

Karthik Reddy Basupally

+91 9063007185 - basupallykarthikreddy@gmail.com - [linkedin](#) - [github](#) - [leetcode](#)

PROFESSIONAL SUMMARY

AI Engineer with 2 years at Nomura, building production AI systems in the AI COE. Designed and shipped Chinou's RAG system over 10,000 indexed documents, Agent Mode with MCP integrations, and grounded Web Search using Python, LangChain, and LangGraph. Previously built high-throughput Java backend systems before developing Sonar AutoFix, an LLM-powered CI/CD remediation platform that resolved 40% of SonarQube issues and earned a quarterly award. Experienced in building scalable, production-grade AI systems from infrastructure to deployment.

EDUCATION

VIT University

B.Tech in Electronics & Communication Engineering, CGPA-8.71/10.0

Vellore, Tamil Nadu, India

Sep 2020 — May 2024

TECHNICAL SKILLS

Programming Languages: Python, Java (8–21), SQL, TypeScript

AI / LLM Engineering: RAG, LangChain, LangGraph, AWS Bedrock, LLMs, Agentic AI / AI Agents, MCP, Tool/Function Calling, Prompt Engineering & Templating, Conversational Memory, Context Ranking & Orchestration

Retrieval & Data: Vector Databases (ChromaDB, PGVector, Weaviate), Embedding Models, Sentence Transformers, Semantic Search, Document Parsing & Chunking Strategies, Ingestion Pipelines

Backend & APIs: FastAPI, Spring Boot, RESTful APIs, OpenAPI/Swagger, Microservices, Distributed Systems

Infrastructure & Observability: Docker, GitLab CI/CD, Grafana, MS SQL Server, PostgreSQL, Git, Jira/Confluence

WORK EXPERIENCE

AI Software Engineer

Nomura, Mumbai, India

Jul 2024 — Present

- Built enterprise AI capabilities for **Chinou**, Nomura's internal AI platform, providing access to **AWS Bedrock** foundation models for **10,000–15,000 daily users** with enterprise-grade security, governance, and observability.
- Designed and implemented an enterprise **RAG** pipeline over **10,000+ documents** with personal knowledge bases and shared vector stores — parsing, chunking, embeddings, semantic retrieval, reranking – extended with a grounded **Web Search** capability that sourced live results for time-sensitive queries outside the knowledge base.
- Developed **Agent Mode** using **MCP**, enabling LLM-driven multi-step workflows and tool orchestration by integrating Confluence, Jira, release management, SDLC, financial data platforms, and enterprise web search.
- Built reusable AI components using **LangChain** and **LangGraph**, implementing prompt engineering, conversational memory, context management, and retrieval pipelines for enterprise AI applications.
- Delivered high-performance backend systems using **Java**, **Spring Boot**, and **Apache Camel**, including a regulatory trade persistence microservice processing **60M+ records** (latency reduced from **3 s to 40 ms**) and an integration gateway benchmarked at 462 messages/sec.

PROJECTS

- SonarAutoFix – LLM-Powered CI/CD Remediation (Nomura)** - Conceptualized and built an agentic AI remediation platform using Dockerized GitLab Runners, LLM workflows, and Javalang parsers to analyze SonarQube findings, generate code fixes, and raise Merge Requests for developer review; remediated **40%** of SonarQube issues across enterprise microservices, earned the **Nomura Star / Quarterly Recognition Award**, and led directly to my transfer to the **AI Center of Excellence**.
- CodeRAG – Code-Aware Retrieval-Augmented Generation (RAG)** - Developed a code-aware RAG system that models source code using method signatures, call graphs, static analysis, and vector embeddings instead of flat-text chunking, enabling semantic code search and conversational querying through a FastAPI backend using Python, Sentence Transformers, ChromaDB/PGVector/Weaviate, React, and TypeScript. [GitHub](#)

EXTRACURRICULAR ACTIVITIES

- Earned 5+ skill badges/certifications from industry-recognized platforms (IBM, Cisco, Google, AWS).
- Led and organized weekly coding sessions for the Tech Enthusiasts Club, fostering skill development and a collaborative learning environment.