Monish Galla

monishgalla18@gmail.com   +1 (940) 843-8147   Linkedin - Monish_Galla

PROFESSIONAL SUMMARY

AI and Machine Learning Engineer with 4+ years of experience designing, training, and deploying production-grade machine learning and generative AI systems. I specialize in building end-to-end ML pipelines for NLP, computer vision, and time-series analytics using frameworks such as PyTorch and TensorFlow. My expertise includes LLM fine-tuning, retrieval-augmented generation (RAG), and agentic AI, along with strong MLOps practices using Docker, Kubernetes, and CI/CD pipelines. I have hands-on experience optimizing models for low-latency, high-throughput inference in real-time, cloud-native environments. I am passionate about engineering scalable AI solutions that solve complex business problems with measurable impact.

TECHNICAL SKILLS

- **Programming Languages:** Python, SQL, Java
- **Frameworks & Libraries:** TensorFlow, PyTorch, Scikit-learn, Keras, OpenCV, Langchain, LlamaIndex, AutoGen, LangGraph, YOLO, CNNs
- **AI/ML & Generative AI:** Deep Learning, Transformers, NLP pipeline development, RAG, LLMfine-tuning (LoRA, QLoRA,PEFT), Azure AI Foundry, AWS Agentic AI, chatbot development & optimization, Prompt Engineering, Multimodal data processing
- **Cloud & MLOps:** AWS, Azure AI Services, Azure ML Studio, Azure OpenAI, Docker, Kubernetes, FastAPI, CI/CD, Model Versioning & Deployment, Workflow Orchestration, Cloud Run, Vertex AI Workbench
- **Data & Visualization:** PostgreSQL, MySQL, OracleDB, Power BI, Tableau, Matplotlib, data Analysis & Visualization, Data Modeling & Pipeline Optimization, Snowflake
- **Agile & Collaboration:** Agile/Scrum Methodologies, JIRA, Cross-Functional Team Collaboration, Production & Operations Alignment, Stakeholder Communication & Business Process Integration

PROFESSIONAL EXPERIENCE

Generative AI Engineer | Fiserv, USA
Sep 2024 – Present

- Implemented MLOps workflows for scalable deployment, monitoring, and retraining of AI/ML models.
- Fine-tuned LLMs using structured and unstructured datasets to improve accuracy and domain alignment.
- Built LLM-powered conversational agents for automating customer interactions and improving support response times.
- Designed and deployed Agentic AI systems enabling autonomous workflow creation for fraud analytics and automated decision-making.
- Developed high-throughput time-series analytics pipelines integrating OLAP batch processing with real-time inference.
- Optimized models for latency, throughput, and cost, implemented drift monitoring and retraining pipelines.
- Developed multi-agent reasoning-based agentic AI processes for independent fraud investigation and decision assistance.
- Integrated Document AI and enterprise content search to automate extraction and retrieval of structured and unstructured financial data.
- Optimized domain-specific LLMs with LoRA and PEFT methods to lower inference costs and increase the accuracy of fraud categorization.
- Created pipelines for end-to-end data ingestion and validation to guarantee high-quality inputs for downstream analytics, inference, and LLM training.
- Integrated vector databases and embedding-based retrieval (RAG) to enhance contextual accuracy of LLM-powered applications.
- Led the design and implementation of secure, scalable GenAI architectures aligned with enterprise compliance standards, enabling reliable deployment of LLM-based applications in regulated financial environments.

Machine Learning Engineer | eBay

Mar 2024 – Aug 2024

- Led end-to-end ML projects, from research and model development to production deployment and monitoring.
- Developed advanced NLP solutions (Transformers, BERT, GPT) for sentiment analysis, chatbots, and document classification. Implemented full MLOps pipelines using Docker, Kubernetes, Jenkins, MLflow, and Model Registry to automate deployment, versioning, drift monitoring, and model lifecycle management.
- Designed and deployed AI models used directly by end users, improving customer satisfaction and interaction quality.
- Developed production-grade AI services that combine vector retrieval and API-based inference with Hugging Face models.
- Enhanced conversational AI foundation through embedding optimization and hybrid RAG methods.
- Developed interactive dashboards and visualizations (Tableau, Power BI, Matplotlib) to communicate insights effectively. Participated in technical planning and project estimation for ML and GenAI initiatives.

Machine Learning Engineer | Kotak Mahindra Bank

July 2022 – June 2023

- Designed and implemented end-to-end ML pipelines, including data ingestion, preprocessing, model training, and deployment. Optimized model performance with hyperparameter tuning, ensemble methods, and cross-validation.
- Built deep learning models (CNNs, RNNs, LSTM) for NLP, computer vision, and time-series forecasting projects.
- Deployed AI/ML solutions to cloud environments (AWS/GCP) with CI/CD pipelines and version control.
- Implemented real-time data pipelines using Apache Kafka and Spark Streaming for ML model consumption.
- Worked with cloud platforms (GCP Vertex AI, Azure ML) to deploy and scale ML solutions. Collaborated with cross-functional teams to translate business requirements into ML solutions with measurable outcomes.

Data Scientist | Wipro

Jan 2021 – Jun 2022

- Assisted in data collection, cleaning, and preprocessing using Python (Pandas, NumPy) for ML model training.
- Implemented basic supervised and unsupervised ML models (regression, classification, clustering) with scikit-learn.
- Developed visualizations with Matplotlib and Seaborn to support model evaluation and reporting. Assisted in building ETL pipelines using Apache NiFi and Airflow to automate data processing workflow.
- Gained hands-on experience with SQL and NoSQL databases for storing and retrieving structured and unstructured data.

CERTIFICATION

- Completed AWS Machine Learning Associate certification - AWS (May 2025)
- Achieved Google Professional Machine Learning Engineer Certification - Google (Jan 2026)

EDUCATION

Master's in Artificial Intelligence
University of North Texas, TX **| Aug 2023 – May 2025**