ABHISHEK KUMAR

Lead Generative AI Engineer — Multi-Agent Systems · Enterprise RAG Platforms · Scalable AI & GPU Infrastructure

Bengaluru, India | +91 72779 05904 | abhi3229@outlook.com | linkedin.com/in/abhishekkumarai | github.com/abhishekkumarai

PROFESSIONAL SUMMARY

Lead Generative AI Engineer with **10+ years** across software, distributed data systems, and machine learning engineering — the last **3+ years architecting and leading production GenAI platforms** for global pharma leaders (**Bristol Myers Squibb, Eli Lilly**). Proven technical leader directing teams of **5–6 engineers** across multi-agent system design (**LangGraph**), enterprise Retrieval-Augmented Generation (**RAG**), continuous LLM evaluation, observability, and guardrails (**LangSmith**), and autoscaled GPU inference infrastructure on **AWS and Kubernetes (EKS)**. Track record of establishing enterprise AI architectural standards, bridging scientific ambiguity with scalable engineering design, slashing inference infrastructure costs by **60–70%**, and accelerating drug discovery screening workflows from months to days.

PROFESSIONAL EXPERIENCE

Lead GenAI Engineer — Syngene International (Client: Bristol Myers Squibb)

Apr 2025 – Present

Bengaluru, India · Technical Lead heading a team of 5 engineers building production GenAI platforms for BMS Drug Discovery R&D

- **Technical Leadership & Architecture:** Led an engineering team of 5 in architecting and delivering enterprise GenAI platforms; defined agentic design principles, coding standards, CI/CD automation, and evaluation benchmarks adopted across multiple research units.
- **Multi-Agent Compound Library Platform (LangGraph):** Built and shipped a multi-agent LLM platform empowering BMS computational chemists for iterative compound analysis; engineered stateful supervisor routing, governed tool execution, and dual-tier memory (short-term conversational context + long-term episodic vector storage); integrated LangSmith evaluation for complete auditability in regulated research environments, **onboarding 5 project teams**.
- **High-Throughput Screening (HTS) Autonomous Triage:** Spearheaded the design of an autonomous supervisor-orchestrated multi-agent reasoning system (advocate, critic, and hypothesis agents) that triaged **3,000,000 compounds down to 3,000 high-confidence hits** with human-in-the-loop calibration, **eliminating ~5 months of manual review effort** and accelerating hit-to-lead discovery.
- **ChatEDA (Conversational Analytics Platform):** Delivered a natural-language data analytics platform over complex R&D datasets; translated user intent into validated transformations, dynamic visualizations, and RDKit cheminformatics operations, enforcing strict sandbox execution and guardrails for regulated research.
- **GPU Inference Infrastructure & Cost Governance:** Championed the migration of production LLM inference from static SageMaker endpoints to an autoscaled Kubernetes (EKS) GPU hosting model (NVIDIA A100/A10); engineered queue-based workload scheduling with Redis and FastAPI, **cutting enterprise inference infrastructure costs by 60–70%** while preserving strict latency SLAs.
- **Standardized Containerized Tooling & MCP:** Standardized model container deployment patterns and Model Context Protocol (MCP) integrations across agent workflows, reducing new agent rollout cycles by 50% and improving operational agility.

Lead GenAI / LLM Engineer (Tech Lead) — Eli Lilly and Company

Nov 2023 – Mar 2025

Bengaluru, India · Led a team of 4 engineers; architected the enterprise GenAI RAG platform adopted across 5 global business units

- **Enterprise RAG Architecture & Governance Standards:** Architected an enterprise-grade Generative AI RAG platform (LangChain, AWS Bedrock, EKS) serving multiple global business units; established company-wide standards for embedding model benchmarking, chunking strategies, and retrieval topologies—**adopted by 5 business units**, accelerating secure GenAI adoption across the enterprise.
- **Hybrid Retrieval & Shared Enterprise Memory:** Engineered high-scale hybrid retrieval combining dense semantic vector search (Pinecone) and sparse keyword search (Elasticsearch); designed a shared enterprise memory layer that substantially enhanced answer relevance while **slashing vector retrieval infrastructure costs by ~90%**.
- **Security, RBAC & GxP Compliance:** Implemented user-aware, role-based document access control (RBAC) at the retrieval layer, ensuring strict data governance, cross-departmental privacy, and audit compliance across sensitive enterprise datasets.
- **Model Gateway & Vendor Independence:** Built a unified model gateway and containerized microservice framework for open-source Hugging Face models on EKS, eliminating vendor lock-in, enabling rapid zero-downtime model onboarding, and establishing enterprise-wide inference cost governance.
- **Automated Evaluation & Governance Framework:** Established automated continuous evaluation and governance pipelines (response grading, regression test suites, hallucination and drift monitoring), enhancing platform reliability and **reducing manual validation overhead by ~40%**.
- **Domain Agent Ecosystem (Agent-to-Agent Orchestration):** Evolved the RAG platform into a domain-specific agent ecosystem (SQL generation, JIRA automation) with structured agent-to-agent (A2A) orchestration, automating mission-critical decision workflows across business units.

