

SANDHYA H

AI Platform Engineer · Staff Software Engineer

sandhyah23@gmail.com · Bangalore, India · +91 96630 75857
linkedin.com/in/sandhyahonnappa · sandhyahonnappa.dev

PROFILE

AI Platform Engineer with 15+ years in enterprise software, including 4+ years architecting production Generative AI infrastructure. Lead architect of JLL GPT's platform layer, owning LLM gateway design, multi-provider routing, RAG pipelines, and cloud infrastructure for a GenAI platform serving 45,000 monthly users. Deep expertise in LLM orchestration, agentic AI (MCP), caching/resilience patterns, and Azure/AWS platform architecture.

CORE COMPETENCIES

AI & GenAI	LLM Integration (Azure OpenAI, AWS Bedrock, Gemini, Baidu Ernie), RAG Pipelines, Agentic AI, MCP, Claude Agent SDK, LiteLLM AI Gateway, Prompt Engineering & Caching
Cloud & Infra	Azure (AKS, Cosmos DB, Front Door, APIM, App Service, Container Apps), AWS (S3, Lambda, EC2, RDS), Azure DevOps CI/CD
Backend	Python, FastAPI, Flask, ASP.NET Core 8, C#, Clean/Onion Architecture, CQRS (MediatR), SignalR, REST APIs, Node.js/Express
Data	PostgreSQL, pgvector, Azure Cosmos DB, Redis, Qdrant (Vector DB), Datadog, Splunk, Azure Monitor
Architecture	System Design, ADR Authoring, Multi-Provider Strategy Pattern, API Gateway Design, Multi-Region Architecture, Zero-Downtime Deployment, Observability & Monitoring Architecture, Identity & Access Architecture, Enterprise Design Systems, Resilience Patterns, Security & Auth
Frontend	React 18, TypeScript, Next.js, Angular, Redux, Design Systems, Micro Frontend, PWA

WORK EXPERIENCE

Staff Software Engineer — JLL (Jones Lang LaSalle) Apr 2022 – Present · Bangalore

JLL GPT — Enterprise Generative AI Platform (Key Architect & Lead Engineer)

- Own end-to-end AI architecture for JLL GPT, an enterprise GenAI platform used by 45,000 employees globally; designed multi-provider routing (Strategy Pattern + ChatBackendServiceLocator) for zero-downtime LLM hot-swapping.
- Built the Assistant module for custom AI assistants with configurable instructions, model selection, and access control, grounded via an internal RAG service across document uploads, Confluence, and SharePoint.
- Built the Prompt Library and Collections feature for sharing reusable prompts across business units.
- Architected the backend on Clean/Onion Architecture, CQRS, JWT/Okta auth, and Cosmos DB; delivered real-time chat via SignalR.
- Architected an MCP Gateway enabling agentic tool use; authored ADRs and mentored engineers on AI architecture and resilience patterns.
- Built observability into the platform with Datadog and Splunk for monitoring, alerting, and incident triage across services.
- **Recognized with the CoE Shining Star Award and a 2024 Stevie Award (Bronze, Technical Innovation) for JLL GPT.**

LeaseLens AI & Outlook Assistant AI — Contributor / Architecture Advisor

- Built Flask API routes for document upload and LLM-interaction features (LeaseLens AI); advised on frontend architecture and UX for an Office 365 email-assistant add-in, reducing defects and improving adoption (Outlook Assistant AI).

Specialist Front End Developer — L&T Technology Services (LTTS)

2019 – 2022 · Bangalore

- Led architectural design across React and Angular projects, including an IoT web application and an enterprise design system (reusable component library).
- Acted as technical consultant for code reviews, performance troubleshooting, and BDD adoption; championed Storybook for component documentation.
- Recognized as Outstanding Performer (Client Award) and on the Stellar Team of the Year, 2019.

Lead Full Stack Developer — SLK Software

2015 – 2019 · Bangalore

- Led a 15-person team building PayAdmin, an insurance and investment plan system, migrating a legacy payroll system to a modern SPA.
- Named Star Performer of the Year, 2018–2019.

Earlier Career — Full-Stack .NET Developer

2010 – 2015 · Bangalore

Xerox, Tricon Infotech, Apex Decisions — ASP.NET/C# and ASP.NET MVC development, legacy-to-MVC migrations, SQL root-cause analysis, and responsive UI redesigns.

KEY PROJECTS

Nudge — AI Habit Tracker (Personal, Python Backend)

Full-stack app on FastAPI, PostgreSQL, pgvector, and MCP with a multi-repo architecture, JWT auth, and an MCP server for agentic habit logging; uses Relay for LLM access; deployed on Railway.

Relay — Self-Hosted LLM Gateway (Personal, Full-Stack)

Architected a self-hosted LLM gateway platform — unifying auth, cost attribution, and prompt management across a FastAPI/LiteLLM/React stack — designed for enterprise-grade multi-tenant LLM access.

TECHNICAL FOCUS AREAS

- Model Context Protocol (MCP) — JSON-RPC transport, tool/resource/prompt primitives, agentic patterns
- Claude Agent SDK — query() patterns, autonomous ADR-writer and resilience-auditor proofs of concept
- LiteLLM AI Gateway — production AKS deployment, semantic caching, cost governance

EDUCATION

Bachelor of Engineering, Computer Science — Visvesvaraya Technological University, Bangalore
2006 – 2010