AKHIL BONTHINAYANIVARI

(470) 940-8044 | akhilme008@gmail.com | akhilportfolio-omega.vercel.app | Tampa, FL (Open to relocation)

Lead AI/ML Engineer | Production agentic AI, LLM applications, computer vision and MLOps

Professional Summary:

- I started in backend Python and moved into AI/ML engineering. Four years building agentic workflows, LLM-powered applications and production AI systems from prototype through deployment, with full lifecycle ownership in cloud environments.
- I run two AI systems in production right now, an agentic LLM assistant with a RAG pipeline and a computer vision system on 3D mesh data. Both are owned end to end, from architecture and deployment through evaluation, observability and incident response.
- Python is the core. LangGraph and LangChain for orchestration, RAG for retrieval, Amazon Bedrock and SageMaker for foundation models. I build evaluation pipelines, CI/CD and observability into everything I ship, and I treat production AI as far more than model building.

Education:

M.S. in Artificial Intelligence, University of North Texas *2023 to 2025*

B.Tech in Computer Science (AI and ML Specialization), Madanapalle Institute of Technology and Science *2020 to 2023*

Technical Skills:

Machine Learning / AI	Large Language Models (LLMs), Generative AI, AI Agents, LangChain, LangGraph, AutoGen, CrewAI, RAG, MCP servers, PyTorch, TensorFlow, scikit-learn, XGBoost, MLflow, MLOps, NLP, prompt engineering, Hugging Face, LoRA/QLoRA, PEFT, CLIP, SAM, OpenCV, Whisper (STT), Kokoro TTS, AWS Polly, RAGAS, DeepEval, NeMo Guardrails
Programming / Infra	Python, FastAPI, SQL, boto3, Git, CI/CD, microservices, REST APIs, Linux/Unix
Data Pipelines / Streaming	PySpark, Apache Kafka, Airflow, dbt, Spark, batch and streaming pipelines
Cloud / Deployment	Docker, Kubernetes, Terraform, AWS (SageMaker, Bedrock, Lambda, EC2, S3, EMR, Glue, Athena, Polly, ECS), Azure (Databricks, AI Foundry, Azure ML), Linux
Databases	PostgreSQL, MySQL, MongoDB, Redis, Elasticsearch, Snowflake, Redshift, Pinecone
Monitoring / Tooling	Grafana, Prometheus, Tableau, Power BI, GitHub Actions, Claude Code, Cursor, Windsurf, GitHub Copilot

Professional Experience:

AI/ML Engineer, DriverAI, Tampa FL (Remote) *January 2024 to Present*

- Leading end-to-end development of Sparke, an agentic AI assistant for DriverAI enterprise customers. LangGraph and LangChain for multi-agent orchestration, RAG with embedding-based retrieval, Amazon Bedrock as the foundation model layer, with architecture through production monitoring all mine.
- Also running a computer vision pipeline for product detection on 3D mesh data using CLIP, SAM, OpenCV and GPU-accelerated CUDA via COLMAP. Both systems include structured evaluation pipelines using RAGAS and DeepEval measuring accuracy, reliability and failure modes before each release.
- Built data pipelines using PySpark, Kafka and Airflow on AWS. Data contracts and schema standards defined with the data team, structured and unstructured data ingested, validated and made AI-ready.

- Owned monitoring and incident response, with drift detection and anomaly flagging via Grafana and Prometheus, MLflow for model versioning and experiment tracking, root cause analysis with documented runbooks and model cards.
- Built CI/CD pipelines for model deployment using Docker, Kubernetes and AWS ECS with SageMaker for fine-tuning with LoRA and QLoRA. Containerized workloads with full observability and automated releases.

Machine Learning Engineer, PwC, Bangalore India *August 2022 to August 2023*

- Built data pipelines across a few client engagements. Ingestion, validation, prod support on both structured and unstructured data. Trained and deployed NLP models on Google Vertex AI for text classification and model lifecycle management. Used Redis for caching and session management and Elasticsearch for search and log analytics across client data platforms.
- Worked with data scientists and app teams in close collaboration to keep training data in shape. Classification, regression, clustering and predictive modeling varied across clients.
- Caught data quality issues before they hit training. Tightened test coverage and the downstream incident count dropped.

ML Engineer, Hosting Raja, Bangalore India *June 2019 to May 2020*

- Started as an intern and moved into an ML Engineer role after 3 months, then took ownership of backend modules and automation scripts used in a live product.
- Built backend features, fixed bugs, handled SQL workflows and supported API integrations in production. That grounding in software engineering still shows up in how I approach ML systems.
- Worked in a shared codebase with code reviews and version control. That's where the discipline for writing clean Python came from.

Projects:

Project: Drug-Drug Interaction Discovery *Independent to 2025*

GNN over a biomedical knowledge graph to flag drug-drug interactions. Claude API surfaces the ranked scores as human-readable output.

Technologies: PyTorch, GNN, Knowledge Graph, Claude API, Python

Project: Signal Desk Agent *Independent to 2025*

Agentic research tool that turns a question into a structured brief. It searches, retrieves and formats relevant information into a usable report.

Technologies: TypeScript, LLM Agents, Tool Use, Structured Output

Project: Reddit Data Engineering Pipeline *Independent to 2025*

End-to-end data pipeline: Reddit API into S3, catalogued with Glue, transformed in Athena, landed in Redshift. Airflow and Celery orchestrating, Tableau for visualization.

Technologies: Airflow, PySpark, AWS Glue, Athena, S3, Redshift, Celery, Tableau

Project: Real-Time Voting Data Pipeline *Independent to 2025*

Kafka and PySpark pipeline aggregating election votes in real time with anomaly detection. Dashboard latency from ingest to display under one second.

Technologies: Kafka, PySpark Streaming, Python, PostgreSQL, Plotly