KARTHIK KOTA

AI Engineer · Machine Learning · GenAI & LLM Applications

Allen, TX | +1 (469) 396-2619 | kotakarthik.ai@gmail.com | LinkedIn | Portfolio | Kaggle

PROFESSIONAL SUMMARY

AI Engineer with 3+ years of experience building and deploying production-grade ML models, RAG architectures, LLM-powered applications, and computer vision systems across financial services, healthcare, and enterprise environments. Delivered **$1.2M+ in annual savings**, **40% latency reduction**, and **91% model accuracy** at scale. Expert in Python, OpenAI API, LangChain, AWS Bedrock, GCP Vertex AI, and MLflow-driven MLOps.

TECHNICAL SKILLS

Languages: Python, Java, C++, JavaScript, SQL, Bash

AI & LLMs: OpenAI API, Claude API, LangChain, Hugging Face, RAG, Prompt Engineering, AWS Bedrock, NER, Text Summarization

ML & Deep Learning: scikit-learn, XGBoost, LightGBM, TensorFlow, Keras, PyTorch, CNN, LSTM, Transformers, SHAP, ARIMA, Prophet, Reinforcement Learning

Computer Vision: OpenCV, YOLO, Detectron2, TensorRT, Transfer Learning, Object Detection, Real-Time Inference

MLOps & APIs: MLflow, FastAPI, REST APIs, FAISS, Pinecone, SageMaker Endpoints, Distributed Training, Model Parallelism, CI/CD, Model Versioning

Cloud & DevOps: AWS (EC2, S3, Lambda, SageMaker, Bedrock, IAM, CloudWatch), GCP (Vertex AI, BigQuery), Docker, GitHub Actions, Terraform

Data Engineering: Pandas, NumPy, Apache Spark, Kafka, ETL Pipelines, Matplotlib, Seaborn, Tableau, Power BI

Databases: PostgreSQL, MySQL, MongoDB, DynamoDB, Cassandra, NoSQL

Compliance: HIPAA, SOC 2, IAM Role-Based Access, Data Encryption

PROFESSIONAL EXPERIENCE

AI Engineer · **Mr. Cooper** *Feb 2025 – Present*
Dallas, TX · LangChain | RAG | AWS SageMaker | OpenAI API | MLflow | Apache Spark

- Deployed LLM-powered mortgage document intelligence using RAG, LangChain, and OpenAI API on SageMaker — **cutting review time 45%** across **500K+ annual transactions**, directly saving $800K+ in operational costs.
- Built real-time XGBoost risk scoring at **sub-100ms latency** processing **1M+ loan records monthly**; MLflow experiment tracking reduced model iteration cycles by 50%.
- Engineered Hugging Face NER pipelines improving loan application data accuracy by **32%** and eliminating **20+ manual review hours weekly**.
- Developed Apache Spark ETL ingesting **200GB+ daily** on AWS S3 at **99.9% uptime**, reducing processing time 60%; GitHub Actions CI/CD compressed deployments from 3 days to 4 hours.
- Shipped LangChain + FAISS mortgage policy Q&A achieving **87% self-resolution**, cutting escalation volume by 35%; delivered ML dashboards backing a **$2M+ efficiency initiative**.

AI/ML Engineer · **Tata Consultancy Services (TCS)** *Apr 2021 – Jun 2023*
Hyderabad, India · XGBoost | Kafka | PyTorch | Hugging Face | AWS EC2 | MLflow

- Delivered XGBoost credit risk classifier at **91% accuracy**, reducing false positives **28%** on **800K+ records**; full model lifecycle managed in MLflow with bi-weekly automated retraining via CI/CD.
- Built LSTM fraud detection pipeline on Kafka streams processing **50K+ transactions/minute** at **94% precision**, cutting fraud losses by 22% across financial portfolios.
- Created Hugging Face + OpenAI NLP pipelines for financial report summarization — **35% reduction in analyst effort** across 10+ enterprise clients.
- Engineered CNN/LSTM deep learning models (TensorFlow, Keras, PyTorch) achieving **95% accuracy** in image recognition and sequence prediction; containerized on Docker reducing latency **40%** and infra costs **22%**.
- Optimized PostgreSQL and DynamoDB queries cutting pipeline runtime **50%** on **5M+ record datasets**; integrated 10+ systems into centralized ML feature store on AWS S3/RDS.

AI ENGINEERING PROJECTS

Enterprise Financial Fraud Intelligence Platform | *AWS Bedrock · LangChain · BERT · FAISS · Pinecone · FastAPI · MLflow · SOC 2*

- Built RAG fraud system integrating BERT embeddings with FAISS + Pinecone across **1M+ daily transactions**; FastAPI on AWS EC2 at **95ms latency, 99.95% uptime — saved $1.2M annually**.
- Reduced false positive fraud alerts **24%** by combining vector similarity search with XGBoost classifiers; MLflow model registry enabled weekly zero-downtime releases.
- CloudWatch monitoring cut MTTD **52%**; SOC 2 compliance enforced via IAM controls and encrypted Pinecone namespaces serving fraud and audit teams.

AI-Powered Enterprise Knowledge Copilot | *LangChain · OpenAI · AWS Bedrock · Pinecone · FastAPI · MLflow · Apache Spark*

- Architected multi-tenant GenAI copilot over **2M+ enterprise documents** for **5K+ concurrent users**; FastAPI on AWS EC2 at **sub-150ms latency** handling **500+ simultaneous requests**.
- Automated Spark/S3 ingestion slashing knowledge base refresh from **8 hours to 20 minutes**; LangChain agents **cut human lookup time 70%**.
- MLflow experiment tracking improved answer relevance **38%**; IAM-enforced tenant isolation and encrypted Pinecone namespaces ensured enterprise-grade security.

IoT Predictive Maintenance & Financial Forecasting | *PyTorch · Apache Spark · Kafka · TensorRT · GCP Vertex AI · BigQuery · LightGBM · Tableau*

- Deployed predictive maintenance models on IoT sensor data (PyTorch + Spark Streaming + Kafka) to **GCP Vertex AI — reduced equipment downtime 18%** across production facilities.
- Implemented financial time-series forecasting with ARIMA, Prophet, XGBoost, and LightGBM; TensorRT-optimized inference pipelines delivered **sub-50ms predictions** on streaming data.
- Automated BigQuery ETL improving data consistency; Tableau dashboards visualizing asset health and portfolio performance adopted by **C-suite stakeholders** across 3 business units.

EDUCATION

Master of Science — Computer and Information Science *May 2025*
University of North Texas · Denton, TX

Bachelor of Technology — CS & Engineering (Artificial Intelligence) *May 2023*
Vellore Institute of Technology · Vellore, India

CERTIFICATIONS

- **AWS Certified Machine Learning – Associate** — *Amazon Web Services*
- **TensorFlow Developer Certificate** — *Google*
- **Deep Learning Specialization** — *Coursera – deeplearning.ai*