

KARTHIKEYA CHEKURI

AI Engineer | Generative AI, RAG & LLM Evaluation

Dallas, TX | (682) 372-6534 | ckvarma579@gmail.com | linkedin.com/in/karthikeya-varma-350635217

PROFESSIONAL SUMMARY

AI Engineer with 4 years of experience taking machine learning and generative AI from prototype to production across healthcare, manufacturing and identity security. At CVS Health, builds Retrieval-Augmented Generation (RAG) assistants and LangGraph agentic workflows on Azure OpenAI and Amazon Bedrock, including an internal healthcare assistant that raised citation-grounding accuracy from 72% to 91% and cut unsupported answers from 11% to under 4%. Focused on grounded answers, rigorous LLM evaluation and safe human escalation, and works closely with business stakeholders to turn their requirements into production AI.

CORE SKILLS

Generative AI and Agents: Large Language Models (LLMs), Azure OpenAI, Amazon Bedrock, Claude, GPT, LLaMA, Prompt Engineering, Structured Outputs, Tool Calling, LangGraph, Agentic AI Workflows, Human-in-the-Loop Escalation

RAG and Retrieval: Retrieval-Augmented Generation (RAG), Embeddings, Vector Databases, Azure AI Search, Bedrock Knowledge Bases, Amazon Titan Embeddings, OpenSearch Serverless, FAISS, BM25 Hybrid Search, Metadata Filtering, Cross-Encoder Reranking

LLM Evaluation and Responsible AI: RAGAS, DeepEval, Golden Datasets, LLM-as-Judge, Groundedness and Faithfulness, Hallucination Mitigation, Bedrock Guardrails, PII/PHI Protection, Role-Based Access Control (RBAC), Audit Trails

Machine Learning: PyTorch, TensorFlow, scikit-learn, XGBoost, SHAP, Probability Calibration, Time-Series Modeling, Anomaly Detection, Recommendation Systems, Learning-to-Rank

Data Engineering: Databricks, PySpark, Spark SQL, Spark Structured Streaming, Delta Lake, Kafka, Feature Engineering

Cloud and MLOps: AWS (S3, Lambda, API Gateway, DynamoDB, EKS, ECR, End User Messaging), Azure (Azure ML, AKS), MLflow, Docker, Kubernetes, Azure DevOps, CI/CD, Git

Languages and APIs: Python, SQL, Bash, FastAPI, Flask, REST APIs, CRM API Integration, PostgreSQL, SQL Server, Redis

PROFESSIONAL EXPERIENCE

CVS Health

Dallas, TX

AI Engineer

Jun 2025 to Present

Internal GenAI Healthcare Assistant: A generative AI assistant built on Retrieval-Augmented Generation (RAG). CVS employees ask natural-language questions about policy, coverage, claims and member-service documentation and get answers grounded in approved sources.

- Improved hybrid retrieval on Azure AI Search by combining BM25 keyword search with vector embeddings, metadata filters for plan, state and policy version, and cross-encoder reranking, raising citation-grounding accuracy from 72% to 91%.
- Traced answers that sounded right but weren't backed by the documents to either retrieval or generation issues, then tightened prompts and added answer validation, cutting unsupported large language model (LLM) responses from 11% to under 4%.
- Built agentic workflows in LangGraph that check for enough evidence before answering and route low-confidence or sensitive questions to human review.
- Created an evaluation framework with golden datasets, RAGAS and DeepEval to measure groundedness and faithfulness every time retrieval or prompts changed.
- Built Python and FastAPI services on Azure OpenAI, deployed with Docker and Kubernetes through Azure DevOps CI/CD.
- Enforced role-based access control (RBAC) in retrieval, along with PII/PHI masking and audit trails, so employees only see information their role allows.

Aetna AI Outreach Assistant: An LLM-powered assistant that texts consented Aetna insurance prospects through an SMS API. It answers their plan questions from approved information and hands complex conversations to human representatives.

- Built an AI-powered customer outreach workflow for Aetna insurance products, automating early-stage customer engagement and reducing manual outreach effort by ~35% through LLM-driven conversations, RAG-grounded responses and human-in-the-loop escalation.
- Built the RAG pipeline on Amazon Bedrock Knowledge Bases, indexing approved Aetna insurance documents from Amazon S3 with Titan embeddings and OpenSearch Serverless, so Claude on Bedrock answers only from retrieved plan information.
- Wrote Python integrations with the CRM API for customer context and with AWS End User Messaging SMS through API Gateway and Lambda, letting the LLM hold personalized two-way text conversations.
- Designed the LangGraph workflow up front so routine replies go to the LLM, plan questions go through RAG, and complex or uncertain conversations move to a human representative instead of getting a hallucinated answer.

- Added plan and state filtering, reranking and Bedrock Guardrails, plus a validation step that escalates to a human whenever the retrieved evidence can't support an answer.
- Stored conversation state in DynamoDB and kept opt-outs and other compliance actions on deterministic rules for consistent, auditable handling.

Harley-Davidson

Milwaukee, WI (Remote)

AI/ML Engineer

Sep 2023 to Jun 2025

Predictive Maintenance and Quality Intelligence: A machine learning platform that flags early equipment failure and quality risk from live plant sensor data so teams can act before downtime or defects.

- Raised failure recall from 74% to 89% with XGBoost and PyTorch models trained on plant sensor data, while keeping false alerts in check.
- Created rolling-window features from sensor streams and tuned alert thresholds by weighing the cost of a missed failure against a false alarm.
- Trained quality-risk models on inspection and production data to spot runs likely to produce defects, giving the quality team time to act.
- Cut failure detection time from 15 minutes to under 3 by moving anomaly scoring onto Kafka and Spark Structured Streaming, with alerts ranked by severity.
- Pulled PLC telemetry, maintenance history and warranty records together into one equipment-health score used across the plant floor.
- Set up reproducible training pipelines on Databricks and Delta Lake, using point-in-time joins to keep future data out of training.
- Wrote FastAPI services for batch and asynchronous predictions, with autoscaling and versioned endpoints.
- Shortened model releases from several days to hours by automating CI/CD with MLflow on AWS EKS, including canary rollback.
- Gave maintenance teams the top SHAP failure drivers behind each alert, and added drift monitoring with scheduled retraining.

Saviynt

Bengaluru, INDIA

Machine Learning Engineer

Oct 2022 to Jul 2023

Identity Risk Analytics and Access Recommendations: A machine learning platform that helps identity and access management teams spot risky access and enforce least-privilege governance.

- Trained access-risk models with XGBoost and PyTorch, using class balancing and probability calibration so scores stayed reliable on heavily imbalanced access data.
- Attached SHAP explanations to every flagged access, so reviewers could see why a permission looked risky instead of trusting a bare score.
- Designed a permission recommendation engine that blended collaborative filtering with XGBoost learning-to-rank to suggest which entitlements each user should hold.
- Turned user, role, entitlement and login history into behavioral features for risk scoring and anomaly detection.
- Automated batch and near-real-time pipelines in Python, PySpark and Kafka, with point-in-time features that kept training reproducible in MLflow.
- Shipped the models as FastAPI services on Kubernetes inside identity governance workflows, and retrained them on reviewer decisions and false-positive analysis.

EDUCATION

Master of Science in Data Science, University of Texas at Arlington

Arlington, TX (2023-2024)

CERTIFICATIONS

Claude Certified Architect – Foundations, Anthropic, 2026

AWS Certified AI Practitioner, Amazon Web Services, 2026