

Mukesh Pandey

Senior AI / ML Engineer — Generative AI & LLM Systems

New Delhi, India | +91 9716966182 | mukesh1q2@gmail.com | [LinkedIn](#) | [GitHub](#) | [theqbitlabs.com](#)

PROFESSIONAL SUMMARY

Senior AI/ML Engineer with extensive experience architecting production-grade Large Language Model (LLM) systems, Generative AI applications, and end-to-end machine learning pipelines. Deep expertise in Python-based ML ecosystems including TensorFlow, PyTorch, and LangChain, with proven success deploying scalable AI solutions on Azure ML and Google Cloud Platform. Specialized in Retrieval Augmented Generation (RAG) pipelines, AI agent orchestration, prompt engineering, and inference optimization for enterprise-scale systems. Track record of reducing LLM hallucinations by 40%, processing 500k+ tokens daily in production, and delivering measurable business impact through advanced NLP, deep learning, and MLOps practices across consulting and leadership roles.

CORE TECHNICAL SKILLS

Programming: Python, SQL, JavaScript, TypeScript

Machine Learning & Data Science: Pandas, NumPy, Scikit-learn, Feature Engineering, Exploratory Data Analysis (EDA), Model Optimization, Model Evaluation, Hyperparameter Tuning

Deep Learning: TensorFlow, Keras, PyTorch, CNNs, RNNs, Graph Neural Networks

NLP & LLM Systems: Transformers, BERT, GPT Architectures, Word Embeddings, NLTK, spaCy, Large Language Models, Prompt Engineering

Generative AI: LangChain, Retrieval Augmented Generation (RAG), AI Agents, Hugging Face, OpenAI APIs, Vector Databases, Embeddings, Langflow, Phidata

MLOps & Cloud AI: Azure ML, Google Cloud Platform, Kubernetes, CI/CD Pipelines, Model Deployment, Inference Optimization, Containerization

Data Engineering: Databricks, Azure Data Factory, Data Pipelines, ML Pipelines, ETL Workflows

Web & Integration: Node.js, Next.js, React, REST API Design, System Architecture, Microservices

PROFESSIONAL EXPERIENCE

Lead AI Engineer & AI Architect

Baranwal Consultancy and Services | Pune, India | Jan 2025 – Present

- Architected the Brahm AI cognitive framework integrating symbolic reasoning with neural networks and Graph Neural Networks, reducing LLM hallucinations by 40% across production workloads.
- Developed production-grade LLM pipelines using Python, LangChain, and OpenAI APIs, processing over 500k tokens daily to power enterprise Generative AI automation workflows.
- Engineered Retrieval Augmented Generation (RAG) systems with vector databases and embeddings to deliver context-aware AI responses, improving answer accuracy by 35%.

- Designed and deployed scalable AI agent orchestration frameworks using Transformers, Hugging Face models, and prompt engineering for multi-step reasoning tasks.
- Built end-to-end ML pipelines encompassing data preprocessing, feature engineering, model optimization, and inference optimization on Azure ML infrastructure.
- Implemented MLOps best practices including CI/CD pipelines, model versioning, and automated model deployment on Kubernetes for continuous AI system delivery.
- Led research in advanced cognitive AI architectures combining deep learning with symbolic AI, published internal frameworks adopted across 3 product teams.

AI Consultant — Generative AI & ML Systems

Freelance AI Consulting | Remote | Jan 2024 – Present

- Delivered 10+ production AI solutions for enterprise clients, integrating Large Language Model APIs, NLP pipelines, and Generative AI systems to automate complex business workflows.
- Built RAG-powered conversational AI platforms using LangChain, vector databases, and OpenAI APIs, reducing client manual operations by 15+ hours weekly per team.
- Developed ML-powered analytics dashboards and predictive models using Python, TensorFlow, and PyTorch for data-driven decision-making across multiple industries.
- Designed and deployed REST APIs and microservices enabling scalable AI system integrations, supporting 50k+ API calls daily in production environments.
- Implemented end-to-end deep learning solutions from EDA and feature engineering through model deployment on cloud AI platforms (Azure ML, GCP).
- Provided technical consulting on AI architecture, LLM systems design, and MLOps strategy for clients transitioning to AI-first operations.

Assistant Manager — AI & Operations Analytics

Infinite Locus | Gurugram, Haryana | Nov 2023 – Jun 2024

- Managed a portfolio of 14 enterprise projects valued at over \$2M for Fortune 500 clients including Pepsico, Ericsson, and TCS, delivering AI-driven analytics solutions.
- Applied Large Language Models and NLP-based pattern recognition to automate data analysis, cutting turnaround time by 60% and saving 200+ analyst hours monthly.
- Designed data pipelines and ML workflows to process large-scale datasets, enabling predictive insights that drove 100% on-time delivery rate across all projects.

Lead Data Analyst — ML & AI Systems

Asahi Travel Service | New Delhi | Jul 2018 – May 2020

- Built machine learning models for anomaly detection and fraud identification using TensorFlow, Keras, and PyTorch, achieving 95% classification accuracy on transaction datasets.
- Designed automated data pipelines and feature engineering workflows processing 1M+ records, powering analytical dashboards in Power BI and Tableau for executive decision-making.
- Applied statistical modeling, EDA, and model evaluation techniques to identify operational patterns, resulting in 30% reduction in fraudulent transactions.

- Implemented cloud-based AI infrastructure on Zoho ecosystem (CRM, Projects), improving cross-team efficiency by 25% through intelligent automation.
-

KEY AI & MACHINE LEARNING PROJECTS

Brahm AI Cognitive Framework

Problem: LLMs suffered from hallucinations and lacked structured reasoning capabilities for complex enterprise tasks.

Technologies: Python, Transformers, Graph Neural Networks, Symbolic AI, PyTorch, LangChain

Architected a novel AI cognitive framework combining neural networks with symbolic reasoning engines. Implemented multi-layer attention mechanisms and knowledge graph integration to ground LLM outputs in verified data.

Results: Reduced LLM hallucinations by 40% and improved reasoning accuracy by 35% across production workloads. Framework adopted by 3 internal product teams.

Enterprise LLM Automation Platform

Problem: Enterprise clients required scalable conversational AI systems with domain-specific knowledge retrieval.

Technologies: LangChain, OpenAI APIs, Hugging Face, Vector Databases, RAG, Embeddings, Kubernetes

Designed and deployed a Retrieval Augmented Generation (RAG) platform integrating vector search, semantic embeddings, and generative models for enterprise automation workflows. Built AI agent orchestration for multi-step reasoning.

Results: Platform processes 500k+ tokens daily, reduced manual workflow time by 15+ hours per week per client, and achieved 92% user satisfaction score.

ML Fraud Detection System

Problem: Financial transaction data contained undetected anomalies causing revenue loss.

Technologies: TensorFlow, PyTorch, Keras, Scikit-learn, Feature Engineering, Model Optimization

Built deep learning classification models with advanced feature engineering and hyperparameter tuning for identifying fraudulent transactions across large-scale financial datasets.

Results: Achieved 95% classification accuracy, reduced fraudulent transactions by 30%, and processed 1M+ records through automated ML pipelines.

EDUCATION

Bachelor of Engineering (Polymer Science & Chemical Technology)

Delhi College of Engineering (DTU) — 2009–2013

CERTIFICATIONS

- Quantum Machine Learning — IBM Qiskit (Score: 100/100)
- Master's Certification in Machine Learning & Artificial Intelligence
- Applied Data Science with Python — WorldQuant University