Kunj Patel AI Engineer

+91 94845 46452 • kunjpatel91012@gmail.com • linkedin.com/in/kunjpatel101 • github.com/kunj10

TECHNICAL EXPERTISE

Core Engineering: Python, FastAPI, Git, Machine Learning, Deep Learning, Docker, CI/CD
Agentic AI: LangChain, LangGraph, Google ADK, Model Context Protocol (MCP), Agent-to-Agent
RAG & Retrieval: Semantic Search, Vector Databases, Embedding Optimization, Knowledge Grounding
Cloud & Ops: Google Vertex AI, Agent Engine, Cloud Run, AWS SageMaker, AWS Bedrock

EXPERIENCE

AI Engineer - Vypar TaxOne *Nov, 2025 - Present*

- Built the "**GST Compliance Assistant,**" a domain-specific RAG system allowing CAs to accurately retrieve context from complex legal documents.
- Benchmarked next-gen RAG frameworks (**HyDE, CAG, KAG, Agentic RAG**) to select the most effective architecture for handling dense regulatory queries.
- Engineered an **OCR-to-LLM pipeline** for financial data, boosting extraction accuracy from 40% to **80%** and cutting down manual data entry time.
- Implemented long-term memory using **MongoDB** to persist user sessions and chat history, enabling context retention across long conversations.
- Deployed the full-stack AI application on **Google Cloud Run** (Serverless) with a CI/CD pipeline for auto-scaling and high availability.

PROJECTS

Autonomous Multi-Agent Code Auditor *Google ADK, CrewAI, LangGraph, Gemini 2.5*

- Orchestrated a multi-agent system (Orchestrator, Auditor, Planner) to automate security code reviews without human oversight.
- Developed a custom **Agent-to-Agent (A2A)** communication protocol using FastAPI, allowing agents to delegate tasks and aggregate findings.
- Integrated **Model Context Protocol (MCP)** for secure file system access, enabling agents to read codebases, spot vulnerabilities, and draft fix reports.

Financial Intelligence Platform *FastAPI, FAISS, Synthetic Data*

- Built a semantic search engine using **vector embeddings** and FAISS, enabling natural language queries on transaction data with sub-2s latency.
- Developed **FastAPI** microservices for real-time spending analysis and context-aware anomaly detection.
- Generated 500K+ synthetic financial transactions to stress-test the model and ensure performance under load while maintaining privacy.

RESEARCH

RAG-Enhanced LLM for Web-Based Assistance *IEEE ICCES 2024*
Built a hybrid RAG pipeline with LLaMA 3, Gemini, and FAISS that improved contextual accuracy by **31% improvement** in contextual accuracy over baseline models. DOI: 10.1109/ICCES63552.2024.10859894

EDUCATION & CERTIFICATIONS

B.E. Computer Science and Design *2021–2025*
A.D. Patel Institute of Technology *CGPA: 8.86 / 10*

Certifications: Google Cloud (ML Engineer), AWS (GenAI with LLMs), IBM (ML Professional)

LEADERSHIP

Training & Placement Coordinator: Coordinated industry outreach and technical recruitment at ADIT.