

Tanmay Jain

Senior AI Engineer

+91-999-319-2485 | tanmayjain717@gmail.com | linkedin.com/in/tj07 | tjdev.in
Indore, Madhya Pradesh, India

PROFESSIONAL SUMMARY

Senior AI/ML Engineer with **4+ years** building enterprise-grade AI systems at scale — from multi-agent orchestration and RAG pipelines to computer vision and voice commerce. Specialized in **LangChain, LangGraph, Google ADK, FastAPI, and React**, with production deployments on **AWS, GCP, and Azure** serving 1,000+ concurrent users at sub-second latency. Expert in **MCP server development, business process automation (Camunda BPM)**, and **AI-assisted development tools** (GitHub Copilot, Claude Code). Proven track record of shipping end-to-end AI solutions in retail, e-commerce, and enterprise SaaS — including multilingual agentic chatbots, intelligent invoice processing, business process automation platforms, and real-time analytics systems.

TECHNICAL SKILLS

AI/ML Expertise: LLMs, RAG, Multi-Agent Systems (LangChain, LangGraph, Google ADK), Prompt Engineering & Caching, Computer Vision (OpenCV, DLib, MediaPipe, PoseNet), NLP, Vector Embeddings, Enterprise Knowledge Graphs
GenAI & LLM Tools: LiteLLM, Langchain, LangGraph, Claude API, Gemini API, Anthropic SDK, OpenAI SDK, Prompt caching & optimization
AI Agent Development: MCP (Model Context Protocol) server development, GitHub/Jira/Figma MCP integrations, Claude Code & GitHub Copilot integration, Custom agent tools for AI-assisted development
ML Frameworks: TensorFlow, PyTorch, Hugging Face Transformers, scikit-learn
Programming Languages: Python, JavaScript, TypeScript, SQL
Backend Technologies: FastAPI, Node.js, Express.js, RESTful APIs, WebSockets, Microservices Architecture
Frontend Technologies: React.js, Next.js, TypeScript, Tailwind CSS, Redux, Material-UI, Responsive Design
Databases & Storage: Vector Databases (Qdrant, Pinecone, Chroma), Firestore, PostgreSQL, DynamoDB, SQLAlchemy, MongoDB
Cloud Platforms: AWS (Textract, Lambda, S3, EC2, IVS, API Gateway), GCP (Vertex AI, Vertex AI Search, Agent Development Kit, Cloud Run, Firestore, GCS, Artifact Registry, Cloud Functions, Gemini API), Microsoft Azure (Azure OpenAI, Cognitive Services, Container Instances)
Workflow Automation & BPM: Camunda BPM, Business process orchestration, LangGraph workflow patterns, Event-driven architecture
DevOps & Tools: Docker, Terraform, Git, GitHub Actions, CI/CD Pipelines, Claude Code, GitHub Copilot, Linux

PROFESSIONAL EXPERIENCE

Senior AI Engineer

Nagarro Software

February 2025 – Present

Indore, India

- Architected and deployed **enterprise-grade AI/ML solutions** focusing on business process automation, achieving **40% reduction** in manual workflow processing time
- Designed **AI Expense Management System** with intelligent invoice processing using AWS Textract and custom ML models, achieving **90%+ categorization accuracy** and sub-second response times for 10,000+ monthly transactions
- Built comprehensive **business process automation platform** integrated with Camunda BPM workflows, streamlining operations for 500+ enterprise users across multiple departments
- Developed **GlowBuddy** computer vision platform using DLib and OpenCV with custom facial recognition algorithms, real-time image processing pipelines, and explainable AI recommendation engine serving 5,000+ users
- Designed and developed **custom MCP (Model Context Protocol) servers** including GitHub and Jira integrations for Claude Code, enabling AI-assisted development workflows with private guidelines, organizational information tools, and issue tracking automation
- Integrated **GitHub Copilot and Claude Code** into development cycle for AI-augmented coding, improving team productivity and code quality through intelligent code completion and suggestions
- Architected **prompt caching strategy** separating static instruction (large, cacheable, amortized) from dynamic per-request context (small, injected at runtime) to optimize latency and token costs
- Implemented **enterprise RBAC security frameworks** across multiple applications, ensuring SOC 2 compliance and data governance standards
- Designed **scalable microservices architecture** supporting thousands of concurrent users while maintaining **sub-second latency** and 99.9% uptime

AI Engineer

Nagarro Software

July 2023 – February 2025

Indore, India

- Engineered **LINE-integrated Multilingual Agentic Chatbot** deployed on Google Cloud Run with JWT-protected FastAPI backend, routing queries through a **Gemini** powered orchestrator agent built on Google ADK
- Designed **single-agent / multi-tool architecture** using Google ADK's FunctionTool with 5 specialized tools – Thai/English product search, Thai/English promotions search, and FAQ search – selected dynamically based on query intent
- Built **Knowledge Base Management System** with full CRUD for PDFs/HTML via GCS resumable uploads, Vertex AI Search ImportDocumentsRequest (sync/async LRO), smart incremental promotions sync with zero-downtime orphan cleanup, and signed URL generation for secure in-browser PDF preview
- Developed **SmartDashboard AI Chatbot** on Cloud Run Gen2 with RS256/JWKS JWT auth, OpenAPIToolset auto-discovery, bilingual Thai/English routing, Gemini **prompt caching**, and Firestore async session persistence
- Built **Voice Commerce and OCR solution** for Fortune 500 clients (healthcare/wellness) with voice-enabled product ordering, hot word detection, and multi-level variant selection, processing **10,000+ orders monthly**
- Designed **vector database solutions** using Qdrant for semantic search, integrating multiple data sources with 300ms average query response time
- Created conversational registration assistant deployed across LINE, mobile, and web platforms, reducing registration time by **60%**

