PRAVEEN LAL T H

+91 9446060094 ◇ Bengaluru, India
lal.praveen123@gmail.com ◇ linkedin.com/in/plal99 ◇ github.com/plal99

SUMMARY

AI Engineer with 4+ years building production-grade GenAI systems and RAG pipelines at scale. Expert in LLMs (LangChain, LlamaIndex, LangGraph), vector databases (semantic search), and multi-agent architectures. Delivered enterprise contract intelligence platform using Vertex AI and BigQuery, reducing manual review by 65%. Proven ability to architect, deploy, and optimize intelligent systems on GCP using modern MLOps practices and retrieval-augmented generation.

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript, Node.js, React
Generative AI & LLMs: Vertex AI, LangChain, LlamaIndex, LangGraph, Embeddings, RAG, FAISS/Pinecone, Prompt Engineering, Structured Outputs, Re-ranking, BLEU/ROUGE/F1
Agents & Backend: Multi-Agent Systems, A2A Orchestration, Google ADK, Event-Driven Architecture, Django, FastAPI, REST APIs, Microservices
Data & Cloud: BigQuery, Cloud Storage, Cloud Run, Cloud Functions, Pub/Sub, Dataflow, PostgreSQL, Elasticsearch
MLOps & DevOps: Model Monitoring, Drift Detection, A/B Testing, Docker, Kubernetes, CI/CD, Git, Cloud IAM

EXPERIENCE

Deloitte Consulting India Pvt Ltd Jan 2022 – Present
AI Engineer Bengaluru, India

- Architected production-grade GenAI platform using Vertex AI, LangChain RAG, and FAISS vector databases for contract intelligence (SoW, Order Forms, Amendments); reduced manual review time by **65%** across legal teams with structured output validation for JSON schema compliance.
- Designed scalable ingestion pipeline (Google Drive → Cloud Storage → Embeddings + BigQuery) processing **2,500+ contracts/month** with **95% automated metadata accuracy**.
- Built **hybrid retrieval engine** combining semantic RAG and direct SQL on BigQuery, improving answer precision by **42%** and reducing hallucination rates; implemented LLM evaluation with BLEU/ROUGE metrics and human feedback loops.
- Engineered modular **Agent-to-Agent (A2A)** architecture using Google ADK and async Pub/Sub event queues; specialized agents for RAG retrieval, analytics, and summarization cut query latency from **1.8s to 0.9s**.
- Developed multi-stage **contract drafting pipeline** using chained LLM prompts to extract clauses, validate business rules, and populate standardized templates, minimizing manual intervention.
- Extracted and unified contract metadata (expiry, value, renewal terms, amendment lineage) in BigQuery, enabling structured reporting and a **5x faster audit cycle**.
- Built library of **reusable agent skills** — modular prompt templates and LLM workflow components — standardizing AI outputs across delivery teams and **cutting ramp-up time by 50%**.
- Created **Django dashboard** for expiry timelines, amendment chains, and renewal alerts supporting **10K+ monthly queries**; owned end-to-end GCP deployment (Cloud Run, Functions, IAM) with CI/CD pipelines.

EDUCATION

Government Engineering College, Thrissur 2017 – 2021
B.Tech in Electronics and Communication Engineering GPA: 8.8/10

CERTIFICATIONS

AWS Certified Developer – Associate (2025)
Deep Learning Specialization – Coursera
CS50's Introduction to Computer Science – HarvardX
Python for Everybody – Coursera