pipelines, ensuring smooth handoff between application logic and analytical services.
- Modernized application workflows by automating backend tasks, streamlining data flows, and reducing manual effort for cross-functional teams.
- Built reusable backend modules, middleware, and utility functions that standardized development patterns, reduced code duplication, and improved maintainability.
- Created CI/CD automation using Docker and GitHub Actions, cutting deployment effort by 60% and enabling predictable, repeatable release cycles.
- Participated in code reviews, sprint planning, and architecture discussions, contributing to cleaner design and development practices.
- Performed database query optimizations (indexing, refactoring joins), reducing key report generation times.
- Built reusable backend components and utility functions that standardized development patterns across teams.

PROJECTS:

Enterprise Analytics Platform Modernization

- Refactored and modernized a large enterprise analytics platform by restructuring monolithic Flask services into modular API components, improving maintainability, reducing dependencies, and increasing developer productivity across teams.
- Optimized backend performance by redesigning SQL queries, implementing caching layers, and adding standardized authentication and validation middleware—reducing report generation times and improving overall system throughput.
- Automated recurring data-processing and API integration tasks using Python scripts and scheduled workflows, reducing manual effort for cross-functional teams and improving system reliability.

EDUCATION

Southern Arkansas University, Arkansas, USA Aug 2023 - May 2025
Master of Science in Computer Science

CERTIFICATES

Infosys - Power Programmer Virtual Experience Program (Link)
IBM - Machine Learning with Python (Link)
Cognizant – Artificial Intelligence Job Simulation (Link)
IBM - SQL and Relational Database (Link)