Abhishek Gupta K

AI Engineer — GenAI & Agentic Systems

2001abhigupta@gmail.com | +91 9248637637 | LinkedIn | GitHub

Professional Summary

AI Engineer with 3 years building production GenAI and ML systems for **Fortune 500** clients. Architects multi-agent LLM platforms, ensemble forecasting pipelines, and production RAG systems on GCP, AWS, and Azure. Hands-on depth from Transformer internals through LLM evaluation, with a track record of measurable business impact. Brings end-to-end production ownership across the full AI lifecycle; from experimentation and fine-tuning to LLMOps observability, drift detection, and CI/CD deployment gates.

Skills

Agentic AI & LLM Systems: LangGraph, LangChain, LlamaIndex, CrewAI, Multi-agent Orchestration, RAG, MCP, Tool-calling, Prompt Engineering, LLM Fine-tuning (LoRA/QLoRA), LLM Evaluation (RAGAS), HuggingFace Transformers, OpenAI/Gemini/Claude APIs

Cloud & MLOps: GCP (Vertex AI, Cloud Run, Cloud Build), AWS (SageMaker, S3, Lambda), Azure (OpenAI Service, AI Search), Docker, FastAPI, REST APIs, Git/GitHub Actions, Langfuse, MLflow, Model Monitoring, ChromaDB

ML & Modeling: Supervised/Unsupervised Learning, XGBoost, Ensemble Methods, Time Series, A/B Testing, Bayesian Inference, Causal Inference, SHAP Explainability

Programming & Data: Python (PyTorch, TensorFlow, Scikit-Learn, Pandas), PostgreSQL, SQL, Power BI, Agile/Scrum

Professional Experience

The Modern Data Company — **June 2023 – Present**

AI Engineer — Hyderabad, India

- Launched a **stateful agentic LLM system** (LangGraph, FastAPI, PostgreSQL) on GCP Cloud Run for natural-language dataset queries; cut analyst turnaround by ∼**40%** and surfaced answers no dashboard could provide.
- Designed a **production RAG pipeline** with hybrid retrieval; applied **RAGAS metrics** for faithfulness and context precision, iterated chunking strategy, and reduced analyst lookup time across the team's shared documentation.
- Shipped a production **multi-model forecasting pipeline** (XGBoost, LightGBM, Deep Learning) on GCP Cloud Run with real-time inference, reducing forecast error by **26%** and cutting inventory costs for global distribution.
- Deployed a brand-level demand model (XGBoost + Deep Learning) with **MLflow** tracking and versioning; achieved a **33% uplift** in shipment accuracy across global distribution and procurement networks.
- Spearheaded **Survival Analysis** with Kaplan-Meier estimators to quantify churn risk; improved churn-model AUC by **21%** through behavioral feature engineering on usage frequency, contract tenure, and product recency signals.

Dr. Reddy's Laboratories — **June 2022 – December 2022**

Data Analyst Intern — Hyderabad, India

- Built **statistical validation pipelines** for pilot-scale ML initiatives, cutting null-rate and schema violations by **30%** and improving downstream feature quality for production model-training pipelines.
- Performed root-cause **data-quality analysis** across 12+ source tables, resolving systemic inconsistencies that reduced downstream model bias and improved prediction reliability across global ingestion pipelines.
- Built automated data aggregation and reporting workflows using **Python** and **SQL**, consolidating outputs from 8+ production and sales systems into weekly decision-ready reports for 4 cross-functional business teams.

Projects

Sports Analytics Intelligence Platform — **Python, XGBoost, PyTorch, LangGraph, Vertex AI, GCP, Docker**

- Designed a GCP Cloud Run **MLOps pipeline** with a **5-model ensemble** (XGBoost, LightGBM, PyTorch, TensorFlow), ONNX export, SHAP explainability, and **Vertex AI Model Registry** with CI/CD retraining.
- Orchestrated a **multi-agent LLM system** (LangGraph + 3 specialist agents) with SSE streaming, PostgreSQL memory, and a 3-tier **RAG pipeline**; instrumented with **Langfuse** for LLMOps tracing and RAGAS-based retrieval evaluation.
- Applied **SPRT** for continuous A/B testing (50–70% fewer samples), Bayesian Bradley-Terry team ratings, and **Difference-in-Differences** causal inference; built a Prophet + ARIMA ensemble for 14-day win-rate forecasting.
- Implemented **LoRA/QLoRA** fine-tuning on NBA play-by-play data, enabling parameter-efficient sports analytics question answering on a foundation LLM; achieved ∼**80% reduction** in trainable parameters versus full fine-tuning.
- Deployed a hybrid serverless stack (Vercel + Cloud Run) with **KL-divergence** drift detection, Prometheus/Grafana observability, and a **GitHub Actions** regression gate enforcing model quality on every merge to main.

Education

Birla Institute of Technology and Science — **August 2019 – May 2023**

Bachelor of Engineering (B.E.) – Mechanical Engineering — Hyderabad, India