Tanish Rajput

Noida,UP
+91-9690190921 Email Linkedin GitHub

Education

Noida Institute of Engineering and Technology **2021 – 2025**
B.Tech in Information Technology ***CGPA: 9.01*** *Greater Noida, UP*

Experience

QualtechEdge **Nov 2025 – Present**
AI Engineer *Noida*

- **Designed** a **LangGraph**-based react-agent **voice loan advisor** using **LiveKit**, **LLMs**, and **FastAPI**, enabling end-to-end automated loan journeys.
- **Built and integrated** multiple middleware and external APIs, including government and partner APIs, with robust **fault tolerance**, **retry logic**, and validation mechanisms.
- **Developed** the agent's complete tool ecosystem, containerized services with **Docker**, and deployed scalable workflows on **AWS**.
- **Collaborated** closely with product and engineering teams to optimize system architecture, improve response accuracy, and enhance overall agent reliability.

Microsoft **June 2024 – August 2024**
Software Engineer Intern *Noida*

- **Integrated** an **audio panel** into the **Microsoft Designer** editor using **React** and **FluentUI v9,** enabling **audio content creation** and **MP4 export** functionality.
- **Redesigned** the panel UI based on **UX feedback** reducing **user friction** by **20%** and improving overall **usability.**
- **Collaborated** with **Design** and **Accessibility (A11y)** teams to enhance **interface performance** and ensure **WCAG-compliant accessibility.**

Projects

AuthoGraph AI

- **Built** a production-grade **multi-agent AI system** using **LangGraph** implementing a Plan–Execute–Distribute workflow to autonomously research, generate, and publish technical blogs to Dev.to, Hashnode, LinkedIn, and Medium in under 5 minutes.
- **Designed** a real-time **RAG pipeline** with **Tavily Search**, vector retrieval, and **LangSmith** monitoring, achieving 95%+ factual grounding through automated LLM-based evaluation.
- **Developed** a scalable full-stack platform using **FastAPI + React**, streaming agent reasoning via **SSE** and integrating **GPT-4o-mini** and **Gemini 2.5 Flash** for content and visual generation.
- **Deployed** on **AWS EC2**, scaling to 25+ active users while implementing a **credit-based monetization system** via Razorpay and database-driven task cancellation logic to optimize inference cost and resource utilization.

OmniBrief

- **Developed** an autonomous intelligence engine using **LangGraph** and **GPT-4o** to curate, rank, and synthesize high-signal technical developments from **ArXiv, GitHub, and 30+ RSS feeds**.
- **Implemented** a state-managed multi-agent pipeline featuring a **Critic-in-the-loop node** for automated quality control and **Playwright-based scraping** to enrich technical summaries.
- **Architected** a production-ready delivery system with **FastAPI** and **PostgreSQL**, optimizing inference costs via a **dual-model routing strategy** (GPT-4o/mini) and custom token-usage auditing.

AI & Technical Skills

Programming: Python
AI & LLM Systems: RAG Pipelines, Multi-Agent Systems, Prompt Engineering, LangChain, LangGraph, CrewAI
ML Frameworks & Libraries: PyTorch, Hugging Face Transformers, NumPy, Pandas
Model Training & Optimization: LoRA, QLoRA, PEFT, Fine-Tuning
Vector Databases: FAISS, ChromaDB, Pinecone
Databases: PostgreSQL, MongoDB
Backend & APIs: FastAPI, Async Programming
Cloud & Deployment: AWS (EC2, S3, ECS), GCP (Compute Engine), Docker
Developer Tools: Git, GitHub, GitHub Actions