Pavan Bharadwaj Annambhotla

Senior AI / ML Engineer

pavanbannam10@gmail.com
+1 334-686-4470
linkedin.com/in/pavanbharadwaj

PROFESSIONAL SUMMARY:

- **Senior AI / ML Engineer** with **4+ years of experience** designing, deploying, and optimizing **production-grade**, **cloud-native ML and LLM systems** at scale.
- Built and optimized **end-to-end ML pipelines** and **distributed inference services**, supporting **500K+ monthly inference requests** with **sub-second latency**.
- Designed **RAG-based** and **LLM-driven systems** using **vector search**, **embeddings**, and **prompt engineering**, improving **accuracy and relevance by up to 30%**.
- Developed **multi-agent workflows** and **LLM-powered automation** that reduced **analyst and decision-review time by nearly 50%**.
- Strong **MLOps / LLMOps** experience across **model monitoring**, **drift detection**, **experiment tracking**, and **CI/CD** using **MLflow, Prometheus, Kubernetes, and Azure DevOps**.
- Hands-on with **GPU-optimized training and inference** on **AWS and Azure**, reducing **latency, training time, and inference costs**.

Core Competencies: Generative AI, Large Language Models (LLMs), Machine Learning Engineering, Distributed Systems, Cloud AI (Azure & AWS), RAG, LLMOps, Scalable ML Systems

TECHNICAL SKILLS:

Programming	Python, C#, Go, JavaScript, TypeScript, SQL, Bash
ML & GenAI	PyTorch, PyTorch Lightning, TensorFlow Extended (TFX), JAX, LangChain, LangGraph, RAG, Vector Search.
AI/ML Tooling	OpenAI API, Azure OpenAI, Hugging Face Hub, Azure AI Studio, Triton Inference Server, MosaicML, MLflow
Frontend	React, Angular, TypeScript, REST API Integration, Data Visualization (Plotly, D3.js)
Data Engineering & Storage	PostgreSQL, MongoDB, Redis, Cosmos DB, Pandas, NumPy.
DevOps & Cloud	Azure, AWS, Docker, Kubernetes, Terraform, Jenkins, GitHub Actions, CI/CD.
Messaging & Orchestration	Kafka, Event-Driven Architecture, gRPC, Dagster.
Observability & Monitoring	Prometheus, Grafana, OpenTelemetry, Model Drift Detection

PROFESSIONAL EXPERIENCE:

Client: - PayPal **Jan 2024 – Present**

Role: - Senior AI / ML Engineer **USA**

- Own the design and deployment of **RAG-based systems** for finance and compliance workflows, improving retrieval accuracy by **38%** and significantly reducing manual review effort.
- Lead development of fraud-risk ML models with optimized feature pipelines, reducing false positives by **22%** while maintaining precision at high transaction volumes.
- Architect **multi-agent automation workflows** for KYC/AML investigations, cutting analyst review time by **50%** within two quarters through structured task orchestration.
- **Design and scale distributed, high-availability inference services** using FastAPI, async execution, vector caching, and GPU-backed inference, reducing API latency from **1.2s to under 500ms**.
- Optimize model training and inference using ONNX Runtime and GPU acceleration, achieving **2× performance gains** for fraud evaluation workloads.
- Build production-grade model monitoring, **model versioning**, and governance, including drift detection, anomaly dashboards, RBAC, encrypted data flows, and audit logging, reducing model-related incidents by **25%** while meeting PayPal's 2024 compliance requirements.

- Deliver LLM-powered case summarization and decision-support tools, incorporating **prompt engineering and LLM evaluation** for accuracy, latency, and safety, improving dispute-resolution speed by **30%** across operational workflows.

Environment:- Python, FastAPI, PyTorch, ONNX Runtime, LangGraph, Azure AI, Kubernetes, Redis, PostgreSQL, Prometheus, Grafana, Azure DevOps.

Client: - Paytm **Mar 2022 – June 2023**

Role: - Software Engineer (AI & Backend) **Hyderabad, India**

- Owned the design of ML-backed risk-scoring services for wallet and payment flows, improving fraud-detection recall by **20%** while maintaining sub**-700ms** latency at production scale.
- Architected **scalable**, Kafka-driven, event-based microservices processing **5M+ transactions per day**, achieving **99.9% uptime** across high-volume payment systems.
- Developed customer behavior and propensity models that increased promo-targeting accuracy by **25%**, directly improving campaign efficiency and ROI.
- Led migration of NLP and analytics workloads to GPU-optimized ONNX inference, delivering **2× performance gains** during peak traffic periods.
- Improved platform reliability by implementing retry orchestration, circuit breakers, and structured logging, reducing incident recovery time by **40%**.
- Drove CI/CD automation and containerized deployments using Docker and Jenkins, increasing deployment frequency from **biweekly to daily** without service disruption.
- Built ML-backed reconciliation and anomaly-detection checks, reducing settlement discrepancies by **28%** across quarterly reporting cycles.

Environment: - Python, FastAPI, Kafka, ONNX Runtime, AWS Lambda, Docker, Airflow, PostgreSQL, Redis, Jenkins, Pandas.

Client: - Paytm **Jan 2021 – Feb 2022**

Role: - Software Engineer **Hyderabad , India**

- Built and optimized high-throughput backend APIs for payments and refunds, improving response times by **35%** through async I/O, query optimization, and efficient connection handling.
- Developed anomaly-detection checks for payment failures and settlement issues, reducing undetected incidents by **30%** across monthly operational cycles.
- Designed and maintained ETL pipelines using Airflow and Pandas to support reporting, reconciliation, and internal analytics workflows.
- Improved system performance during peak traffic by implementing caching, batching, and concurrency controls across critical payment paths.
- Implemented secure API authentication, encrypted payload handling, and audit logging to support compliance-ready payment systems.

Environment: - Python, C#, FastAPI, Airflow, PostgreSQL, Redis, Pandas, AWS EC2, Docker, Jenkins, Git.

Education:

Master's: Computer Science **2025**

Auburn University at Montgomery United States