Rakshith Roy Gantagogula

AI/ML Engineer

+1(972) 730-4157 | rakshithroygantagogula@gmail.com | LinkedIn

SUMMARY

AI/ML Engineer with 3+ years of experience designing and deploying machine learning and GenAI solutions across finance, healthcare, and technology sectors. Experienced in distributed training of LLMs, high-performance inference (AWS ECS), and model optimization (post-training quantization, 4x size reduction). Skilled in MLOps best practices, including CI/CD pipelines, infrastructure-as-code (Pulumi), and observability (AWS CloudWatch), achieving up to 20% cost savings. Proficient in Python and Golang, with hands-on expertise in LangChain, FastAI, CrewAI, Opensearch, AWS Bedrock, Azure ML, and RabbitMQ. Delivered AI-driven automation such as GPT-4 powered agents and no-code AI services, resulting in measurable productivity gains and business impact.

SKILLS

Programming Languages: Python, Golang, TypeScript, C++, JavaScript, Java, SQL

Machine Learning & AI: TensorFlow, PyTorch, Pandas, NumPy, OpenAI, Claude, LangChain, LangGraph, huggingFace, Onnx, Scikit-learn, XGBoost, ElevenLabs, DeepSpeed, RAG

MLOps & Deployment: Flask, Docker, ECS, RESTful APIs, CI/CD (GitHub Actions), NodeJS, ReactJS, Next.JS, NestJS

Cloud & Infrastructure: AWS (EC2, ECS, Lambda, SageMaker, Bedrock etc), Azure (Azure Functions, Azure ML), Pulumi

Data Engineering: Apache Spark, PostgreSQL, DynamoDB, ETL Pipelines, Kinesis, RedShift, Amazon RDS, Neon

EXPERIENCE

AI Application Developer, Advanced Management, USA — Sep 2025 - Present

- Architected and deployed a production-grade AI chatbot integrated into a customer-facing application using RAG and YAML-configured GPT tool endpoints enabling real-time data retrieval from BigQuery and Neon (PostgreSQL), including autonomous SQL query generation.
- Designed and implemented orchestration workflows using LangGraph, enabling multi-step reasoning, tool usage, and agent-based execution.
- Integrated Azure AI Foundry-hosted LLMs to power conversational intelligence with secure and scalable inference pipelines.
- Deployed the application on Google Cloud Run using containerized services (Docker), ensuring high availability and scalable serverless execution.
- Led AI Operations initiatives as Project Manager, managing a cross-functional team of 4–5 engineers and overseeing roadmap planning, sprint execution, and production reliability.
- Contributed to both backend (Python, APIs, agent architecture) and frontend (React, TypeScript, Java) development, delivering full-stack AI-enabled application features.

AI/ML Engineer, CGI Incorp, USA — Jul 2024 – Sep 2025

- Worked on AI/ML platform for scalable distributed training, finetuning, and inference supporting multi-billion parameter LLMs, computational models, and hosting hundreds of AI Agents.
- Engineered a GPT-4/Claude powered AI Agent using LangChain, automating information access and saving key departments 3+ hours/day, scaling GenAI adoption from 1000 to 4500 employees.
- Built an AI agent to review GitHub pull requests with CrewAI, reducing code review time by 30% and post deployment defects by 15%.
- Enhanced ML job monitoring via reducing retrieval times by 40%, by building an optimized ETL pipeline (DynamoDB to PostgreSQL) with an FastAPI Backend and interactive ReactJS UI for searching/filtering jobs.
- Deployed fine-tuned LLMs on AWS Bedrock and Azure ML, applying quantization and autoscaling to improve cost efficiency in production environments.
- Automated infrastructure provisioning and CI/CD pipelines using Pulumi, Docker, and GitHub Actions, ensuring consistent and reliable deployments

AI Engineer Intern, Xnode.ai, USA — Jan 2024 – Jul 2024

- Developed secure microservices in FastAPI for Slack integration, enabling storage and retrieval of messages in SQLite.
- Implemented Azure Key Vault for credential management and JWT authentication to strengthen API security.
- Built and tested LangChain pipelines for structured prompt storage and response generation in GenAI workflows.

Machine Learning Engineer Intern, Hexaware Technologies, INDIA — Dec 2021 – Mar 2022

- Designed and deployed an automated churn prediction system using ML algorithms like Logistic Regression, Decision Trees, and Gradient Boosting, improving customer retention by 25%.
- Developed a BERT-based sentiment analysis model with 90% accuracy and automated data pipelines for continuous extraction and retraining to assess customer satisfaction.
- Developed RESTful APIs using Flask to integrate the ML model into the bank's CRM system, enabling real-time churn prediction alerts for proactive customer retention efforts.

PROJECTS

LLM Assistant, The University of Texas at Dallas, Texas — **Jan 2023 – May 2023**

- Built a **retrieval-augmented generation (RAG) pipeline** using LangChain and Opensearch to reduce hallucinations in LLM outputs. Designed prompt templates and evaluators to improve response accuracy and relevance.

EDUCATION

Master's in Information Technology and Management, The University of Texas at Dallas, Texas — Aug 2022 – May 2024

Bachelor's in Computer Science and Engineering, Pragati Engineering College, AP, India — Jun 2018 – May 2022

CERTIFICATIONS

AWS Certified Solution Associate | HashiCorp Terraform Associate (003) | Databricks Generative AI Fundamentals