Dheeraj Rahul Reddy Piduru

AI/ML Engineer | 4 years

Dallas, Texas |

Professional Summary

AI/ML Engineer with **4 years of experience** building and deploying scalable machine learning and GenAI solutions across enterprise environments. Strong focus on production systems, including RAG pipelines, LLM-powered applications, and end-to-end ML lifecycle using AWS and MLOps frameworks. Proven track record of improving model performance, reducing manual workflows, and delivering measurable business impact through data-driven solutions. Experienced in translating complex business problems into reliable, production-ready AI systems.

Technical Skills

- **Programming & Querying:** Python (NumPy, Pandas, PySpark), SQL (Joins, Window Functions, Query Optimization), REST APIs, Data Structures, Feature Engineering
- **Machine Learning:** Supervised Learning, Unsupervised Learning, Model Selection, Hyperparameter Tuning, Ensemble Methods (XGBoost, LightGBM), Model Evaluation
- **Deep Learning:** PyTorch, TensorFlow, Neural Networks, CNN, RNN, Transformers, Transfer Learning
- **Generative AI & LLMs:** Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Prompt Engineering, Hugging Face Transformers, OpenAI API, Azure OpenAI, LangChain, LlamaIndex, Embeddings, Semantic Search, LLM Evaluation, Guardrails
- **LLM Fine-Tuning & Optimization:** LoRA (Low-Rank Adaptation), QLoRA, Parameter-Efficient Fine-Tuning (PEFT), Instruction Tuning, Model Quantization
- **Vector Databases & Search:** FAISS, Pinecone, Weaviate, ChromaDB, Similarity Search, Embedding Pipelines
- **MLOps & LLMOps:** Model Deployment (Batch & Real-Time), ML Pipelines, CI/CD (GitHub Actions), MLflow, Experiment Tracking, Model Versioning, Monitoring (Data Drift, Performance), A/B Testing, Retraining Pipelines
- **Cloud Platforms:** AWS (S3, SageMaker, EC2, Lambda), Azure (Azure ML, Azure OpenAI), GCP (Vertex AI)
- **Big Data & Data Engineering:** Apache Spark, Databricks, Airflow, ETL Pipelines, Data Pipelines, Snowflake, BigQuery
- **Software Engineering & Systems:** Scalable ML Systems, Microservices Architecture, API Development (FastAPI), Docker, Kubernetes, Redis (Caching), Asynchronous Processing, System Design
- **NLP & Conversational AI:** BERT, Tokenization, Text Classification, Named Entity Recognition (NER), Sentiment Analysis, Topic Modeling, Document Processing, Conversational AI, Chatbots
- **Data Visualization & Analytics:** Tableau, Power BI, Matplotlib, Seaborn, Dashboarding, KPI Tracking

Professional Experience

Adobe | AI/ML Engineer **Jul 2025 – Present | USA**

- Rolled out a production RAG pipeline combining embeddings and vector search, cutting document retrieval time by 40% while improving answer accuracy and making internal knowledge access faster for business users
- Shipped LLM-driven automation features using OpenAI and Hugging Face, reducing manual content work by 45% and improving consistency across high-volume workflows used by multiple internal product teams
- Set up AWS-based ML pipelines with SageMaker and Airflow to handle large-scale data processing, supporting over 10M records monthly and stabilizing batch and near real-time inference operations
- Introduced MLflow tracking for experiments, model versions, and performance monitoring, helping teams catch drift issues early and lowering production failures by around 25%
- Built FastAPI-based ML services and integrated them into existing Adobe systems, supporting high request volumes daily while keeping latency low and improving service reliability
- Worked closely with product managers and engineers to turn vague requirements into working AI solutions, contributing to measurable gains in user engagement and improving key product metrics by ~20%

Accenture | AI/ML Engineer **Sep 2020 – Oct 2023 | India**

- Delivered machine learning solutions for customer analytics and forecasting using Python and SQL, improving prediction accuracy by nearly 18% and enabling more reliable planning decisions for enterprise clients
- Engineered data pipelines using PySpark and Airflow to process millions of records daily, cutting data preparation time by 35% and improving stability of downstream ML workflows
- Built NLP pipelines using BERT and spaCy for classification and entity extraction, reducing manual document review effort by 40% and speeding up processing across business operations teams
- Deployed models using Docker and REST APIs, supporting both batch and real-time inference, and reducing release cycles by about 30% across multiple client environments
- Applied feature engineering and model tuning techniques using XGBoost and LightGBM, improving model performance by 20% and ensuring consistent results across different datasets
- Coordinated with business stakeholders to define use cases and KPIs, translating requirements into production-ready ML solutions that improved operational efficiency by roughly 15%

Education

Master of Science in Business Analytics & Artificial Intelligence **Dec 2025**
The University of Texas at Dallas, USA

Bachelor of Technology in Computer Science & Engineering **Jun 2023**
Mahindra University, Hyderabad, India