GAURAV RAJ

Software Engineer | AI Agents, RAG & Python Backends

+91 8969786627 | Bengaluru | graj.work2706@gmail.com | https://www.linkedin.com/in/gaurav-raj-405a96237/ | https://portfolio.gaurav-raj.in/

SUMMARY

Software engineer with 2+ years building RAG pipelines, NL-to-SQL agents and human-in-the-loop agent orchestration in Python and FastAPI, plus .NET 8 microservices. Shipped a RAG assistant for 150+ field engineers and an agentic SDLC workflow used by multiple teams. Owns modules end to end, from architecture to delivery, explaining decisions to clients.

SKILLS

Languages: Python, C#, SQL | **Applied AI:** LLMs, RAG, AI Agents, Agent orchestration, Single-agent and multi-agent architecture, Claude Agent SDK, LangGraph, Tool calling, LLM evaluation | **Backend & Distributed Systems:** FastAPI, API design, Kafka, Message queues, Redis, Scalability, Failure handling | **Databases & Vector Stores:** PostgreSQL, MongoDB (NoSQL), Schema design, Query optimisation | **Cloud & DevOps:** Docker, Kubernetes, Azure AKS, CI/CD

EXPERIENCE

Honeywell - Software Engineer I *Aug 2024 - Present*

- **Delivered under-5-second answers for 150+ field engineers** through cloud and air-gapped RAG pipelines on Vertex AI, Gemini and Ollama (7B-32B models) with pgvector and FAISS.
- **Managed a 4-agent human-in-the-loop SDLC workflow** orchestrated across GitHub Copilot, Claude and Jira MCP, saving ~$6K in one release with 24 stories and 60+ test cases; adopted by multiple teams.
- **Built an NL-to-SQL agent** in Python and FastAPI with few-shot prompting and structured outputs over PostgreSQL nurse-call schemas, so non-technical users generate KPI reports.
- **Lowered failed or stuck message deliveries ~40%** by building fault-tolerant Kafka consumer retries and zero-downtime service updates in core .NET 8 services.
- **Migrated 5+ microservices from MongoDB to PostgreSQL** by redesigning database schemas and using the factory design pattern to abstract data access.
- **Cut dashboard response times from 20 to under 3 seconds** for 50K records using GraphQL filtering and pagination, and offloaded PostgreSQL with Redis caching.
- **Scaled production Docker and Kubernetes workloads** on Azure and GCP with HPA/VPA autoscaling and Prometheus alerting, and piloted GitHub Actions CI/CD that halved release cycles.
- **Architected scalable software across 15+ .NET 8 microservices** on gRPC and Kafka, defining technical architecture for the API Gateway, Rules Engine and Connector services.

Honeywell - Bachelor Intern Software Engineer *Jan 2024 - Jul 2024*

- **Decreased error rate from 36% to 3%** in a 10K-concurrent-user AKS load test (1M requests, Locust) by adding Kubernetes HPA/VPA autoscaling for scalability; earned the Dashing Debut Award.
- **Instrumented 5+ microservices with Prometheus** monitoring and GitHub Actions CI/CD to resolve incidents.

PROJECTS

SupportRAG | Python, FastAPI, Qdrant, Groq, Docker, AKS | *https://www.github.com/Gaurav12341/SupportRAG*

- **Raised answer accuracy 57.5% to 82.5%** for a citation-grounded RAG copilot over 5,400 Kubernetes, PostgreSQL, Kafka and GitHub docs with hybrid BM25 + dense retrieval (RRF, Qdrant), access filters and claim verification.
- **Built an LLM evaluation harness** (108 questions, 29/30 hand-label agreement): 78.9% on a one-shot holdout, all 82 unanswerable questions refused; ablated reranker, HyDE and agentic retry, kept only wins.

ACHIEVEMENTS & RECOGNITION

- **Honeywell Recognitions:** Top Performer of the Month, plus Bronze and Win Together awards for zero-defect releases, SSO modernization and cross-functional delivery.
- **Certifications & Programs:** Six Sigma Green Belt; recent Claude Academy badges (Claude Code in Action, MCP Advanced Topics, Human-Agent Teams); Google Cohort Hackathon.

EDUCATION

SRM Institute of Science & Technology | B.Tech CSE - Software Engineering | CGPA: 9.44 / 10 *2020 - 2024*

DAV Public School, Bokaro Steel City | 10th - PCM with Computer Science | 94.00% / 100 *2017 – 2018*