VIVEK GUPTA

Staff Machine Learning Engineer at PayPal · IIT Kanpur · 9+ Years Experience

Bangalore, India · gvivek2692@gmail.com · +91 98156 19602

SUMMARY

Staff-level engineer with 9+ years of building and architecting **production-grade AI systems on top of distributed-systems backends** — across **fintech (PayPal, IndMoney)**, marketplaces, and ad-tech. At PayPal, leading 0-to-1 delivery of an agentic AI assistant (LangGraph, CrewAI, **RAG, MCP-based tool orchestration**) and building an **internal LLM evaluation and observability platform** used across DS / ML teams. Earlier shipped **personalisation / recommendation engines, transaction-data enrichment and classification systems**, and a **bot-detection classifier at 10M+ sessions/day**. Lead Data Scientist titles at three companies; **managed 5 data scientists at Cars24**; 2 years at Bain advising Fortune-500 leadership. Comfortable owning AI-native architecture end-to-end with explicit tradeoffs across latency, cost, reliability, and model performance.

TECHNICAL SKILLS

- **Languages: Python (expert)**, SQL (advanced), PySpark, JavaScript / TypeScript (working knowledge)
- **AI Agents & Orchestration: agentic workflows, orchestration of models / tools / memory / decision-making** — LangGraph, LangChain, CrewAI, OpenAI Agents SDK; **MCP tool integration / tool-calling frameworks**; multi-step reasoning, prompt chains, refinement loops; OpenAI / Claude / Bedrock APIs, function calling, structured outputs
- **RAG, Embeddings & Vector DBs:** end-to-end RAG (chunking, **embeddings**, hybrid search, re-ranking, citation enforcement); **vector stores** (Pinecone, Weaviate, ChromaDB, FAISS, Elasticsearch); HuggingFace Transformers, PyTorch, TensorFlow; fine-tuning (SFT, LoRA, RLHF — PPO, GRPO)
- **Evaluation, Observability & Guardrails:** internal LLM eval platform — agent evals, regression harnesses, simulated trajectories, **LLM-as-judge, rubric-based scoring**, guardrails, hallucination control, quality / latency / cost dashboards; fallback / degradation logic
- **Backend & Distributed Systems:** REST APIs and microservices for AI capabilities, **async / event-driven pipelines**, schema validation, **Spark / PySpark at 10M+ events/day**, Kafka (familiar), Hadoop, Hive, BigQuery; large-scale ETL on structured + unstructured data
- **Personalisation & ML: recommendation and ranking engines**, transaction-data enrichment and classification, NER, embeddings-based similarity, XGBoost, LightGBM, CatBoost, scikit-learn, SHAP
- **MLOps / LLMOps & Cloud:** MLflow, Kubeflow, Docker, Kubernetes (familiar), CI/CD for ML, model versioning, drift detection, monitoring; AWS (SageMaker, S3, EC2, Lambda), GCP (Vertex, BigQuery), Azure; multi-GPU training
- **Leadership:** Managed 5 data scientists; mentored backend and AI engineers on system tradeoffs, evaluation rigor, and production reliability; influenced technical roadmaps across product, data, and leadership

EXPERIENCE

PayPal — MTS-1 Machine Learning Engineer

Bangalore *Dec 2024 – Present*

- **Led architecture and 0-to-1 delivery of the KYC Ops Assistant** — production **agentic AI service** (LangGraph, CrewAI) on a fintech compliance surface; defined the **orchestration layer for models, tools, memory, and decision-making**. **Doubled KYC conversions in 3 months**.
- Designed the **RAG pipeline** (chunking, embeddings, vector retrieval, hybrid search + re-ranking, citation enforcement) and **integrated multiple MCP tools** for workflow automation — exposing internal services as guarded, schema-validated agent tools.
- Building an **internal LLM evaluation and observability platform** used across DS / ML teams — agent evals, regression harnesses, LLM-as-judge, guardrails, and offline + online evaluation tied to **quality, latency, hallucination, and cost**; informs fallback and reliability decisions.
- Fine-tuning LLMs as digital agents (Venmo) using SFT and RLHF (PPO, GRPO) on multi-GPU clusters — owning training, evaluation, monitoring, and production rollout.

Cars24 — Lead Data Scientist

Bangalore *Nov 2021 – Dec 2024*

- **Managed a team of 5 data scientists** — hired, set roadmap and priorities, ran code / model reviews, mentored on system tradeoffs and production reliability.
- Owned the **Dynamic Pricing Engine** end-to-end (CatBoost) — **4% gross-margin lift** at scale with SHAP explainability and A/B testing as source of truth; designed around latency, cost, and conversion-velocity tradeoffs.
- Built and rolled out **reusable team-level backend platforms** for feature engineering, experimentation, deployment, and monitoring — adopted across product squads, reducing time-to-production.

Bain & Company — Senior Associate, Data Science

Bangalore *Mar 2020 – Oct 2021*

- Embedded technical advisor for **Fortune-500 retail and telecom leadership** — scoped workstreams, translated model behavior into executive decisions.
- Built **demand-forecasting and price-optimization models** (LightGBM, Hyperopt) — **5% reduction in promotion spend**; NLP pipelines (topic modeling, sentiment) for a telecom client's churn strategy.

IndMoney — Lead Data Scientist *(10th employee)*
Gurgaon *Aug 2018 – Feb 2020*

- Built and shipped the **Mutual Fund Recommendation & Ranking Engine** (XGBoost + similarity ranking) — production fintech **personalisation surface** serving end users daily.
- Designed an **AI-driven SMS transaction-parsing system** (NER + supervised text classification) — **large-scale data enrichment, classification, and correlation across transaction data**, with a labeling pipeline that continuously improved precision / recall on edge cases.
- Established the data-science function from scratch as the 10th employee, owning the end-to-end lifecycle.

Cogneau Systems — Data Scientist *Gurgaon · Oct 2017 – Jul 2018*

- Optimized a distribution-center network for a Malaysian logistics customer using **DBSCAN** — **30% reduction in operations cost**.

Simpli5d Technologies — Data Scientist *Gurgaon · Nov 2016 – Sep 2017*

- Built a **real-time bot-detection classifier** (Random Forest on **Spark / PySpark**) processing **10M+ user sessions / day across 120+ publishers**; designed the Spark feature pipeline and distributed-data infra — **high-throughput, low-latency event processing**.

Mu Sigma — Trainee Decision Scientist *Bangalore · Aug 2014 – Feb 2015*

- Analyzed search-advertising KPIs across Asia Pacific for Bing Ads.

EDUCATION

Indian Institute of Technology (IIT) Kanpur — B.Tech., Bio Science · Jul 2010 – May 2014 · CGPA 7.0