Pavan Bharadwaj Annambhotla **pavanbannam10@gmail.com**
AI / MI Engineer **+1 334-686-4470**

PROFESSIONAL SUMMARY :-

- AI/ML Engineer with 4+ years of experience designing intelligent systems that blend machine learning, LLMs, and cloud-native engineering to streamline operations and improve decision automation across enterprise environments.
- Delivered production-grade ML pipelines that reduced model training time by 35% (2021–2024) through optimized GPU utilization, automated experiment tracking, and refactored data flows.
- Built end-to-end AI services using Python, FastAPI, and React that supported more than 500K monthly inference requests while maintaining reliable sub-second response times.
- Designed RAG-based search and question-answering systems using vector indexing and custom scoring logic, improving retrieval accuracy by 30% in enterprise knowledge tools.
- Integrated Azure OpenAI, domain-tuned embeddings, and open-source LLMs into analytics applications, cutting manual content triage and research time by 40%.
- Developed multi-agent automation workflows that reduced human decision-review time by nearly half within a 12-month production cycle.
- Built real-time drift monitoring, metric tracking, and model lineage using MLflow and Prometheus, reducing model-related incidents by 25%.
- Engineered microservices and event-driven pipelines with Kafka and Kubernetes, enabling smooth horizontal scaling and processing millions of data events per week.
- Automated CI/CD for AI workloads using GitHub Actions and Azure DevOps, accelerating deployment cycles from several days to a few hours.
- Implemented A/B testing frameworks that improved predictive model accuracy by 20% across quarterly evaluation cycles.
- Designed AWS and Azure inference architectures that reduced operational costs by 15% through adaptive resource scheduling and optimized compute usage.
- Embedded explainable AI components for risk-sensitive decisions, creating clear audit trails and improving stakeholder trust across analytics teams.
- Built interactive visualization dashboards with React, Plotly, and D3.js to track model behavior, system health, and business KPIs in real time.
- Developed structured and unstructured data pipelines using Spark, Airflow, and PostgreSQL to support analytics, model training, and high-throughput ingestion.
- Created inference-optimized runtimes with ONNX Runtime and TensorRT, improving GPU throughput by 40% across production services.
- Authored reusable playbooks and templates for ML deployment, monitoring, and feature engineering, reducing onboarding and delivery time by 30%.
- Delivered multiple end-to-end AI solutions—spanning ingestion, training, deployment, and observability—that elevated automation, efficiency, and decision intelligence.
- Built evaluation harnesses for LLM and ML models that benchmarked latency and accuracy, reducing model comparison time by 45%.
- Collaborated with security teams to implement validation, red-team checks, and output-filtering safeguards that strengthened safe AI deployment.
- Mentored junior engineers and data scientists on deployment, optimization, and debugging practices, improving team execution quality and delivery speed.

TECHNICAL SKILLS: -

Programming	Python, C#, Go, JavaScript, TypeScript, SQL, Bash, Shell Scripting
ML Frameworks	LangChain, PyTorch Lightning, TensorFlow Extended (TFX), JAX, React, Angular
AI/ML Tooling	OpenAI API, Hugging Face Hub, Azure AI Studio, LangGraph, Triton InferServer, MosaicML
Generative AI	Agentic Workflow Design, Azure OpenAI, Custom LLM Tooling, Vector Sear Orchestration
Data Engineering & Storage	PostgreSQL, MongoDB, Redis, Cosmos DB, Pandas, NumPy, MLflow
DevOps & Cloud	Azure, AWS, Docker, Kubernetes, Terraform, Jenkins, GitHub Actions, CI/CD
Messaging & Orchestration	Kafka, Event-Driven Architecture, gRPC, Dagster, MLflow Tracking
Observability & Monitoring	Prometheus, Grafana, OpenTelemetry, Model Drift Detection

PROFESSIONAL EXPERIENCE :-

Client: - PayPal **Jan 2024 – Present**
Role: - AI / ML Engineer **USA**

Roles & Responsibilities:-

- Designed RAG pipelines for finance and compliance workflows, improving retrieval accuracy by **38%** and reducing manual review cycles across 2024.
- Developed fraud-risk ML models with optimized feature pipelines, lowering false positives by **22%** and improving precision in high-volume transaction streams.
- Built multi-agent automation for KYC/AML investigations, cutting analyst review time by **50%** over two quarters through structured task orchestration.
- Implemented high-throughput inference services with FastAPI, async processing, and vector caching, reducing API latency from **1.2s to under 500ms**.
- Built GPU-optimized training and ONNX inference achieving **2×** speedup for fraud-rule evaluation.
- Created drift and anomaly dashboards using Prometheus/Grafana, reducing model incidents by **25%**.
- Containerized AI microservices with Kubernetes and automated CI/CD pipelines, accelerating deployment cycles from monthly to **weekly**.
- Implemented RBAC, encrypted data flows, and audit logging to meet PayPal's 2024 compliance requirements.
  Delivered LLM-powered case summarization boosting dispute-resolution speed **30%**.

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 Specialist **Hyderabad, India**

Roles & Responsibilities:-

- Designed ML-backed risk-scoring services for wallet and payment flows, improving fraud-detection recall **20%** while maintaining sub-700ms latency across production traffic.
- Built Kafka-driven, event-based microservices handling **5M+** transactions/day with 99.9% service uptime.
- Developed customer-behavior scoring models that increased promo-targeting accuracy by **25%**, boosting campaign efficiency.
- Migrated NLP and analytics workloads to GPU-optimized ONNX inference, achieving **2×** faster processing under peak load.
- Implemented retry orchestration, circuit breaking, and structured logging to reduce incident recovery time **40%**.
- Automated CI/CD pipelines and containerized deployments with Docker/Jenkins, raising deployment frequency from biweekly to **daily**.
- Built ML-backed reconciliation and anomaly checks, reducing settlement discrepancies **28%** across quarterly cycles.
- Optimized FastAPI services using async I/O, caching, and connection pooling for higher throughput during surge traffic.
- Collaborated with product and data teams to embed backend APIs with analytics signals, improving internal decision workflows and reducing data lookup time by **35%**.

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**

Roles & Responsibilities:-

- Built high-throughput backend APIs for payments and refunds, improving response speed **35%** through async I/O and optimized SQL queries.
- Developed anomaly-detection checks for payment failures, reducing undetected issues **30%** across monthly cycles.
- Developed ETL jobs using Airflow and Pandas to streamline data preparation for reporting and internal analytics.
- Improved overall system throughput during peak hours through caching, batching, and better connection management.
- Created internal dashboards that gave real-time visibility into payment and settlement flows, helping ops teams respond more quickly.
- Implemented secure API authentication, encrypted payload handling, and audit trail logging for compliance readiness.
  Containerized services on AWS/Docker and strengthened deployment pipelines, lowering release failures **35%**.

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