Dheeraj Bhaskaruni

Auburn, AL | 4083345204 | dzb0098@auburn.edu | LinkedIn | GitHub | Portfolio Website

Summary

AI Engineer with 2+ years building end-to-end ML and LLM solutions, from data processing, data preparation, and feature engineering to model development, production deployment, and monitoring. Skilled in designing real-time streaming ML pipelines processing millions of events daily, fine-tuning compact LLMs, and implementing production anomaly-detection services. Experienced in Python, TensorFlow/PyTorch, cloud AI services (AWS, Azure, Vertex AI, GCP), and scalable ETL pipelines within agile teams, collaborating cross-functionally.

Skills

Machine Learning: Machine Learning Algorithms, Model Development, ML Pipelines, Model Evaluation, Feature Engineering, Classification, Anomaly Detection, Time-Series Modeling, Statistical Modeling, Computer Vision
Deep Learning & LLMs: PyTorch, TensorFlow, Keras, LLM Fine-Tuning, Prompt Engineering, Generative AI, LangChain, Embeddings, NLP Pipelines, OpenAI/Gemini API
MLOps & Cloud: AWS, Azure, GCP, Docker, Kubernetes, MLflow (Experiment Tracking), CI/CD Pipelines, FastAPI, Snowflake/Databricks, Production Deployment
Programming: Python, pandas, NumPy, scikit-learn, PySpark, SQL, APIs/Flask, JavaScript/TypeScript, Java/Spring Boot, C++/C#
Data Engineering & Governance: ETL/ELT Pipelines, Spark/PySpark, SQL, MongoDB, InfluxDB, Data Lake, Parquet/Delta, Airflow, Data Quality Gates, Access Controls
Methodologies & Practices: Data Processing, Model Deployment, Agile Teams, Prototyping Solutions

Work Experience

Auburn University RFID Lab — **Nov 2023 - Present**
Graduate Assistant/Automation Developer & ML Engineer — *Auburn, AL*

- Built and shipped real-time ML pipelines processing 2M+ events/day with <2s latency (Flask + InfluxDB), including data preprocessing + feature engineering for model-ready datasets powering dashboards and rapid iteration, and data processing for model readiness and downstream pipelines.
- Mined historical fraud cases using similarity-based retrieval (vector search) to surface recurring fraud patterns and nearest-neighbor explanations for analysts, with relevant GLM statistical analysis.
- Fine-tuned LLMs with custom token grammar and JSONL pipelines, achieving Acc 0.971 / F1 0.973 / AUC 0.993; rapid prototyping of generative AI microservices, deployed with mixed precision and PEFT for reproducible, low-latency inference integrated into production scoring services.
- Developed a streaming anomaly-detection service (Autoencoder + Bi-LSTM) with 97% precision / 92% recall, adding LLM-generated root-cause summaries for analysts.
- Productionized ML systems as governed data products using AWS Sagemaker and MLflow/CI/CD pipelines: event-driven data lake, near-real-time dashboards, feature stores, lineage and quality gates, and observability across sites, documented model behavior and experiments, improving reproducibility and hand-off.
- Engineered and collaborated in agile teams with cross-functional stakeholders across Walmart to deliver a real-time barcode-to-RFID reconciler (Zebra FX9600), proving value in one store and scaling to nine Walmart Retail stores, establishing a deployable pattern for retail AI systems. Collaborated with senior engineers and cross-functional teams to implement ML solutions, enabling scalable store deployments.
- Designed A/B tests with power and confidence-interval guardrails; performed cohort/time-series analysis on multi-million-event logs to quantify lift and de-noise seasonal effects.
- Implemented data validation, schema/version control, and access controls for governed datasets.

Auburn University Library — **May 2023 - 2024**
Graduate Research Assistant — *Auburn, AL*

- Automated metadata extraction for 10,000+ artifacts using spaCy NER, cutting manual processing effort by 50% and boosting search recall by 40% through improved entity recognition, ETL pipelines, and data integration workflows.
- Mentored two graduate students to refine NLP pipelines, enhancing text-analysis accuracy and contributing to more reliable prompt-engineering and linguistic preprocessing methods across research projects. Supported research by exploring new techniques, tools, and frameworks.

IBM India **Jan 2021 - Dec 2023**

Azure Full Stack Developer & Cloud Consultant *Bengaluru*

- Developed and deployed 20+ .NET Core microservices on Azure, reducing deployment time 40% via automated CI/CD pipelines and scalable service architecture; integrating AI/ML APIs for real-time inference in enterprise applications.
- Led 7 hybrid-cloud migrations, lowering operational costs 22% while maintaining high-availability API integrations and enterprise-scale performance across distributed systems.

Projects

GuardianAI

- Built GuardianAI, a real-time fraud detector using XGBoost + Qdrant vector search, deployed as a Snowflake-native ML service. Used Python for model development and evaluation phases.
- Mined historical fraud cases using similarity-based retrieval (vector search) to surface recurring fraud patterns and nearest-neighbor explanations for analysts.
- Productionized with FastAPI, MLflow, and Docker, integrated directly with Snowpark for scalable inference, suitable for all data protection laws.
- Azure OpenAI and Gemini give responses based on available scenarios in RAG data.

LLM RFID Event-Sequence Classifier (LoRA-Tuned Phi-3 Mini)

- Designed a custom RFID event-token grammar and fine-tuned Phi-3 Mini with LoRA to classify shrink-risk sequences, achieving Acc 0.971 / F1 0.973 / AUC 0.993, demonstrating compact LLMs can model subtle retail event patterns.

Anomaly Detection for IoT / RFID Streams (Autoencoder + Bi-LSTM)

- Built a streaming anomaly detection system (Autoencoder + Bi-LSTM) for RFID/IoT event streams, achieving 97% precision / 92% recall with real-time scoring and LLM-based root-cause summaries.

End-to-End RFID Data Pipeline (Store - ML - Dashboard)

- Built an end-to-end streaming pipeline (FX9600 - MQTT - InfluxDB - Python ETL - ML Models - Power BI) supporting near-real-time shrink analytics across store pilots, in Azure. Used Pandas and NumPy for data processing and transformation.

Deepfake Detection (CNN-based Computer Vision Project)

- Fine-tuned a VGG16-based classifier with focal loss to detect real vs deepfake frames, achieving 86.9% precision on 20k labeled frames and exposing the model via a FastAPI inference service. Used Docker for production deployment and scalable serving.

Education

Auburn University **Feb 2023 - May 2025**

Master of Science *Auburn*

- Graduate Certificates: Data Engineering; Artificial Intelligence; Cybersecurity Engineering
- GPA: 3.8/4.0

SASTR A University **May 2016 - Sep 2020**

Bachelor of Technology

- GPA: 7.58/10

Certifications

- Machine Learning / AI Engineer - Codecademy (2025)
- Data Scientist: Machine Learning - Codecademy (2025)
- Generative AI with Large Language Models - Coursera (2025)
- Azure Cloud Microservices - IBM (2023)

Publications

- C. Turner, D. Bhaskaruni, X. Wang, J. Zhang, S. Mao, S. Periaswamy, J. Patton. Digital Twin of Retail Environment Using RFID Particle-Filter Localisation.