THANUSH PK

Nagoya, Japan | +81 09056198331 | thanushpk016@gmail.com| LinkedIn | GitHub| Portfolio

PROFESSIONAL SUMMARY

Applied AI Engineer with 4+ years of robust backend experience (Java/Spring Boot) bridging the gap to the modern GenAI stack. Specializing in designing scalable, LLM-powered agentic workflows, custom RAG pipelines, and fine-tuning open-source models. Adept at leveraging containerized environments (Docker) and advanced orchestration platforms (Dify) to deploy local LLMs and integrate emerging agent protocols (MCP). Passionate about building secure, scalable AI solutions that drive measurable business value.

TECHNICAL SKILLS

AI / GenAI : Large Language Models (LLMs), RAG pipelines, AI Agents, LLM Orchestration (Dify), LoRA Fine-tuning, Prompt Engineering, MCP, AI Gateways (Kong), LLM APIs (Claude, OpenAI, self-hosted open-source models)
Full-Stack : Java, Spring Boot 3, Spring MVC, RESTful APIs, Microservices, Python , JavaScript, TypeScript, React
Databases : MySQL, PostgreSQL, MongoDB, Vector Databases
Tools & Cloud : AWS, Docker, Git, CI/CD, Jira, Postman, IntelliJ IDEA, Maven/Gradle

EXPERIENCE

AI Engineer (Team Lead) | NTT DATA Group Dec 2025 – Present
- Lead a team of AI engineers delivering custom AI agent solutions for enterprise clients in the automotive and manufacturing sectors, accelerating internal data retrieval processes.
- Architected and orchestrated end-to-end LLM workflows using self-hosted Dify and Docker, enabling efficient processing of daily queries through custom RAG pipelines across enterprise knowledge bases
- Integrated emerging GenAI infrastructure including AI gateways, Model Context Protocols (MCP), and local GPU environments, reducing third-party API dependencies and associated costs.
- Drive model fine-tuning efforts (LoRA) to improve tool/function-calling reliability, and mentor team members on AI development and testing practices.

Software Developer | Sonata Software Aug 2022 – Jun 2025
- Architected and maintained Java/Spring Boot backend services for a global enterprise e-commerce platform, managing scalable APIs and improving deployment pipeline efficiency.
- Migrated legacy CMS components to a Micro Frontend (MFE) / Backend-for-Frontend architecture using GraphQL, TypeScript, and headless CMS, improving scalability and maintainability.
- Transitioned complex business logic from legacy monolithic controllers to a Backend-for-Frontend (BFF) layer, enhancing system scalability and reducing technical debt.

PROJECTS

Agentic Application with MCP Tools | Python, Pydantic AI, MCP, JSON Schema 2026
- Built an AI agent using Pydantic AI that orchestrates multiple custom MCP (Model Context Protocol) tool servers to complete multi-step tasks, implemented against the new stateless MCP specification.
- Implemented tool-call authorization and guardrails for safe agent execution, with typed, structured agent workflows validated through Pydantic schemas.

Fine-Tuning a Small LLM for Tool and Function Calling | QLoRA, Llama-3.2-1B, Hugging Face, PyTorch 2026
- Fine-tuned Llama-3.2-1B-Instruct with QLoRA on the Salesforce xLAM dataset, correcting malformed tool-call output into consistent, parseable function calls, with a documented before/after evaluation baseline

RAG vs OKF Retrieval Benchmark | Python, Ollama, FAISS, nomic-embed-text, phi3 2026
- Built an end-to-end benchmark comparing embedding-based RAG vs Google's OKF over a shared 20-document knowledge base on local infrastructure, with a reusable evaluation harness: curated golden dataset, retrieval metrics, and pluggable LLM-as-judge scoring.

EDUCATION

Master of Computer Application (MCA) 2020 – 2022
St Joseph Engineering College
Bachelor of Computer Application (BCA) 2017 – 2020
Field Marshal K. M. Cariappa College

CERTIFICATIONS & LANGUAGES

AWS Certified AI Practitioner (AIF-C01) July 2026
Languages : English (fluent, professional); Japanese (JLPT N3 — working proficiency in a Japanese enterprise environment)