ABHISHEK KUMAR

Senior ML/LLM Engineer

Bangalore, India | +91-83103 38891 | abhishek.lagran@gmail.com | linkedin.com/in/abhishek-kumar-aiml

PROFESSIONAL SUMMARY

Senior ML/LLM Engineer with 10+ years building production AI systems, specialized in Retrieval-Augmented Generation (RAG), retrieval quality tuning, and knowledge-grounded LLM applications across Azure OpenAI and AWS Bedrock. Deep expertise in hybrid retrieval (semantic + BM25), reranking, embedding and chunking strategy, prompt orchestration, and RAG evaluation frameworks (RAGAS, TruLens, LLM-as-judge). Currently leading production RAG delivery for a large enterprise engagement — evaluating PageIndex vs chunked RAG on 300+ page compliance documents, tuning retrieval precision, and implementing citation-grounded response generation. Proven record taking AI pilots from prototype to production with measurable quality gates, cost controls, and governance (GDPR, EU AI Act). AWS Global GenAI Hackathon winner. MS in ML/AI from Liverpool John Moores University.

CORE COMPETENCIES

Retrieval & RAG

- RAG pipeline design (chunking, embedding, retrieval, rerank, grounding)
- Hybrid retrieval (semantic + BM25), Cohere/cross-encoder reranking
- Azure AI Search, Qdrant, Pinecone, FAISS, OpenSearch
- Contextual retrieval, HyDE, query rewriting, parent-document retrieval
- PageIndex, LlamaIndex, LangChain, LangGraph, DSPy

LLM & Prompt Engineering

- Azure OpenAI (GPT-4o, GPT-4.1), AWS Bedrock (Claude Sonnet/Opus)
- Prompt orchestration, structured outputs, tool calling, function calling
- Citation strategy, inline grounding, provenance tracking
- RAG evaluation: RAGAS, TruLens, DeepEval, LLM-as-judge
- Faithfulness, context precision/recall, answer relevancy metrics

Cloud & MLOps

- Azure: AI Search, ML Studio, OpenAI, Cognitive Services, AKS
- AWS: Bedrock, SageMaker, Lambda, OpenSearch, ECS
- CI/CD, containerization (Docker, Kubernetes), Terraform
- Model lifecycle, A/B testing, drift detection, observability

Languages & Governance

- Python, TypeScript, SQL, PySpark, C#/.NET
- FastAPI, REST APIs, async orchestration
- AI governance (GDPR, EU AI Act), responsible AI, bias detection
- Microsoft Copilot Studio, Power Platform, Dataverse

PROFESSIONAL EXPERIENCE

Digital Engineering Staff Engineer — AI & GenAI *April 2025 – Present*

NTT DATA *Bangalore, India*

- Architecting and delivering production RAG systems for automated compliance (PIA) review — Claude Sonnet via AWS Bedrock, Qdrant vector store, FastAPI orchestration — processing 300+ page regulatory documents with citation-grounded, source-attributed responses for auditor-grade traceability.
- Leading formal head-to-head evaluation of retrieval strategies (PageIndex vs chunked RAG with reranking) for production selection; built evaluation harness measuring retrieval precision@k, context faithfulness, answer relevancy, latency, and per-query cost — producing the decision framework adopted for downstream compliance workflows.
- Tuning chunking strategy (semantic, recursive, parent-document), embedding model selection, and hybrid retrieval (dense + BM25 fusion) to maximize context precision on long-form compliance documents; implemented cross-encoder reranking for top-k refinement and reduced hallucination rate in grounded responses.

- Designed an LLM-based classification system for 222 enterprise asset classes using Claude on AWS Bedrock — with structured output validation, confidence scoring, and human-in-the-loop review for audit compliance; replaced a multi-week manual process with near-real-time categorization.
- Built RAG evaluation pipeline using RAGAS and LLM-as-judge patterns covering faithfulness, context precision, and answer relevancy — enabling quantitative pilot-to-production quality gates and regression testing on every prompt and retrieval config change.
- Architecting parallel Azure-native RAG reference implementations using Azure OpenAI (GPT-4o), Azure AI Search (hybrid + semantic ranker), and Azure ML — ensuring retrieval patterns transfer cleanly across AWS Bedrock and Azure for multi-cloud enterprise deployments.
- Led governance automation initiative implementing CR-CT-Asset mapping with LLM-powered document understanding on Microsoft Copilot Studio and Power Platform, achieving 95% audit-readiness automation and 40% reduction in deployment time.
- Implemented AI governance guardrails addressing compliance (GDPR, EU AI Act), model risk assessment, PII redaction, bias detection, and explainability across all production LLM workflows.
- Championed GenAI-first development culture across 50+ engineering teams through adoption of GitHub Copilot, Cursor, and v0.dev, reducing development cycles by 45%.

