ROHIT GOSWAMI

East Newark, NJ | +1 551-260-1266 | rg57@njit.edu | linkedin.com/in/rohit-goswami07 | github.com/rohit07cf
AWS Certified Machine Learning – Specialty | Certified Kubernetes Administrator (CKA)

AI Engineer with 8+ years of experience architecting agentic AI frameworks, fine-tuning LLMs, RAG and deploying Kubernetes-based microservices at scale across enterprise, healthcare, and smart city domains. Certified Kubernetes Administrator and AWS ML Specialist with a proven track record of delivering production-grade intelligent systems.

TECHNICAL SKILLS

Programming: Python, Go, SQL, Java
Gen AI & Agents: OpenAI Agents SDK, LangChain, LangGraph, LiteLLM, FastMCP (MCP), A2A, Pydantic, tiktoken
LLM Providers: OpenAI, Anthropic (Claude), Google Gemini, Azure OpenAI
ML & NLP: PyTorch, Transformers, scikit-learn, SpaCy, NLTK, ONNX Runtime, XGBoost, TensorFlow, Keras, OpenCV
Fine-tuning & Safety: Llama Factory, LoRA, QLoRA, MLflow, LLM Guard, Guardrails AI
RAG & Retrieval: Pinecone, Elasticsearch, OpenSearch, BM25, Unstructured, PyMuPDF, Azure AI Document Intelligence
Infrastructure: Kubernetes, Docker, Helm, KEDA, Temporal, FastAPI, gRPC, MSAL, Redis Streams, AWS (SageMaker, S3, Glue, Redshift), Azure ML
Databases & Tools: Redis, PostgreSQL, MS SQL Server, Hive, HBase, AWS RDS, Git, GitHub Actions, Power BI

PROFESSIONAL EXPERIENCE

Lead Analyst – Data Science | *Infosys Limited* Jul 2022 – Present
- Architected a proprietary Agentic AI Framework with multi-step reasoning, model-native thinking (Claude), Agent/Tool Hooks, and AgentTree orchestration, powering autonomous enterprise workflows and reducing manual effort by 60%.
- Designed the Vibe Working agent on Microsoft Teams with MSAL auth, OBO token exchange via gRPC, and MS Graph APIs to auto-draft responses with Adaptive Cards for user approval.
- Launched a fine-tuning service (LoRA, QLoRA, Full) via Llama Factory with Temporal multi-queue workflows, KEDA ScaledJobs for on-demand GPU training, and MLflow registry cutting model turnaround by 50%.
- Created MCP tools (FastMCP) for Workday data retrieval, MSGraph calendar/supervisor lookups, and Bing Search grounding with asyncio concurrency.
- Engineered guardrails (LLM Guard + Guardrails AI) for real-time content safety and cortex-models hosting DeepSeek, BioMistral, Llama 2 with multi-GPU inference and ONNX Runtime, achieving 3x faster CPU throughput.
- Leveraged OpenAI LLMs for semantic search and document summarization, improving accuracy by 70% over keyword-based retrieval.
- Devised fraud detection models on ACH ODFI batch transactions with new risk elements, reducing false positives by 25%.
- Managed Kubernetes infrastructure (Helm, HPAs, KEDA, StatefulSets, Init Containers) for 15+ production microservices.

Research Assistant | *New Jersey Institute of Technology* Mar 2021 – Dec 2021
- Implemented CNN-based face emotion recognition achieving 95% accuracy through data augmentation and hyperparameter tuning.
- Produced interactive Power BI dashboards with KPI scorecards for NJ traffic analytics, enabling data-driven state policy decisions.

Senior Data Science Associate | *TheMathCompany Inc* Aug 2020 – Dec 2020
- Conducted A/B tests yielding 2% sales lift from web redesigns; forecasted sales via LSTM time series with 8% MAPE across 30+ countries.
- Spearheaded automated ETL pipelines using S3, AWS Glue, and Redshift, reducing data processing time by 40%.

Data Scientist | *Trinity Mobility Pvt Ltd* Apr 2017 – Apr 2020
- Awarded “Trinity Innovator” for predictive policing project (DBScan, Neural Network, KNN ensemble) reducing monthly crimes by 25% for Bangalore City Police.
- Engineered reusable Python packages for data quality, EDA, and model evaluation, accelerating baseline model development by 70%.
- Orchestrated ML workflows via AWS Step Functions and SageMaker; automated SQL-to-AWS-RDS migration using DMS, saving 20 hrs/month.

Data Analyst | *Envibyte Technologies Pvt Ltd* May 2016 – Feb 2017
- Delivered server-side APIs (Python, Flask) and stored procedures for an application serving 7,000+ daily users; created Power BI KPI scorecards.

KEY PROJECTS

Predictive Policing (Public Safety)
- Conceptualized an ensemble of ML models (DBScan, Neural Network, KNN) to reduce monthly crimes by 25%; established CI/CD pipelines and hosted models as inference pipeline behind a single endpoint.

Parking Revenue Optimization (AIoT Smart City)
- Pioneered LSTM-based occupancy and demand forecasting models for parking lots, optimizing parking revenue by 30%.

Solid Waste Management
- Constructed a Random Forest model predicting smart bin fill levels with 9% MAPE, reducing waste management budget by 35%.

EDUCATION

New Jersey Institute of Technology – *M.S. in Data Science* Jan 2021 – May 2022
Marathwada University – *B.E. in Electrical & Electronics Engineering* Jun 2010 – Nov 2015