

VARIKUPPALA RAVINDER

Senior AI Platform Engineer | LLM Agents | RAG | MCP Tooling | System Architecture | AWS
+91 9515295330 | it.ravinder.456@gmail.com | Hyderabad, India | linkedin | Github

CAREER SUMMARY

Senior AI Platform Engineer with 6+ years architecting production LLM systems, multi-agent pipelines, and full-stack distributed platforms. Delivered 40% reduction in manual workflows, 30–40% retrieval accuracy improvement, and sub-2s RAG response latency at enterprise scale. Deep expertise in end-to-end system architecture from agent orchestration and RAG design to frontend delivery and cloud infrastructure.

TECHNICAL SKILLS

AI / LLM: RAG, GPT-4/Claude integration, LangGraph multi-agent orchestration, MCP Servers, Tool & Function Calling, Vector Search, Embeddings, Prompt Engineering, Guardrails & Fallback Strategies

Architecture: Distributed Systems, Microservices, Event-Driven Architecture, API Gateway Patterns, Planner–Executor Agent Design, BFF Pattern, System Design & Review

Backend: FastAPI (Python), Node.js, Express.js, REST & WebSocket APIs

Frontend: React.js, Next.js (SSR/SSG/ISR), TypeScript, Real-time UI (WebSockets), Analytics & Audit Dashboards

Data & Cloud: MySQL, MongoDB, ChromaDB, Apache Kafka, AWS (ECS, Lambda, API Gateway, CloudFormation), Docker, CI/CD (GitHub Actions through OIDC)

WORK EXPERIENCE

Senior Software Engineer | EPAM Systems

2024- Present

- Designed end-to-end system architecture for enterprise AI platforms - defining service boundaries, agent state management, tool routing patterns, and cloud infrastructure across the full stack
- Architected LangGraph planner–executor multi-agent systems with a supervisor coordinating specialised sub-agents for retrieval, Jira automation, analytics, and summarisation
- Built MCP tool servers enabling agents to autonomously create Jira tickets, trigger triage workflows, and invoke internal APIs without human intervention
- Designed RAG pipelines over millions of documents - improving retrieval accuracy by 30–40% and cutting response latency through optimised chunking, metadata enrichment, and vector indexing strategies
- Built real-time agent streaming interfaces, analytics dashboards, and audit trail views with seamless backend integration
- Led development of **Elsa AI Assistant** - flagship enterprise AI platform combining RAG, MCP tooling, and multi-agent orchestration
- Implemented production-grade guardrails, retry logic, and fallback strategies reducing hallucination-related failures in live environments
- Established AI platform engineering standards, conducted architecture reviews, and designed secure CI/CD pipelines on AWS

Specialist – Software Engineering | LTIMindtree

2022- 2024

- Architected data-intensive microservices with clear service boundaries, API contracts, and data flow patterns serving enterprise clients at scale

- Built full-stack features across frontend and backend layers, data transformation pipelines improved throughput by 20–30% and reduced end-to-end latency for downstream consumers
- Led system design reviews and drove production deployments across multiple client engagements

Senior Developer | DevRabbit IT Solutions

2019- 2022

- Delivered full-stack applications end-to-end - owning frontend, backend, and data layer across the complete delivery lifecycle for enterprise clients
- Designed RESTful APIs and reusable component libraries reducing codebase complexity and cutting feature delivery time by ~20%
- Introduced automation scripts for build tooling and data processing, establishing a clean polyglot architecture pattern

PROJECTS

Elsa AI Assistant — Enterprise AI Platform

Production-grade enterprise AI assistant with a real-time streaming frontend, API gateway and MCP layer, and a LangGraph multi-agent core. A supervisor agent orchestrates specialized sub-agents for RAG retrieval (ChromaDB), Jira automation, analytics, and summarization. Features session-aware multi-turn conversation, SSE streaming, audit dashboard, RBAC, and async event ingestion for document updates and agent activity.

Outcomes: 35% reduction in Jira ticket creation time · 40% improvement in query resolution accuracy · Sub-2s RAG response latency

Stack: Python · FastAPI · LangGraph · Node.js · Next.js · ChromaDB · Kafka · AWS ECS · Docker

AI Web Agent for Intelligent Search & Data Extraction

Autonomous web agent for multi-source search and data synthesis, using LLM-based decision workflows and Scrapy-powered data pipelines to generate structured, cited responses.

Stack: Python · LLM APIs · FastAPI · Scrapy · Next.js

Healthcare & Enterprise Platforms

Fleet Manager (IoT monitoring) · Ayurway (EMR system) · iDestination (location intelligence) · Notifii (enterprise package tracking) - full-stack systems with real-time dashboards, REST APIs, and data processing microservices.

ACHIEVEMENTS

- Ace of Delivery Award- EPAM
- Client Hero Award– EPAM
- Fearless Problem Solver Award (3x) – EPAM & LTIMindtree
- A-TEAM SPOT ON Award – LTIMindtree
- Super Crew Award – LTIMindtree

EDUCATION AND CERTIFICATION

Bachelor of Technology (CSE) | Guru Nanak Institutions Technical Campus, JNTU

2019

Diploma (CSE) | Govt Polytechnic Masabtank, SBTET

2016