Manager — Data Science & Analytics *May 2023 – April 2025*

Factspan Analytics *Bangalore, India*

- Built production RAG-based clinical decision support for a large US healthcare deployment using Azure OpenAI (GPT-4) and Azure AI Search, integrating LangChain orchestration with Dataverse and SQL Server for real-time evidence retrieval from patient records; deployed on Azure Kubernetes Service with end-to-end monitoring.
- Designed PharmaGraph — a Neo4j knowledge graph with LLM-based entity linking and a RAG retrieval interface for complex pharmacological queries, implementing provenance tracking and inline citation so every LLM response was grounded to a source document with auditor-visible attribution.
- Won AWS Global GenAI Hackathon ($10K prize) building a Claude-based interview platform on AWS Bedrock with dynamic, context-aware question generation and RAG-grounded transcript evaluation against role competencies.
- Designed scalable MLOps infrastructure on Azure and GCP with containerized microservices and automated CI/CD pipelines, cutting model deployment time by 40%.
- Generated $1.25M+ annual savings through AI-driven optimization across ED staffing ($500K), inventory ($300K), process automation ($250K), and operational efficiency ($200K).

Lead Business Analyst *January 2021 – April 2023*

Practo Technologies *Bangalore, India*

- Implemented real-time ML-powered campaign optimization using Power BI and Azure Analytics, achieving ₹4.5L monthly cost savings via dynamic budget allocation.
- Designed Marketing Mix Model (MMM) optimizing ₹40L+ monthly ad spend, improving ROAS by 32% through attribution modeling.
- Built user segmentation engine with predictive modeling, driving 10% increase in high-value transactions and ₹2.5L monthly margin improvement.

Progressive Analytics Leadership *2015 – 2020*

Gambit Sports | Rooster Properties | Jupiter Infrastructure *India*

- Developed predictive models achieving 15% accuracy improvement in sports analytics using deep learning and statistical modeling.
- Streamlined ETL pipelines, improving data accuracy by 25% and reducing processing time by 30%.

KEY AI/ML PROJECTS

PIA Compliance RAG Pipeline (current)

End-to-end production RAG system processing 300+ page Privacy Impact Assessment documents for a Tier-1 enterprise deployment. Stack: Claude Sonnet via AWS Bedrock, Qdrant for hybrid retrieval, FastAPI orchestration, RAGAS for evaluation. Implemented semantic + recursive chunking, cross-encoder reranking, inline citation grounding, and a PageIndex-vs-RAG

comparative evaluation harness with faithfulness and context-precision quality gates. Designing Azure-native parallel using Azure OpenAI + Azure AI Search hybrid retrieval.

Enterprise Asset Classification at Scale

LLM classification system categorizing 222 enterprise asset classes for regulatory compliance using Claude on AWS Bedrock. Structured output validation, confidence scoring, rationale generation, and human-in-the-loop review loop for audit defensibility. Cut a multi-week manual categorization process to near-real-time.

PharmaGraph — Knowledge-Grounded Clinical RAG

Neo4j knowledge graph with LLM-based entity linking and RAG retrieval interface for complex pharmacological queries. Every response grounded to source evidence with inline citation and provenance tracking. Built on Azure OpenAI with Azure AI Search for hybrid keyword + semantic retrieval.

Claude Interview Platform — AWS GenAI Hackathon Winner

$10K global prize. Claude on AWS Bedrock with dynamic, context-aware question generation and RAG-grounded transcript evaluation against role competencies. Prompt orchestration, structured output scoring, and quality validation across generated and evaluated content.

EDUCATION & CERTIFICATIONS

Master of Science — Machine Learning & AI *2019 – 2021*

Liverpool John Moores University, UK

Post Graduate Diploma — Machine Learning & AI *2019 – 2020*

IIIT Bangalore

Bachelor of Technology — Engineering *2012 – 2015*

Manipal Institute of Technology

Professional Certifications

- Microsoft Azure AI Fundamentals & Generative AI (2024)
- AWS Certified Machine Learning – Specialty (In Progress)
- Google Cloud — Introduction to Generative AI (2024)
- Deep Learning Specialization — deeplearning.ai (2020)
- Certified Scrum Product Owner (CSPO®) — Scrum Alliance (2023)

TECHNICAL EXPERTISE

Retrieval & RAG: Azure AI Search (hybrid + semantic ranker), Qdrant, Pinecone, FAISS, OpenSearch; LangChain, LangGraph, LlamaIndex, PageIndex, DSPy; semantic/recursive/parent-document chunking; Cohere Rerank, cross-encoder reranking; HyDE, query rewriting, contextual retrieval

LLM Platforms: Azure OpenAI (GPT-4o, GPT-4.1, embeddings), AWS Bedrock (Claude Sonnet/Opus, Titan), Anthropic API, OpenAI API

Evaluation: RAGAS, TruLens, DeepEval, LLM-as-judge; faithfulness, context precision/recall, answer relevancy, groundedness

Cloud & MLOps: Azure Machine Learning, AWS SageMaker, GCP Vertex AI; AKS, Docker, Kubernetes; Azure DevOps, GitHub Actions, Terraform

Microsoft Stack: Copilot Studio, Power Platform (Power Apps, Power Automate, Power BI), Cognitive Services, Visual Studio, C#/.NET, SQL Server, Dataverse

Languages & Data: Python, TypeScript, SQL, PySpark, C#; Neo4j, PostgreSQL, MongoDB; FastAPI, async orchestration; Git, JIRA

AI/ML: TensorFlow, PyTorch, XGBoost, Hugging Face Transformers, sentence-transformers