NAVEEN YALAKALA

Forward Deployed Engineer | AI Engineer — Agentic AI in Production, Generative AI, Machine Learning

New Jersey, United States | (201) 736-6669 | yalakalanaveen789@gmail.com | linkedin.com/in/naveen-y-b47090122

Portfolio: https://muse.ai/s/naveen-yalakala-profile-ej6qxx6bc2pxs/CTil50l4R99DwRSZnAnDwOznm2jJ31g_-BPWaH0P9bg

Professional Summary

Forward Deployed Engineer at ADP — architect of the production AI Search and AI Assist platforms behind payroll workflows for thousands of enterprise users; hybrid retrieval (OpenSearch vector + keyword) improved search precision by 35%.

LangGraph multi-agent workflows route Payroll, Tax, Benefits, and HR requests in production at sub-second latency, held to 99.9% uptime by Airflow evaluation and monitoring pipelines fed by production feedback.

At DBS Bank, built RAG and document-intelligence systems over millions of financial records — NLP/OCR automation cut manual document effort and improved extraction accuracy, with hybrid-search retrieval serving risk modeling and analyst workflows.

Professional Experience

ADP — Forward Deployed Engineer, New Jersey, United States | 07/2025 – Present

- Search precision improved 35% and enterprise knowledge-discovery relevance raised — built scalable hybrid retrieval (OpenSearch vector + keyword search) at the core of production AI Search.
- Payroll workflows for thousands of enterprise users — led architecture and end-to-end development of enterprise-scale AI/ML systems, including the production AI Search and AI Assist platforms.
- Manual triage replaced by automated orchestration — designed multi-step, tool-oriented agentic AI workflows with LangGraph routing requests across Payroll, Tax, Benefits, and HR domains.
- Sub-second response latency for conversational AI use cases — designed production-ready ML/AI microservices with FastAPI, Docker, Kubernetes, and AWS.
- 99.9% uptime maintained across releases — supported testing, deployment, monitoring, and rollback readiness for production AI services on Azure Kubernetes clusters.
- Error rates reduced through iterative model improvement — built Airflow-based evaluation and monitoring pipelines (relevance, reliability, latency, logging, feedback signals).
- Quality held across model and prompt changes — expanded LLM evaluation with structured quality checks, production feedback, and relevance/reliability metrics with regression monitoring.
- Latency lowered while relevance, reliability, and consistency improved — tuned retrieval behavior, prompts, context management, and caching in Airflow-managed pipelines.
- Compliance, traceability, and data protection enforced — built secure enterprise AI applications with audit logging, guardrails, and role-based access controls (RBAC).
- Ambiguous requirements translated into scalable production AI capabilities — partnered with product managers, engineers, and stakeholders to map high-friction workflows.

DBS Bank — Senior Data Scientist, Hyderabad, India | 02/2022 – 04/2025

- Faster analyst insight generation — developed end-to-end enterprise ML/AI systems for financial intelligence, document understanding, search, and conversational applications (Python, PyTorch, MLOps).
- Fraud-detection model accuracy increased — developed and evaluated predictive and classification models in Python with PyTorch, applying feature engineering and statistical analysis.
- Millions of financial records delivered as timely analytics for risk modeling — designed scalable data pipelines (Python, SQL, Spark, PySpark).
- Manual document effort reduced and extraction accuracy improved — automated document processing with NLP and OCR.
- Production inference latency improved — GPU optimization, efficient model serving, and performance tuning.
- Knowledge retrieval and response relevance improved — built RAG solutions with complex architectures using vector databases and hybrid search.

Infosys (Client: Walmart) — Data Scientist, Hyderabad, India | 08/2018 – 02/2022

- Forecast accuracy improved by 20%+ — built time-series forecasting models using XGBoost, LightGBM, and Scikit-Learn.
- Inventory planning made more accurate across stores, warehouses, and fulfillment centers — developed demand-forecasting models (Python, Scikit-Learn) on sales and inventory data.
- Stockouts reduced while excess inventory was minimized — designed inventory optimization algorithms improving supply chain efficiency.
- Processing time significantly reduced, enabling timely model training — engineered large-scale ETL and feature pipelines with PySpark for sales, inventory, and logistics data.
- Ordering efficiency improved — built predictive models (Python, PyTorch, Scikit-Learn) for replenishment planning, supplier lead-time analysis, and fulfillment optimization.

ExaWizards Inc — Artificial Intelligence Intern, Visakhapatnam, India | 01/2018 – 06/2018

- 92%+ accuracy on real-time face recognition across challenging datasets with varying lighting and occlusion — developed and optimized CNN models using transfer learning.
- Fraud review efficiency increased by 35% — built fraud investigation dashboards in Power BI and Tableau.

Selected Platforms & Projects

Enterprise AI Search & AI Assist (ADP): Hybrid-retrieval search + conversational assist for payroll workflows; +35% precision; sub-second latency; guardrails, RBAC, audit logging.

LangGraph Multi-Agent Routing (ADP): Tool-oriented agent workflows routing Payroll, Tax, Benefits, HR requests; orchestration replacing manual triage.

Financial Document Intelligence & RAG (DBS Bank): Complex RAG over financial documents, hybrid search; NLP + OCR automation; fraud modeling at scale.

Demand Forecasting & Inventory Optimization (Walmart): Time-series forecasting (+20% accuracy); replenishment/inventory optimization across stores and fulfillment centers.

Core Skills

- **Production LLM Systems:** Generative AI (GenAI), Large Language Models (LLMs), LLM-Powered Applications, Retrieval-Augmented Generation (RAG), Hybrid Search, Vector Databases & Search, OpenSearch, Embeddings, Query Understanding, Context Management, Prompt Engineering, Guardrails, Vertex AI, Gemini
- **Evaluation, Monitoring & Reliability:** LLM Evaluation, Model Evaluation, Model Monitoring, Drift Detection, Regression Monitoring, A/B Testing, Experimentation, Observability, Logging & Monitoring, Rollback Readiness, 99.9% Uptime
- **Agentic AI:** Agentic AI Systems, AI Agents, Multi-Agent Systems & Workflows, LangGraph, LangChain, LlamaIndex, Tool-Oriented Workflows, Workflow Orchestration, AI Assistants, Conversational AI
- **Machine Learning & Data Science:** Python, SQL, Machine Learning, Deep Learning, Statistical Analysis, Predictive Modeling, Classification, Regression, Time-Series & Demand Forecasting, Feature Engineering, Scikit-Learn, XGBoost, LightGBM, PyTorch, TensorFlow
- **NLP & Document AI:** Natural Language Processing (NLP), Transformer-Based Models, Text Classification, Document Understanding & Intelligence, OCR, Automated Document Processing, AI Search, Search Relevance & Precision
- **Platform, Data & Deployment:** FastAPI, REST APIs, Microservices Architecture, Docker, Kubernetes (GKE, AKS), AWS, Azure (Azure ML), Google Cloud (GCP), CI/CD, MLOps, Model Serving, Production Inference, GPU & Latency Optimization, PySpark, Apache Spark, Airflow, ETL & Data Pipelines, Role-Based Access Control (RBAC), Audit Logging, Compliance
- **Domain Expertise:** Enterprise HR and Payroll Technology (Payroll, Tax, Benefits, HR), Banking and Financial Services (Fraud Detection, Risk Modeling), Retail and Supply Chain (Inventory Optimization, Replenishment Planning, Fulfillment Optimization)

Education

Jawaharlal Nehru Technological University (JNTU) — B.Tech, Computer Science, India | 2014 – 2018