

Ravi Kumar Mishra

+91 7012257034 | ravikumar89386@gmail.com | [LinkedIn](#) | [GitHub](#)

SUMMARY

AI Engineer with 2+ years of experience building and deploying production Generative AI systems in the BFSI domain, specializing in **RAG, agentic workflows, multi-agent systems, and document intelligence**. Experienced in designing custom and managed RAG pipelines, stateful agentic workflows using Google ADK and LangGraph, and LLM-powered automation integrated with enterprise systems and business workflows. Strong focus on retrieval quality, evaluation, observability, guardrails, and production reliability, with hands-on experience deploying AI solutions on Google Cloud.

EXPERIENCE

Tata Consultancy Services AI Engineer

Pune, India
Jun 2024 – Present

- Multi-Agent Incident Management System:** Automated Tier-1 login/MFA incident resolution using LangGraph orchestration and human-in-the-loop (HITL) approval, combining LLM-based intent classification, RAG-based retrieval of previously solved ServiceNow incidents, and controlled diagnostic workflows. Internally integrated Client Registry, Co-Admin, and ServiceNow systems for user and incident validation, with task creation for escalating unresolved cases to L2 support teams. A scheduled Cloud Job refreshes the RAG corpus with newly resolved ServiceNow incidents.
Tech: LangGraph · Gemini · RAG Engine · FastAPI · Cloud Run · Firestore · ServiceNow API
- Customer Engagement Suite – Call Transcription & Summarization:** Built a real-time customer-call GenAI pipeline integrating Genesys Cloud AudioHook with Google Cloud Agent Assist for live speech transcription, knowledge assistance, and Gemini-based conversation summarization. Developed Cloud Run middleware to stream dual-channel call audio to Agent Assist, process Pub/Sub events, and deliver transcripts and suggestions to the Salesforce agent workspace, with Redis-based session mapping for conversation state.
Tech: Genesys Cloud AudioHook · Google Cloud Agent Assist / CCAI · Vertex AI (Gemini) · Cloud Run
- Insurance QA Bot – Multi-Agent RAG:** Built a production multi-agent RAG system supporting 6 insurance product lines, selecting custom Pinecone-based RAG and managed Vertex AI RAG Engine based on product requirements. Implemented hybrid retrieval with Dense Search, BM25, RRF fusion, and cross-encoder reranking, alongside iterative answer review, golden-dataset evaluation, and session memory to improve answer grounding and retrieval reliability.
Tech: Google ADK · Vertex AI Agent Engine · Pinecone · Vertex AI RAG Engine · Gemini
- AI OCR & Document Intelligence Platform:** Contributed to an 8-service, configuration-driven insurance document processing platform on Google Cloud Run, integrating Document AI OCR and Gemini for document classification, semantic field mapping, and source-document validation. Implemented insurer-specific extraction workflows using prompts and field definitions managed through GCP Parameter Manager. Built an offline ground-truth evaluation workflow to measure field-level extraction quality, improving results from an 87% baseline to **98.7% accuracy** across insurer-provided evaluation datasets, without model retraining.
Tech: Python · FastAPI · Document AI · Gemini · firestore · GCS · Cloud Run

TECHNICAL SKILLS

Languages: Python, SQL

Generative AI & Agents: Gemini, Google ADK, LangGraph, Vertex AI Agent Engine, Multi-Agent Orchestration

RAG & Retrieval: Custom RAG, Managed RAG, Pinecone, Vertex AI RAG Engine, Hybrid Search (Dense + BM25), Reciprocal Rank Fusion (RRF), Cross-Encoder Reranking, Clause-Aware Chunking, Vertex AI Search

Evaluation, Observability & Guardrails: Golden Datasets, LLM Evaluation, LLM-as-Judge, OpenTelemetry, Cloud Trace, Cloud Monitoring, Cloud Armor, Cloud DLP, PII Masking, Model Armor

Google Cloud: Vertex AI, Gemini, Document AI, Agent Assist/CCAI, Dialogflow, Cloud Run, Firestore, Cloud Storage, Speech-to-Text

Backend & Architecture: FastAPI, REST, gRPC, Pydantic, Async, Microservices, Clean Architecture, Event-Driven
DevOps: Docker, Git, Pytest

EDUCATION

Cochin University of Science and Technology
B.Tech, Electronics and Communication Engineering

2020 – 2024
CGPA: 8.25 / 10