Paritosh Sharma

Senior AI Engineer | Generative AI Workflow Architect

paritoshsharma072000@gmail.com | linkedin.com/in/paritosh-sharma | github.com/paritosh0707 |
+91 8427176868

Crafting robust GenAI platforms with LangChain, LangGraph, and FastAPI — focused on agentic orchestration and developer enablement.

Senior AI Engineer with 3+ years of experience designing and delivering enterprise-grade NLP and LLM-based systems, combining agentic multi-LLM workflows with deep expertise in LangChain, LangGraph, RAG, and prompt engineering.

Specialized in building end-to-end platforms using transformer fine-tuning, vector-based retrieval, and scalable backend services with FastAPI, Docker, and Kubernetes.

Led the development and deployment of an LLM-powered QA automation platform—managing architecture, implementation, and team coordination—while integrating cloud-native MLOps (Azure, AWS) to drive intelligent automation across UI, API, and data validation layers.

Skills

- **GenAI & AI Agents:** LangChain, LangGraph, RAG, Agentic RAG, Tool Calling, LangSmith
- **LLMOps & MLOps:** Fine-tuning, PEFT, QLoRA, MLflow, DVC, Airflow, BentoML
- **Cloud & Infrastructure:** FastAPI, Docker, Kubernetes, Azure OpenAI, AWS Bedrock, Azure DevOps
- **Programming:** Python, SQL, PySpark, Git, CircleCI
- **ML & NLP:** Transformers, NER, Sentiment Analysis, Scikit-learn, PyTorch
- **Soft Skills:** Team Leadership, Mentorship, Collaboration, Technical Writing

Experience

Senior AI Engineer at Incedo Solutions Ltd **Oct 2024 — Present**

- Orchestrated the scale-up and enterprise rollout of an LLM-powered QA automation platform across **3 major clients**, streamlining test creation, validation, and reporting across UI, API, and data pipelines.
- Directed a cross-functional engineering team of 5 and partnered with leadership to architect platform roadmap, drive delivery velocity, and ensure scalable GenAI integration into client SDLCs.
- Delivered a production-ready, multi-agent **User Story Intelligence system** that parsed 50-page PRDs into clustered epics and sprint-ready stories; generated **180+ user stories in under 3 minutes**.
- Engineered an agentic **API Testing framework** with support for OpenAPI, Postman, and legacy SOAP; automated test generation and batch execution with **850+ test cases generated in under 2 minutes**.
- Spearheaded a full-stack **UI Testing pipeline**, including dynamic test generation and codebase-aware artifact reuse via an in-house **MCP (Model Context Protocol) server**; improved productivity by **83% (manual)** and **78% (automation)**.
- Developed and deployed a **multimodal QA chatbot** supporting documents, screenshots, and contextual queries with long-memory tracking and semantic context handling.

- Launched a no-code **Agent Builder Platform** enabling non-AI engineers to assemble LangGraph workflows via drag-and-drop interface with self-improvement loops, retry policies, and shared toolsets.
- Architected a reusable internal **AI Framework** to standardize GenAI infrastructure across the org, including:
  - **RagTune** – Configurable RAG engine with retriever routing and fallback pipelines.
  - **FineTune** – Modular LLM fine-tuning toolkit with PEFT, QLoRA, and custom training lifecycle APIs.
  - Embedding validators, agent tracing utilities, and vector index managers.
  - Unified via **MCP servers** for internal reuse via NL-based interface calls.
- Championed developer enablement by documenting prompt patterns, agent chaining logic, and RAG tuning recipes; reduced onboarding time for new AI engineers by streamlining internal knowledge assets.
- Liaised with product managers and QA leads to align GenAI-driven test pipelines with domain-specific QA strategy—ensuring adoption across technical and non-technical teams.

AI Engineer at Incedo Solutions Ltd **Jul 2022 — Oct 2024**

- Conceptualized and architected the company's first LLM-powered QA automation platform from scratch; implemented scalable backend APIs using FastAPI, Docker, and Kubernetes.
- Engineered the initial **test case generation engine** using Corrective RAG with hierarchical embeddings; integrated JIRA workflows to generate positive, negative, and edge case scenarios, and connected outputs to TestRail.
- Prototyped early LangChain agents for document parsing, prompt chaining, and semantic story mapping—laying the foundation for agentic workflow orchestration.
- Developed the initial **Data Testing module**, generating PySpark and SQL transformation logic from mapping specs; implemented auto-mapping, record comparison, and end-to-end validation reporting.
- Built the first version of the **API Testing engine**, capable of parsing Swagger/Postman inputs, auto-generating payloads, and executing one-click test runs with structured test result logs.
- Piloted and delivered the platform's first enterprise POC, achieving a **43% reduction in manual QA effort within 2 sprints** and securing leadership approval to scale across client projects.

Leadership & Technical Initiatives

- Built and maintain a technical blog focused on ML, Deep Learning, and NLP fundamentals with math explanations; breakdowns used by learners and professionals across India. *(paritosh0707.github.io/paritosh-tech-journal)*
- Mentored over **50+ students** in the AI field across two years, providing structured guidance on projects, roadmaps, and conceptual depth in NLP and Generative AI.
- Built a personal portfolio platform to showcase AI and GenAI projects with detailed case studies, workflows, and tooling; integrates live demos and GitHub code. *(paritoshsharma.dev)*

Education

BE in Computer Engineering, Thapar Institute of Engineering and Technology, Patiala, **2018 — 2022**