Associate Engineer

Nagarro Software

February 2022 – July 2023

Gurugram, India

- Architected **Live Commerce Platform** combining real-time video streaming with integrated e-commerce using AWS Interactive Video Services, supporting **1,000+ concurrent users** with sub-second latency
- Developed interactive chat system using WebSockets with live shopping integration, real-time product showcase, and synchronized video-commerce capabilities
- Built comprehensive admin dashboard for stream management, analytics, multi-language support, and cross-platform compatibility (web, mobile, OBS)
- Developed robust RESTful APIs and efficient socket connections enabling real-time functionality for 50+ microservices

KEY PROJECTS

SmartDashboard AI Chatbot | *Python, Google ADK, Gemini, GCP, Firestore*

- Built FastAPI backend on Cloud Run Gen2 with RS256/JWKS JWT auth, routing ABO queries through a Gemini orchestrator using **OpenAPIToolset for zero-code tool auto-discovery** from an OpenAPI 3.1 spec
- Enriched tool schemas with vendor extensions (**x-workflow-chain**, **x-ai-patterns**, **x-business-rules**) enabling autonomous **multi-step API workflow orchestration** by the LLM without code changes
- Optimized token costs via **Gemini explicit prompt caching** with stale-cache auto-recovery; integrated Firestore async session persistence and fire-and-forget push notifications with exponential-backoff retry and dead-letter fallback

LINE Multilingual Agentic Chatbot | *Python, Google ADK, Vertex AI Search, Firestore, GCS, GCP*

- Developed **single-agent / multi-tool architecture** with 5 FunctionTools for bilingual (Thai/English) product, promotions, and FAQ search using separate Vertex AI Search datastores per language and domain
- Built **hybrid relevance scoring** combining semantic similarity, keyword match, and position rank with dynamic drop-factor filtering; integrated Vertex AI Discovery Engine for **citation-backed FAQ answers**
- Implemented **bilingual (Thai/English) search routing**, Firestore session storage, GCS document syncing, and HTML-to-plaintext preprocessing (html2text) to improve Vertex AI Search indexing relevance

Knowledge Base Management System | *Python, FastAPI, Vertex AI Search, GCS, Firestore*

- Built **document ingestion pipeline** for PDFs and HTML files via GCS resumable upload sessions, published into Vertex AI Search QA datastores using ImportDocumentsRequest with JSONL metadata in sync and async LRO modes
- Implemented **full CRUD** for product catalogs (ProductDatastoreService) against Vertex AI Search with upsert support and auto-enrichment of product links
- Designed **smart incremental promotions sync** (PromotionsSyncService) using INCREMENTAL reconciliation mode with automatic orphaned-document deletion for zero-downtime updates; exposed admin REST endpoints with signed URL generation for secure in-browser PDF preview

AI Expense Management System | *Python, LangChain, LangGraph, AWS Textract, FastAPI, React*

- Engineered intelligent invoice processing with **90%+ automated categorization accuracy**, reducing manual processing time by 70%
- Implemented LangGraph conversational AI assistant for natural language expense queries with persistent context management
- Built comprehensive analytics dashboard with real-time reporting, trend analysis, and predictive expense forecasting

GlowBuddy – AI Skin Analysis Platform | *Python, DLib, OpenCV, FastAPI, React, AWS*

- Architected custom facial recognition algorithms achieving **92% detection accuracy** across diverse skin types and conditions
- Designed real-time image processing pipelines analyzing multiple skin parameters simultaneously with 200ms latency
- Built explainable AI recommendation engine providing transparent reasoning for personalized skincare suggestions

Live Commerce Platform | *AWS IVS, React, Node.js, WebSockets, DynamoDB, Tailwind CSS, OBS*

- Architected and developed comprehensive live commerce platform combining real-time video streaming with integrated e-commerce, achieving **sub-second latency** and supporting **1,000+ concurrent users**
- Implemented **interactive WebSocket-based chat system** with real-time message delivery, supporting high-concurrency streaming events and integrated live shopping purchase capabilities
- Built **live shopping integration** with real-time product showcase, dynamic pricing, and one-click purchase flow; created comprehensive admin dashboard for stream management, analytics, and multi-language support across web, mobile, and OBS integrations

EDUCATION

Acropolis Institute of Technology and Research

Bachelor of Technology in Information Technology

Indore, India

2018 – 2022

RECOGNITION AND AWARDS

- Recognized as an ‘**Extraordinary Role Model**’ consistently for 4+ years (top 1% of Nagarro employees) for consistently exceeding performance expectations, demonstrating technical excellence, and delivering high-impact solutions
- Contributed to product development showcased at the AWS R&D event, representing the organization’s innovation in AI/ML and real-time commerce systems