JAYAKRISHNA KONDA

Data Scientist · ML Engineer · GenAI Engineer

+1 (667) 910-1092 | jayakrishnakonda10@gmail.com | linkedin.com/in/jaya-krishna-konda | github.com/jay739 | jay739.dev

PROFESSIONAL SUMMARY

ML Engineer and Data Scientist with 4+ years building production ML systems across NLP, computer vision, and GenAI. Experienced deploying scalable ML pipelines, RAG systems, and LLM evaluation frameworks on AWS. Delivered ML systems improving workflow efficiency by 30–40% across NLP and document-processing pipelines. Builder and operator of "Batcave" — a 56-container, hybrid-cloud AI platform running 36+ services across 177GB of production data. Proficient in Python, SQL, PySpark, PyTorch, TensorFlow, LangChain, Docker, Kubernetes, and AWS.

TECHNICAL SKILLS

GenAI / LLM Systems
- **RAG pipelines:** LangChain, LangGraph, LlamaIndex, FAISS, Vector Databases, Semantic Search, Embeddings
- **Fine-tuning:** LoRA, QLoRA, Instruction Tuning
- **Local inference:** Ollama, LLaMA, Mistral, Phi
- **Evaluation:** BLEU, ROUGE, BERTScore, Human-in-the-Loop, A/B Testing
- **Speech & Audio:** TTS pipelines, speech synthesis, OCR

Machine Learning
- **Frameworks:** PyTorch, TensorFlow, Scikit-learn, XGBoost, Keras
- **Architectures:** CNNs, Transformers, RNNs/LSTMs
- **Computer Vision:** YOLOv8, AllCNN, Sentinel-2/MODIS Satellite Imagery, NDVI/NBR
- **Edge AI:** TensorFlow Lite, ESP32
- **Model Rigor:** Cross-validation, Hyperparameter Tuning, Ablation Studies, SHAP, Model Calibration, Bias Evaluation

Programming
- **Core:** Python, SQL, PySpark, Scala, C++, Node.js, Bash/Shell
- **Data:** Pandas, NumPy, SciPy

Data Engineering & MLOps
- **Big Data:** Apache Spark (PySpark), Hadoop, ETL Pipelines, Feature Engineering
- **Databases:** PostgreSQL, MariaDB, Redis (multi-instance), SQLite
- **MLOps:** MLflow, Weights & Biases, CI/CD (GitHub Actions, Jenkins), Model Versioning, Drift Detection
- **Experimentation:** A/B Testing, Causal Analysis, Statistical Significance Testing

Infrastructure & Cloud
- **Cloud:** AWS (SageMaker, Lambda, S3, Step Functions, DynamoDB, Redshift), Oracle Cloud (hybrid caching)
- **Containers:** Docker, Kubernetes, Portainer, Docker Compose
- **Observability:** Netdata, Telegram alerting, structured logging
- **Security:** Authentik SSO/2FA, Tailscale VPN, Zero-trust access, CWE/CVE vulnerability scanning
- **APIs & UI:** FastAPI, RESTful APIs, Next.js, TypeScript, Tailwind CSS

PROFESSIONAL EXPERIENCE

Data Scientist — GenAI & ML | **Enigma Technologies** · Remote, USA *Jun 2025 – Present*
- **GenAI & RAG:** Designed and deployed production RAG pipelines and LLM fine-tuning workflows to automate large-scale document classification and entity extraction — reducing analyst review time by 30%+ and accelerating downstream reporting.
- **LLM Pipelines:** Designed multi-step LLM-based document intelligence pipelines using LangChain and LangGraph with tool orchestration and task decomposition, automating multi-stage business intelligence workflows consistently and at scale.
- **ML Modeling:** Built regression, classification, and clustering models for customer segmentation and risk scoring on 10M+ records with rigorous feature engineering and statistical validation — improving scoring precision by 20% in production.
- **Model Rigor:** Applied cross-validation, hyperparameter tuning (Optuna), ablation studies, and SHAP-based feature importance analysis to ensure model interpretability and production readiness before every deployment.
- **Collaboration:** Partnered with product and data teams to translate business requirements into ML model features and evaluation metrics; presented SHAP-based model explanations to non-technical stakeholders for informed decision-making.
- **MLOps:** Delivered cloud-native ML platforms on AWS using Docker, Kubernetes, GitHub Actions CI/CD, MLflow versioning, and drift detection — sustaining reliability SLAs while reducing compute costs through autoscaling.

AI/ML Programming Intern | **R/SEEK — UMBC** · Baltimore, MD *Jan 2025 – Jun 2025*
- **Computer Vision:** Developed real-time object recognition and motion tracking for autonomous RC car navigation using YOLOv8, achieving 95% detection accuracy across dynamic test environments.
- **Edge AI:** Optimized and deployed TensorFlow Lite models on resource-constrained ESP32 microcontrollers for real-time autonomous navigation with Aruco marker-based positioning — no cloud dependency.

