Vijay Kumar

Tel: (614)-706-7742 | Email: vijaykumar0071120@gmail.com

Summary

- **AI/ML Engineer** with 4+ years of hands-on experience in building, deploying, and maintaining end-to-end machine learning systems across **AWS, Azure, and GCP**.
- Specialized in **LLM fine-tuning**, **prompt engineering**, and **generative AI** with proven success in productionizing models like **GPT-4, LLaMA, and Claude**.
- Proficient in **Python**, with deep expertise in **PyTorch**, **TensorFlow**, **Scikit-learn**, and **Hugging Face Transformers** for scalable AI development.
- Implemented robust **MLOps pipelines** using **MLflow**, **Airflow**, **CI/CD**, **Docker**, and **Kubernetes**, ensuring high availability and fast iteration cycles.
- Integrated ML models with business applications using **FastAPI**, **OpenAPI**, and **REST APIs**, reducing latency and boosting operational efficiency.
- Built and optimized data pipelines using **PySpark**, **Delta Lake**, and **vector databases (Pinecone, ChromaDB)** to support real-time AI applications.
- Experienced in **model monitoring, drift detection, A/B testing**, and compliance with **AI governance and SOC 2** standards.
- Collaborative team player with a strong track record of working with cross-functional teams (engineering, product, business) to deliver AI solutions with measurable ROI.

Skills

- **Programming Languages:** Python, SQL, Java, Bash
- **ML & DL Frameworks:** TensorFlow, PyTorch, Scikit-learn, Keras, Hugging Face Transformers
- **AI/ML Techniques:** Supervised Learning, Unsupervised Learning, Deep Learning, NLP, LLMs (GPT, LLaMA, Claude), Prompt Engineering, Fine-tuning, Embedding, RAG
- **MLOps & Deployment:** MLflow, Airflow, CI/CD, FastAPI, Flask, Docker, Kubernetes, ONNX Runtime, Model Monitoring, A/B Testing
- **Cloud Platforms & Services:** AWS (SageMaker, S3, Lambda, Step Functions), Azure (Azure ML, OpenAI, Cognitive Services), GCP (Vertex AI, Cloud Functions)
- **Data Engineering:** PySpark, Apache Spark, Delta Lake, ETL, Pandas, NumPy, Vector DBs (Pinecone, ChromaDB)
- **Statistics & Analytics:** Hypothesis Testing, A/B Testing, Regression, Clustering, SHAP, LIME
- **Infrastructure as Code:** Terraform, AWS CloudFormation
- **Version Control & DevOps Tools:** Git, GitHub Actions, Jenkins
- **Data Visualization:** Power BI, Tableau, Matplotlib
- **API & Integration:** REST APIs, OpenAPI, Swagger, Postman
- **Collaboration & Agile:** Agile, Scrum, Jira, Confluence
- **Security & Compliance:** AI Governance, Explainability, Zero-Trust Architecture, IAM
- **Model Hosting & Serving:** SageMaker, Azure ML Pipelines, Docker, Kubernetes
- **Generative AI Tools:** Stable Diffusion, LoRA, DreamBooth
- **Productivity Tools:** Databricks, Snowflake, VS Code

Professional Experience

Verizon Wireless Systems — **Aug 2023 – Present**

AI/ML Engineer

- Spearheaded the deployment of scalable ML workflows on **Databricks** and **Azure ML**, reducing model release cycle by **60%** and supporting **24/7 high-availability AI services**.
- Fine-tuned **GPT-4 and LLaMA** using advanced **prompt engineering** and **retrieval-augmented generation (RAG)**, boosting LLM accuracy for internal knowledge search by **40%**.
- Led integration of **OpenAI APIs** into business operations, cutting customer service response times by **65%** and improving agent productivity across 3 global teams.
- Built secure, reusable **CI/CD pipelines** with **MLflow, Airflow**, and **GitHub Actions**, achieving **99.5% deployment success rate** with zero downtime in production.
- Optimized inference workloads using **ONNX Runtime** and **vLLM**, reducing GPU memory usage by **50%** without impacting model latency.
- Designed and containerized API endpoints via **FastAPI** on **Kubernetes**, improving model accessibility and reducing service latency by **30ms** per call.
- Implemented model monitoring with real-time **A/B testing** and drift detection, driving a **25% uplift** in conversion rates through retrained versions.
- Collaborated cross-functionally with product and data engineering teams to automate ETL pipelines in **PySpark** and **Delta Lake**, cutting feature readiness delays by **70%**.
- Published internal documentation for **explainable AI (XAI)** using **SHAP** and **LIME**, ensuring **100% audit compliance** with regulatory and governance policies.

Coforge **Jan 2020 – Aug 2022**

AI/ML Engineer

- Designed and launched predictive ML models to reduce customer churn, delivering **$1.2M+ annual savings** by increasing retention by **18%**.
- Built production-grade **fraud detection systems** using **AWS SageMaker**, integrated with real-time transaction flows via **Lambda** and **Step Functions**, detecting threats with **92% precision**.
- Automated high-volume **ETL pipelines** using **Apache Spark**, improving data processing throughput by **3x** and reducing model training time by **40%**.
- Developed robust **Flask-based REST APIs** for ML inference, enabling integration with enterprise platforms and serving **500K+ requests/month**.
- Deployed NLP-based document classifiers using **Hugging Face Transformers**, reducing manual review workload by **80%** across compliance teams.
- Standardized infrastructure using **Terraform** and **CloudFormation**, enabling secure, scalable ML environments across staging and production.
- Delivered stakeholder dashboards in **Tableau** to visualize model KPIs and business impact, influencing decision-making at the C-suite level.
- Enhanced model reliability with **version-controlled Git pipelines** and **Jenkins CI**, reducing rollback incidents by **90%** and improving deployment confidence.
- Mentored a team of 4 junior data scientists and engineers, accelerating their productivity and contributing to a **30% increase in team output**.

Education

- **Master of Science in Information Technology**
  University of Cincinnati, Ohio, USA
- **Bachelor of Technology in Computer Science and Engineering**
  Gitam Institute of Technology Visakhapatnam, India