

NIKHILNATH K V

SENIOR AI / ML ENGINEER

PROFESSIONAL SUMMARY

Senior AI/ML Engineer with 7+ years designing, deploying, and owning large-scale AI systems in production — Generative AI, agentic and multi-agent architectures, RAG and semantic retrieval, NLP, computer vision, and classical ML. Delivered AI solutions across financial services, oil & gas, education, and audit & compliance, working directly with global business, IT, and data-engineering stakeholders to turn ambiguous problems into scalable, governed systems on Azure and Databricks. Experienced leading projects end to end and mentoring junior engineers.

EXPERIENCE

Senior Associate, AI/ML Engineering, 09/2022 to 07/2026

EY - Kochi, Kerala

- Own the full delivery lifecycle for enterprise **ML, Generative AI, and agentic AI platforms** — data strategy, model/LLM selection, architecture, development, deployment, and post-production optimization — across financial services, energy, education, and compliance engagements.
- Architect AI systems integrating with enterprise IT through **API-first microservices, cloud data platforms, agentic workflows, and downstream business applications**, designed for scale, fault tolerance, security, and traceability.
- Drive **agentic SDLC/PDLC adoption using 8090 Software Factory and AI coding agents such as Droid**, translating business intent and product requirements into technical blueprints, structured work orders, implementation tasks, code, tests, and production-ready software.
- Use **8090** to establish a context-driven development workflow connecting requirements, architecture, implementation, testing, and feedback, enabling AI agents and engineering teams to work from a shared and traceable source of truth across the product lifecycle.
- Leverage **Droid as an AI coding agent** for agent-assisted planning, implementation, codebase analysis, testing, debugging, and iterative development, with human review and engineering quality gates throughout the delivery process.
- Design and optimize production **RAG and agentic AI pipelines** — embeddings, vector databases, hybrid retrieval, LLM orchestration, tool integration, and contextual workflows — with content-level deduplication and incremental synchronization for high-relevance, low-latency responses.
- Select, fine-tune, prompt-engineer, and evaluate **ML and LLM models** for domain-specific use cases, benchmarking accuracy, latency, cost, safety, and reliability to establish production readiness.
- Establish **MLOps and AI lifecycle management** through experiment tracking, model/version management, CI/CD, automated validation, deployment workflows, observability, and post-deployment monitoring.
- Implement **ML/GenAI evaluation and monitoring frameworks** covering model quality, data quality, drift, latency, cost, and operational health, enabling continuous feedback and optimization of production AI systems.
- Build governed ETL and streaming pipelines with data-engineering teams, enforcing **schema tracking, lineage, reproducibility, and auditability** for production AI workloads.

Associate, AI/ML Engineering, 06/2019 to 09/2022

EY - Kochi, Kerala

- Designed AI-driven automation for audit and compliance workflows, achieving ~70% efficiency gain.
- Built document intelligence pipelines with Azure Form Recognizer and NLP, cutting extraction time by 60%.
- Designed AI-driven automation for audit and compliance workflows, improving process efficiency by ~70%.
- Delivered cloud-native ML services on Azure, enabling scalable API-based consumption for internal applications.

Associate, 06/2018 to 05/2019

SpiaTech - Kochi, India

- Worked on a **time-series forecasting project** to predict solar energy generation from photovoltaic (PV) cells deployed at the company's location in the **UAE**.
- Collected, cleaned, and preprocessed historical solar-energy generation data for analysis and forecasting.
- Performed **exploratory data analysis (EDA)** to identify trends, seasonality, temporal patterns, and variability in solar power generation.
- Experimented with multiple forecasting and machine-learning algorithms, including **ARIMA, SARIMA, Exponential Smoothing, Random Forest, XGBoost, and LSTM** models.

SELECTED PROJECTS

Financial Document Intelligence & Risk Screening Platform

Python · Azure OpenAI · NLP · Azure AI Document Intelligence · FastAPI · SQL · MLflow

- Architected an **AI-powered financial document intelligence platform** processing statements, contracts, and counterparty documentation to extract entities, obligations, clauses, and risk signals into a structured, queryable knowledge store.
- Orchestrated **rules, ML classifiers, and LLM reasoning** to identify exceptions and high-risk items, reducing manual review effort and improving consistency across analyst decisions.
- Built an **explainable AI decision layer** with confidence scoring, human-in-the-loop approvals, and end-to-end audit trails, enabling traceable and compliance-ready risk assessments.