Machine Learning Engineer | **Cognizant** · India *Jan 2020 – Jul 2023*

- **NLP Systems:** Delivered production NLP pipelines for contract and document intelligence — context extraction, semantic search, and text classification across 10M+ records — reducing manual legal and compliance review effort by 35% while meeting SLA targets.
- **ML Pipelines:** Designed end-to-end ML pipelines (training, evaluation, batch and real-time inference) using Python, SQL, Scikit-learn, PySpark, TensorFlow, and PyTorch; reduced inference latency from 450ms to under 90ms through batching and model optimization.
- **MLOps & CI/CD:** Implemented enterprise MLOps — Jenkins CI/CD automation, automated retraining, model versioning, performance monitoring, and drift alerting — cutting model release cycles by 30% and reducing production incidents.
- **Cloud Inference:** Engineered low-latency inference services on AWS Lambda and SageMaker with event-driven microservices, reducing p99 latency by 40% and cutting monthly infrastructure cost by 25%.
- **Distributed Data:** Optimized fault-tolerant PySpark workflows for NLP feature extraction and batch inference; reliably processed 1M+ text records/day with consistent throughput and high availability.

Software Engineering Intern | **Infosys** · India *Jan 2019 – May 2019*

- **Full-Stack & Data:** Built a real-time movie database application using Node.js and MongoDB; applied data analysis to evaluate user interaction patterns and rating trends, surfacing actionable engagement insights.

FEATURED PROJECTS

AI Platform Infrastructure — "Batcave" Personal ML Cloud *2023 – Present*

CPU-Optimized private AI server — designed, built, and operated solo on a Beelink mini PC. 56 production containers, 36+ services, 99.9% uptime. Documented in a 3-part engineering blog series at jay739.dev/blog.

- **LLM & RAG Infrastructure:** Deployed multiple LLMs (LLaMA, Mistral, Phi via Ollama) with FAISS-backed RAG pipelines for private semantic Q&A; 177GB multi-database data layer (PostgreSQL, MariaDB, Redis, SQLite) supporting media processing, RAG retrieval, and ML inference services. Hybrid-cloud caching via Oracle Cloud reduced retrieval latency by 60%.
- **Security & Access:** Authentik SSO with 2FA across all 36+ services, Tailscale mesh VPN for encrypted remote access, and Nginx reverse proxy with SSL/HTTPS — zero-trust access enforced at network and application layers.
- **Data Automation:** Built Python pipelines processing 10,000+ audio files with metadata extraction, fuzzy deduplication (95% accuracy), and 8-endpoint REST API orchestration — reducing manual configuration effort by 90%.
- **Observability & CI/CD:** Netdata monitoring with Telegram bot alerting for anomaly detection; GitHub Actions CI/CD on self-hosted VPS with zero-downtime rollouts and automated Watchtower container update polling.

Infrastructure Security Audit & Compliance Scanner · Python · Docker SDK · CWE/CVE · 2026 *2026*

- **Vulnerability Assessment:** Automated security assessment of 56-container production infrastructure, identifying 27 hardcoded credential exposures (CWE-798) through Docker SDK and grep-based scanning across all Compose files, systemd units, and network configurations.
- **Quantitative Scoring:** Developed an 8-category weighted security scoring methodology; calculated baseline security posture of 37.5% and generated a prioritized remediation roadmap targeting 92% — a 54.5-point quantified improvement with effort estimates and measurable score deltas per fix.

Wildfire Detection via Multispectral Satellite Imagery · CNN · AllCNN · Sentinel-2 · NDVI/NBR · Python *2024*

- **Model Development:** Built AllCNN-based wildfire classifier on Sentinel-2 satellite imagery with MODIS labels; performed ablation studies and k-fold cross-validation — achieving 91% accuracy and 88% F1-score on held-out test data.
- **Feature Engineering:** Processed 2,400+ multispectral image patches using NDVI and NBR spectral indices for burn severity quantification; domain-informed feature selection improved model signal-to-noise ratio significantly.

RAG-Powered Podcast Generator · RAG · LangChain · Ollama · OCR · TTS · Fine-Tuning *2024*

- **Document AI Pipeline:** LLM-powered PDF processor achieving 98% OCR accuracy on 200+ page documents; transformer NER for 90% character identification across literary genres — full write-up at jay739.dev/blog.
- **Optimization:** Fine-tuned LLMs for narrative tone consistency; optimized LangChain + Ollama pipeline achieving 4x throughput improvement — deployed as a containerized multimodal AI service.

Financial Risk & Sentiment Analysis Suite · FinBERT · Monte Carlo · CAPM · Python *2023*

- **NLP Sentiment:** FinBERT sentiment classifier achieving 87% accuracy across 10,000+ financial news articles; automated text analytics pipeline for data-driven investment decision workflows.
- **Quantitative Risk:** Ran 1,000+ Monte Carlo simulations for portfolio VaR estimation at 95% confidence; automated CAPM alpha/beta and Sharpe/Treynor computation for 50+ stocks — reducing analyst effort by 80%.

EDUCATION

M.S., Data Science — University of Maryland, Baltimore County (UMBC) *Aug 2023 – Jun 2025*

GPA: 3.91 / 4.0

B.E., Computer Science & Engineering — Sreenidhi Institute of Science and Technology *Aug 2015 – Aug 2019*