

Vaibhav Kumar

vaibhavpro9210@gmail.com | [LinkedIn](#) | [GitHub](#) | vaibhavgit9210.github.io

AI / GenAI Engineer | LLM Agents & RAG | Multi-Agent Systems | Data Engineering

EXPERIENCE

PLIVO

May 2024 – Present

Data Scientist / AI Engineer (titled Business Analyst — legacy classification)

Bangalore, India

- Architected a production **multi-agent AI system** for US compliance — agents that *reason, plan, and take actions across tools and APIs*. 4 specialized agents on **Claude, OpenAI, Azure OpenAI** with **structured tool-calling**, event-driven concurrency (atomic SQL), webhook callbacks. Defined **KPIs** and **optimized agent behavior**; lifted classification accuracy **62%→84%** via **prompt-template design, structured prompting**, A/B eval, and feedback-driven re-prompting; automated **100% of US vetting**. *Thanks Award — GenAI in KYC automation.*
- Built and shipped a production **RAG pipeline** grounding LLMs in structured + domain data — **FastAPI** + Docker on **HuggingFace Spaces**. **LanceDB** vector store, **Voyage AI voyage-3-large** embeddings (1024-dim), 3 indexes (past tickets, Confluence runbooks, public docs via **Firecrawl**); **14,492 vectors**; hybrid filter + semantic search with quality-weighted reranking; tested against **hallucinations** and edge cases. Exposed as 3 tool-callable **/search** endpoints; retrieval evaluation via **Recall@k, F1**, end-to-end answer quality on golden sets.
- Built a Claude-powered **AI agent with memory and multi-step task strategies — 12-step orchestration** (LangGraph-equivalent state machine) with **11 tools**: Postgres / Redshift retrieval, Anthropic Vision document parsing, Zendesk + Slack. Auto-resolves customer tickets across **7 categories**. Established **safety & alignment guardrails**: JSON-schema output validation, structured outputs, low-confidence routing to human review, **human-in-the-loop evaluation** for generative outputs.
- Designed batch **anomaly-detection + classification** on **2M+ daily / 10M+ weekly** telemetry — Python parallel processing, **DynamoDB** caching (TTL, 80%+ hit rate), LLM-based risk scoring with **calibrated confidence**, REST API, **AWS QuickSight** + SNS → Slack alerting. **-70% audit load, -85% response time**. Established **LLMOps**: prompt versioning, A/B eval harness, golden sets, drift / feedback. *Applause Award.*

RAZORPAY

Jul 2022 – May 2024

Technical Consultant — Integrations

Bangalore, India

- Built payments analytics + reconciliation on **AWS — dimensional data model** (star-schema in PostgreSQL), ETL of **50M+ daily transactions** into **Redshift** (MPP / columnar warehouse), reconciliation across MySQL/Postgres/Mongo at **99.97% accuracy** on **200GB+ daily** (advanced SQL: CTEs, window functions, FULL OUTER JOIN). Cut API P95 from **850ms→320ms** via composite indexes, query rewrites, and **Redis** caching.
- Owned integration lifecycle for **20+ enterprise fintech customers** — REST APIs, webhooks (idempotency, HMAC, retries), JSON/XML schema normalization in **Python**; **99.9% delivery** on **2M+ daily API calls, -15% incidents/month** via log-based RCA and A/B experiments. *MVP Award — integration lifecycle and pipeline quality.*

PROJECTS

Sign Language Translator (*Project of the Year*) | *Python, OpenCV, MediaPipe, Bi-LSTM, TensorFlow*

2022

- End-to-end **deep learning** pipeline on **30K+ video samples** (augmented to **150K+**); 543 MediaPipe landmarks/frame, **bidirectional LSTM** at **85% accuracy**, real-time serving **30 FPS / <50ms latency**.

fcstudio.co.in — live e-commerce platform, sole engineer | *Next.js 14, MongoDB, Razorpay, Shiprocket*

2024 – Present

TECHNICAL SKILLS

AI / LLM: LLMs & transformer models (Claude, GPT-4 / 4o, LLaMA, Azure OpenAI — prompt engineering, structured prompting, attention fundamentals); **Agents** that reason, plan, and take actions across tools and APIs; **memory strategies** for multi-step tasks; **safety & alignment guardrails**; hallucination / edge-case testing

Fine-tuning & Adaptation: LoRA / prompt-tuning / RAG (familiar — production RAG shipped; LoRA conceptual); deep learning (Bi-LSTM — TensorFlow, academic); classical ML (gradient boosting, isolation forest, autoencoders — familiar)

Evaluation Metrics: perplexity, BLEU, ROUGE, accuracy, F1, calibration, Recall@k, NDCG; human-in-the-loop evaluation for generative outputs; golden sets + A/B harness; drift / feedback loops

RAG & Vector Search: LanceDB (production), Pinecone, FAISS, pgvector; **Voyage AI**, OpenAI embeddings, sentence-transformers, BGE / E5 (familiar); chunking, hybrid filter + semantic search, reranking; knowledge-graph-style retrieval over relational sources

Languages & Stack: Python (pandas, NumPy, asyncio, parallel); **Java** (familiar); **FastAPI, Docker**; PyTorch (familiar), TensorFlow, scikit-learn; **LangChain / LlamaIndex / Haystack** (familiar); **Advanced SQL** (CTEs, window functions);

AWS (Lambda, S3, DynamoDB, Redshift, QuickSight); cloud-native development

EDUCATION

N.M.A.M. Institute of Technology

Bangalore, India

Bachelor of Engineering in Computer Science

2018 – 2022, CGPA: 7.25