

MANAS RAI

SOFTWARE ENGINEER | GENAI ENGINEER

rai.manas12@gmail.com | +91 8860190239 | Bengaluru, India | [LinkedIn](#) | [GitHub](#)

PROFESSIONAL SUMMARY

Software Engineer with ~5 years building production backend systems, the last ~2 focused on Generative AI — Large Language Model (LLM) applications, Retrieval-Augmented Generation (RAG) pipelines, and multi-agent systems — built on an earlier foundation in async Python microservices, distributed data pipelines, and semantic search. Owns systems end-to-end: architecture, backend services, cloud infrastructure, and production launch, including a multi-tenant healthcare RAG platform load-tested for 2,000+ concurrent sessions at sub-second latency. Comfortable operating on either track — core backend/API engineering in Python and FastAPI, or agentic/LLM system design — with hands-on depth in vector databases, prompt engineering, and cloud-native deployment on AWS and Azure. Certified across Anthropic's Claude ecosystem: API development, Model Context Protocol (MCP), and Claude Code.

TECHNICAL SKILLS

Backend & Languages: Python, Go (Golang), FastAPI, Flask, Django, REST APIs, Microservices Architecture, WebSockets, Async Programming, SQLAlchemy (ORM), Alembic (DB Migrations), System Design, Software Architecture

Generative AI & LLMs: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), AI Agents, Agentic Architecture, Multi-Agent Orchestration, Prompt Engineering, LangChain, LangGraph, LangSmith, OpenAI API, Anthropic Claude API, Model Context Protocol (MCP), Google Agent Development Kit (ADK), Agent-to-Agent (A2A) Protocol, Embeddings, Vector Databases, NLP, LLM Evaluation, RAGAS, LLM Observability, LLM Cost Monitoring

Data Science & ML: pandas, numpy, scikit-learn, Exploratory Data Analysis (EDA), Feature Engineering, Hugging Face (embedding models)

Frontend: React, JavaScript, HTML5

Databases & Vector Stores: PostgreSQL, MySQL, MongoDB, Neo4j, Snowflake, Aurora, Redis, SQL, NoSQL, Pinecone, pgvector, FAISS, MongoDB Atlas Vector Search

Cloud & DevOps: AWS (EC2, RDS, ECS, S3, SQS, Lambda, Bedrock, Cognito), Azure (Azure OpenAI), Docker, Kubernetes, Jenkins, CI/CD, Git, Datadog, MLOps, Cloud Deployment

Tools & Practices: Claude Code, VS Code, Jira, GitHub, Agile, Scrum

PROFESSIONAL EXPERIENCE

Tech Lead | [Tricon Infotech](#)

Jul 2024 – Present · Bengaluru, India

- **Launched a healthcare chatbot from zero to production** by architecting a multi-tenant RAG pipeline on Azure with tenant-isolated vector stores, secure authentication, and data partitioning — directly enabling the product's first paying customers
- **Engineered an LLM-powered clinical simulation platform** using prompt engineering and RAG to replicate real patient interactions for physician training; load-tested for 2,000 concurrent sessions at sub-second response latency
- **Built a multi-tenant VS Code agent platform** with a meta "maker" agent that generates governed agents, skills, hooks, and prompts — interviewing the developer, planning, and checking new instructions against the existing set before implementing
- **Reduced manual developer workflow time by ~60%** by automating multi-step SDLC tasks (code-review scaffolding, test generation, documentation) through agentic orchestration patterns
- **Shipped a multi-tenant lead-generation data pipeline** with SQLAlchemy and AWS Cognito, enforcing per-account data isolation across 5+ client accounts
- **Built an automated NLP-powered web crawler** that extracts and structures institutional data, powering a qualification-matching recommendation engine — cutting 15+ hours/week of manual research

Solution Leader | [Brane Enterprises](#)

Oct 2021 – Jul 2024 · Hyderabad, India

- **Promoted 3 times in 3 years** (Trainee → Developer → Senior Developer → Solution Leader), the fastest progression in the engineering org at the time
- **Cut data retrieval time by 75% (4x faster)** by implementing semantic search with Neo4j graph traversal and MongoDB vector indexing across 500K+ embeddings and data points
- **Improved real-time audio transmission stability by 35%** by engineering bidirectional WebSocket streaming with adaptive buffering and error recovery for NLP voice applications

- **Designed and deployed async Python microservices** for non-blocking concurrent request handling, increasing API throughput and resolving request-timeout failures seen under production load
- **Built a FastAPI persistence service with JWT authentication** serving as the data backbone for multiple product interfaces, with zero-downtime deployments and CI/CD integration

PROJECTS

DevFlow Kit — *Multi-Agent SDLC Automation* [Project Site](#)

- **GitHub Actions, LangGraph, provider-agnostic (Anthropic/OpenAI/Google), Claude Code** — Turns Jira tickets into production PRs with zero added infrastructure; refinement, implementation, and Jira-sync agents cut the ticket-to-PR cycle from days to hours

RegLens — *Multi-Agent Regulatory Compliance Automation* [Project Site](#)

- **LangGraph, Google ADK, A2A Protocol, RAGAS, LangSmith** — Automated regulatory intelligence system with a compliance research agent, gap analyzer, and report generator, plus a drift-detection evaluation harness (in progress)

LLM Cost Tracking SDK — *Open Source* [GitHub](#)

- **OpenAI, Anthropic, Groq, AWS Bedrock APIs, direct-to-database pipeline** — Self-hosted instrumentation layer for unified LLM cost monitoring with real-time token metering and per-request cost attribution

Cloud Waste Hunter — *Open Source* [GitHub](#)

- **Multi-cloud APIs, human-in-the-loop approval workflows** — Cloud-agnostic resource monitor that detects idle resources across providers via a unified cost-optimization dashboard

CERTIFICATIONS

- Building with the Claude API — Anthropic Academy
- Introduction to Model Context Protocol (MCP) — Anthropic Academy
- Claude Code in Action — Anthropic Academy

EDUCATION

B.Tech., Mechanical Engineering — UPTU, Uttar Pradesh, India 2014

Competitive Examination Preparation & Career Transition 2014 – 2021

Full-time preparation for India's UPSC Engineering Services Examination, followed by a self-directed transition into software engineering and applied AI.

AWARDS

- **Outstanding Performer Award** — **Tricon Infotech** (Dec 2025). Recognized for delivering 2 end-to-end AI products in a single year.
- **Spot Award** — **Tricon Infotech**. For shipping the healthcare chatbot from architecture to production launch.