PRATIK SONTAKKE

AI Engineer | +91 8888648824 | pratikass488@gmail.com | Linkedin | Github | Portfolio | YouTube

Professional Summary

Generative AI Engineer specializing in production LLM applications, RAG systems, and AI agents using LangChain and LangGraph. Deployed systems serving 5,000+ daily users with Python, FastAPI, and AWS. Expert in building scalable generative AI workflows backed by robust backend engineering.

Technical Skills

AI & Machine Learning: LLMs, RAG, AI Agents, Fine-Tuning (PEFT), LangChain, LangGraph
Backend & Databases: Python, Java, FastAPI, Spring Boot, SQL, Vector DBs (Chroma, Pinecone), PostgreSQL, Redis
Cloud & DevOps: AWS, Docker, Kubernetes (k8s), Terraform, GitHub Actions, Vercel

Projects

Multi-Tenant RAG-as-a-Service Platform |

- Engineered a RAG SaaS platform **projected to save clients over $20,000 and 1-2 months in development costs** by providing a ready-to-use, production-grade conversational AI.
- Architected a secure, multi-tenant backend on AWS **to guarantee complete data isolation and build client trust**, using a schema-per-tenant model in PostgreSQL and a fine-grained RBAC system.
- Collaborated on a high-performance FastAPI API, **enabling real-time, context-aware responses that ensure a seamless and interactive user experience.**

Professional Experience

Senior Software Engineer AI - 5C Network Sept 2025 - Present | Bengaluru, India

- Developing an end-to-end workflow tool that automates model training from annotated datasets, enabling rapid, repeatable experiments and analysis for ML team leads.
- Engineered an AI agent to automatically verify and validate over 5,000+ radiology reports daily, decreasing manual review time by 30 % and enhancing report reliability.
- Developed and deployed a RAG model for the company's landing page to automate user query responses, reducing the workload for the technical support team.

Cloud & AI Backend Engineer - Freelance Nov 2024 - Jun 2025 | Remote

- Engineered a Text-to-SQL agent using LangChain and integrated it into a Spring Boot service, empowering non-technical teams with self-service analytics and saving 5-10 hours weekly.
- Architected core cloud infrastructure on AWS using Terraform and established a full CI/CD pipeline, slashing service deployment times from hours to under 10 minutes.
- Developed and containerized key Java/Spring Boot microservices to serve AI models, ensuring a resilient and responsive backend system capable of handling production traffic.

Software Engineer - Guenstiger Apr 2023 - Oct 2024 | Delhi, India

- Engineered an AI-driven system to auto-generate hundreds of SEO-optimized product pages, boosting the company's visibility and ranking on Google
- Led backend enhancements that improved system throughput by 15 %, supporting a growing user base without additional infrastructure costs.

Junior Technical Consultant - Edifition Aug 2021 - Apr 2022 | Remote

- Translated client business needs into detailed technical specifications for 5+ projects, leading to a 15 % reduction in requirement-related change requests post-development.

Education

AI Engineering
MisogiAI By Masai, Bengaluru, India Jun 2025 - Sept 2025

Bachelor of Science (B.Sc), Computer Science
Sant Gadge Baba Amravati University, Maharashtra, India Apr 2017 - Nov 2020