**Amulya Kantamaneni**

**AI / ML Engineer • LLM Developer • Full-Stack AI Systems Engineer**

**amulya.kantamaneni098@gmail.com | (470)-815-6486 | Github | Linkedin | Website**

## PROFESSIONAL SUMMARY:

AI Engineer with 3+ years building end-to-end AI systems, LLM applications, RAG pipelines, and scalable backend architectures. Expert in Python, PyTorch, TensorFlow, React, Node.js, and cloud-native AI engineering. Proven ability to build low-latency, production ML pipelines, optimize inference by 35–50%, and deploy secure AI services on AWS, Docker, and Kubernetes. Skilled in LangChain, Hugging Face, vector search, and CI/CD, delivering high-traffic, business-ready AI products at scale.

## TECHNICALSKILLS:

**Languages:** Python, TypeScript, JavaScript, SQL, Bash
**AI/ML:** PyTorch, TensorFlow, Hugging Face, Transformers, LLMs (GPT, Llama, Mistral), XGBoost, LSTMs, CV models, Generative AI
**LLM Development:** LangChain, RAG, Prompt Engineering, Embeddings, Fine-Tuning, Multi-Agent Workflows
**Backend/Full-Stack:** Node.js, Express.js, React, Next.js, Tailwind CSS, REST APIs, WebSockets, JWT Auth
**MLOps & Deployment:** AWS (SageMaker, ECS, Lambda, S3), Kubernetes, Docker, GitHub Actions, Jenkins, MLflow, Model Registry
**Vector Databases:** Pinecone, Qdrant, Milvus, ElasticSearch, ChromaDB, FAISS
**Monitoring:** Prometheus, Grafana, Kibana, Model Drift Monitoring
**Data & Visualization:** Pandas, NumPy, Polars, Tableau, Matplotlib, Seaborn, PowerBI

## EXPERIENCE:

**Machine Learning Engineer | PayPal | San Jose, CA | March 2024 – Present**

- Built and deployed LLM-powered financial reasoning systems, including FinBERT, custom classification models, and RAG pipelines integrated into PayPal's global transaction workflow.
- Engineered scalable backend systems using Node.js + Python microservices, serving model inference to customer-facing apps with <200ms latency.
- Developed secure REST API + async inference pipelines for classification and chatbot agents, reducing response time by 35% via batching, caching, and optimized GPU utilization.
- Integrated ML services into frontend applications by building reusable React components for AI-powered insights, search experiences, and automated support flows.
- Automated model deployment with AWS SageMaker, ECS, Lambda, improving release efficiency and reducing operational overhead.
- Built real-time monitoring dashboards using Prometheus + Grafana to track drift, latency, precision, and fraud detection accuracy.
- Designed microservices architecture supporting multi-model orchestration, A/B testing, versioning, and secure LLM access across PayPal's ecosystem.

**AI Engineer | Accenture | Hyderabad, India | Feb 2022 – Jul 2023**

- Developed production-grade AI applications using Python, Flask, FastAPI, and Node.js services, deployed on AWS, Azure, and Kubernetes.
- Built enterprise-level RAG chatbots, recommendation engines, and segmentation models, improving personalization and decision-making for BFSI and healthcare clients.
- Created modular ML libraries for data pipelines, feature engineering, training workflows, reducing developer workload across teams.
- Integrated AI services into web platforms via REST/gRPC APIs, improving client adoption and reducing integration friction by 40%.
- Led development of a computer vision risk-detection pipeline for 10,000+ annotated industrial images, improving workplace safety incident detection accuracy.
- Established CI/CD pipelines for ML using GitHub Actions, Jenkins, reducing deployment effort and preventing regression issues.
- Deployed containerized AI systems with Docker + Kubernetes autoscaling, maintaining 99.9% uptime and stable performance under load.

## EDUCATION

Kennesaw State University - **Master of Science - MS, Computer Science AI** Kennesaw, GA
CMR College of Engineering & Technology - **Bachelor of Technology - Information Technology** Telangana, India