Enterprise Multi-Agent HR Assistant

LangGraph · Azure OpenAI · Qdrant · FastAPI · Redis · AKS

- Architected a **production-grade multi-agent enterprise assistant** using LangGraph to route employee queries across specialized agents, tools, and retrieval domains for policies, benefits, and HR processes.
- Implemented **grounded RAG with citation-backed responses**, role-based document access, confidence thresholds, and human escalation to improve factual reliability and enterprise governance.
- Deployed agent services as **asynchronous microservices on AKS**, with conversational memory, token-budget controls, and agent-level evaluation traces for monitoring quality, latency, and LLM usage.

Digital Operations Agent — Databricks Lakehouse Platform

Azure Databricks · PySpark · Delta Lake · ADLS Gen2 · Structured Streaming · Lakeflow Declarative Pipelines

- Designed a **production-grade Bronze/Silver/Gold/Ops lakehouse architecture** for document intelligence workloads using Delta Lake and Lakeflow Declarative Pipelines.
- Engineered **incremental streaming ingestion with Auto Loader** from ADLS Gen2 for binary documents and JSON lifecycle events, incorporating schema evolution, validation, and resilient processing.
- Built an **event-driven document lifecycle framework** supporting UPDATE/DELETE operations with cascading Delta-table changes, Change Data Feed (CDF), targeted deletes, and fully auditable processing.
- Developed distributed extraction and re-ingestion workflows for **PDF, DOCX, PPTX, and XLSX**, with secure ADLS access, operational observability, and Lakeview dashboards.

Educational AI Golden Source Platform

Python · FastAPI · LangGraph · Azure OpenAI · Qdrant · AKS · Docker · Helm

- Led the **architecture of an enterprise Generative AI platform** for automated educational-content generation from multimodal source material using LangGraph-based agentic workflows.
- Engineered a **scalable microservices platform processing 10K+ documents/day across 12+ formats**, spanning ingestion, document parsing, embeddings, vector indexing, speech-to-text, retrieval, and content generation.
- Built **semantic retrieval and knowledge-quality pipelines** using Qdrant, zero-shot retrieval, and similarity-graph deduplication to identify near-duplicate content and improve source quality.
- Designed a **multi-LLM abstraction layer** with centralized token governance and rate limiting, reducing manual content-creation effort by **50%+** and improving retrieval relevance by **~35%**.

CONTACT

Address: Bengaluru, India 560100
Phone: +91 77364 82642
Email: nikhilnathkv@gmail.com
WWW: [linkedin.com/in/nikhilnathkv](https://www.linkedin.com/in/nikhilnathkv)

SKILLS

Machine Learning & AI: Machine Learning, Deep Learning, NLP, Generative AI, Retrieval-Augmented Generation (RAG), Embeddings, Semantic Search

Deep Learning & Frameworks: PyTorch, TensorFlow, Keras, scikit-learn(classification, clustering, dimensionality reduction, regression)

Backend & Orchestration: Python, FastAPI, SQL, Celery, Redis, Async Microservices, Workflow Orchestration Patterns

Cloud & Deployment: Azure OpenAI, Azure Data Lake Storage, Azure Cognitive Services, Docker, Kubernetes (AKS), Helm

EDUCATION

M.Sc., Computer Science, Data Analytics, 01/2018
IIITM-K - Trivandrum

B.Tech., Computer Science, 01/2016
SNGIST - Kochi

ADDITIONAL CONTRIBUTIONS

- NLP classification pipelines for healthcare document processing — 85–90% accuracy across multiple categories.
 - Reinforcement-learning experimentation for sequential decision and optimisation problems, evaluating feasibility ahead of production adoption.
 - Consistent delivery of production AI systems to US and EU stakeholders in 2–4 week agile cycles.
-

AWARDS

EY Hackfest Winner — 2021, 2024, Spot Award — 2019, 2020, 2021, 2022, 2024, Extra Miler Award, EY — 2020
