Pavan K

📍 Atlanta, United States | 📱 470-233-9292 | ✉ pavandsml1@gmail.com | 🔗 LinkedIn | 🌐 Portfolio | 💻 GitHub

Senior Generative AI & MLOps Engineer

Enterprise AI | LLM Engineering | RAG | AI Agents | Computer Vision | Cloud-Native ML | Secure AI Deployments

PROFESSIONAL SUMMARY

Senior AI & MLOps Engineer with 10 years of experience designing, deploying, and scaling enterprise-grade Machine Learning, Computer Vision, and Generative AI solutions in regulated environments. Experienced in building fraud detection engines, LLM-powered AI agents, Retrieval-Augmented Generation (RAG) systems, and production MLOps pipelines across AWS, Azure, and GCP. Known for converting ambiguous business requirements into secure, measurable AI outcomes (fraud loss reduction, operational efficiency, and customer experience). Proficient in AI-assisted development workflows using GitHub Copilot, Claude, Cursor, and GPT-based copilots to accelerate delivery while maintaining secure SDLC, code quality, and governance.

VALUE DELIVERED

- Delivered production AI solutions end-to-end: discovery → POC → hardened service → release → monitoring → iterative improvement.
- Implemented LLM + RAG copilots for fraud operations and document workflows to reduce investigation time and improve decision consistency.
- Built scalable, observable ML services on Kubernetes with CI/CD, IaC, and model governance (versioning, approvals, rollback).
- Developed CV-based document verification pipelines to improve onboarding/transaction integrity and reduce manual exceptions.

CORE TECHNICAL EXPERTISE

- Generative AI: GPT-4/4o, Azure OpenAI, AWS Bedrock, Claude, HuggingFace, LangChain, Embeddings, RAG, Prompting, Guardrails
- AI Agents & Orchestration: Tool-calling agents, function/API integration, workflow automation, async task pipelines
- Computer Vision: OpenCV, PyTorch/TensorFlow, CNNs, document verification, ID/receipt validation, OCR pipelines
- Machine Learning: XGBoost, Random Forest, Logistic Regression, Gradient Boosting, Time Series, Anomaly Detection, Imbalanced Learning

- MLOps: MLflow, model registry, CI/CD (GitHub Actions, Jenkins), Docker, Kubernetes, Helm, Terraform, model monitoring/drift
- Cloud & Data: AWS (SageMaker, Lambda, S3, Glue, EKS, SQS, CloudWatch), Azure (Databricks, Azure ML, OpenAI), GCP (Vertex AI, BigQuery)
- Productivity Tools: GitHub Copilot, Claude, Cursor IDE, GPT copilots for code review, test generation, docs, and refactoring

SELECTED DEPLOYMENTS & RELEASES

- Released LLM-powered case summarization and risk-scoring assistant integrated with internal APIs; improved analyst throughput and reduced repetitive manual steps.
- Deployed RAG-based knowledge assistant for policy/procedures and investigation playbooks using vector search + access controls; improved response consistency for operations teams.
- Shipped CV-based document verification service (ID / payment slip validation) with automated QA checks and monitoring dashboards; reduced manual verification workload.
- Productionized fraud detection model upgrades with automated training/evaluation pipelines, gated rollouts, and rollback strategy via CI/CD and model registry.

PROFESSIONAL EXPERIENCE

MoneyGram — Senior Generative AI & MLOps Engineer

Texas | Sep 2021 – Present

- Designed and deployed large-scale fraud detection ML systems for high-volume transactions; improved detection precision and reduced false positives through feature engineering and ensemble modeling.
- Built domain-specific RAG architectures integrating LangChain, FAISS/Pinecone, and GPT models to analyze financial documents and support investigation workflows.
- Developed AI agents using GPT-4/Azure OpenAI for automated case summarization, transaction classification, and fraud risk scoring; integrated tool-calling with internal APIs for end-to-end workflow automation.
- Implemented LoRA/PEFT fine-tuning for domain adaptation and cost-efficient inference; established evaluation checks (hallucination tests, retrieval quality, regression suites).
- Developed Computer Vision pipelines using OpenCV and CNN-based models for document verification, ID validation, and payment slip authentication; implemented quality checks and escalation paths.
- Established secure prompt engineering patterns and guardrails: PII redaction, policy filters, deterministic templates for high-risk decisions, and audit-friendly logging.
- Used GitHub Copilot, Claude, Cursor, and GPT copilots to accelerate development (unit tests, refactors, documentation, code review notes) while adhering to secure SDLC and code scanning policies.

- Implemented MLflow for experiment tracking, model versioning, and reproducibility; standardized model packaging and promotion workflows.
- Built Kubernetes-based scalable inference systems (Docker + Helm) with autoscaling, health checks, and blue/green deployment patterns.
- Automated infrastructure provisioning using Terraform across AWS environments (networking, IAM, compute, storage, observability) and established release pipelines.
- Implemented monitoring for AI services and pipelines using Prometheus/Grafana and CloudWatch/Splunk; added alerting for latency, errors, and model performance drift.
- Partnered with product and operations stakeholders to define success metrics and prioritize releases; delivered measurable operational efficiency improvements.

Autodesk — MLOps Engineer

San Francisco, CA | Jul 2019 – Aug 2021

- Designed scalable ML infrastructure and automated ML deployment pipelines using Azure DevOps and Jenkins; improved release consistency and reduced manual deployment effort.
- Built and supported production recommender systems deployed on Kubernetes; implemented CI/CD gates (tests, security scans, quality checks) and rollback procedures.
- Containerized services with Docker and managed lifecycle via Kubernetes (resource tuning, autoscaling, monitoring, incident response).
- Implemented model lifecycle management using Azure ML and MLflow (tracking, registry, promotions); partnered with data scientists to productionize models.
- Improved system reliability with infrastructure automation and observability dashboards; reduced downtime through proactive monitoring and alerting.

Cummins — Machine Learning Engineer

Maharashtra, India | Oct 2016 – Mar 2019

- Developed predictive ML models (XGBoost, clustering) for customer analytics and segmentation; improved targeting and business insight quality through robust validation.
- Built Spark/Hadoop data pipelines on AWS for large-scale processing; implemented feature pipelines and batch scoring jobs.
- Designed serverless data/ML workflows using AWS Lambda, S3, and CloudWatch; automated recurring jobs and operational tasks.
- Established CI/CD patterns for ML deployments and introduced monitoring checks to detect data quality issues and model drift.
- Partnered with stakeholders to translate operational questions into measurable ML outputs and dashboards.

Bluecoat Systems — Data Scientist

Bangalore, India | Aug 2014 – Sep 2016

- Built fraud detection and revenue forecasting models using ensemble methods; improved model quality through hyperparameter tuning and cross-validation.
- Implemented time-series forecasting (ARIMA/Exponential Smoothing) for financial trend analysis and demand estimation.
- Created Tableau dashboards and executive reporting to communicate model outputs and KPIs to business stakeholders.
- Performed A/B testing and statistical analysis to validate impact and guide rollout decisions.

SECURITY, COMPLIANCE & RELIABILITY

- Applied secure SDLC practices: code reviews, dependency scanning, secrets handling, least-privilege access patterns, and audit-ready logging.
- Designed AI solutions with data protection in mind: PII-aware prompting, access-controlled retrieval, and safe output policies for regulated workflows.
- Operational readiness: runbooks, monitoring/alerts, on-call support practices, and post-release validation to sustain production performance.

EDUCATION

Bachelor of Information Technology — 2014