Senior Data Engineer — Biofourmis

Sep 2020 – Jun 2023

Bengaluru, India · Real-time streaming platform and data infrastructure for continuous patient health monitoring

- **Streaming Infrastructure at Scale:** Led development of a real-time data streaming platform using PySpark, serverless AWS components (Lambda, Redshift), and Kubernetes (EKS), enabling low-latency ingestion and processing of medical and IoT sensor telemetry for continuous patient monitoring.
- **Data Quality & Reliability Frameworks:** Built automated data quality and validation pipelines across heterogeneous relational, NoSQL, and file-based data sources, eliminating manual validation overhead and boosting downstream analytics reliability.
- **Engineering Mentorship & Operations:** Mentored mid-level and junior engineers on distributed systems design, memory optimization, and production best practices; resolved critical production bottlenecks spanning PySpark memory tuning and Airflow orchestration reliability.

Software Engineer — Maple Labs (Client: Data Capital Management, US)

Aug 2019 – Aug 2020

- **Financial Analytics & Backtesting:** Engineered data pipelines and backtesting workflows for stock-market analytics using Python and cloud-native tooling, evaluating algorithmic trading strategies against historical datasets.
- **Monolith Modernization & Full-Stack Delivery:** Refactored a legacy monolithic platform into scalable Python/AWS microservices; developed the Snappy Flow "Poller" observability application (Python/Django backend, Next.js/React frontend) with automated Jenkins CI/CD on AWS.

Application Development Analyst / ML Engineer — Accenture

Oct 2015 – Aug 2019

- **Production ML API Deployment:** Developed and deployed predictive machine learning models (Python, R) for Data Science Lab initiatives, exposing them as containerized REST APIs (Docker, Django, RPlumber) within an MLOps framework.
- **Enterprise Data Governance & Ingestion:** Engineered GDPR compliance pipelines for the IDAG (GDPR-as-a-Service) platform; designed Python/Flask data ingestion services for NBN Australia's GDSS project in PostgreSQL.

TECHNICAL SKILLS

Generative AI & Multi-Agent Systems: Multi-Agent Architectures (LangGraph, Supervisor & Debate Loops, Human-in-the-Loop), Agent-to-Agent (A2A), Model Context Protocol (MCP), Governed Tool Calling, Enterprise RAG, LangChain, AWS Bedrock Agents, Prompt Engineering, Memory Systems

Foundation Models, Evaluation & Observability: Anthropic Claude, OpenAI (GPT-4o/o1), Open-Source LLMs (Llama, Mistral via Hugging Face), LLM Evaluation & Observability (LangSmith, Ragas), Semantic Drift Monitoring, Trajectory Grading, Guardrails, GxP Compliance & Auditability

Retrieval & Vector Search: Hybrid Retrieval (Dense Semantic + Sparse Keyword / BM25), Vector Databases (Pinecone), Elasticsearch, Embedding Model Selection & Benchmarking, Re-ranking (Cohere Rerank), Chunking Optimization

Cloud, GPU & Infrastructure: Amazon Web Services / AWS (EKS, SageMaker, Lambda, S3, Redshift), Kubernetes (EKS), GPU Inference Optimization (NVIDIA A100/A10), Autoscaled Serving, Redis (Queue Orchestration & Caching), Docker, Airflow, CI/CD

Languages & Core Frameworks: Python, SQL, PySpark, FastAPI, FastMCP, Pandas, Django, Flask, PostgreSQL, MongoDB

Technical Leadership & Delivery: Technical Strategy & Architecture, Team Leadership & Mentorship (5+ Engineers), Cross-Functional Stakeholder Alignment (Computational Chemists, Biologists, Business Units), System Design Reviews, Cost Governance

EDUCATION

Bachelor of Engineering (B.E.) in Information Science & Engineering — Dayananda Sagar College of Engineering (VTU), Bengaluru, India · 2011–2015