Sarthak Agrawal

AI/ML Engineer (Agentic Systems)
sarthak.agrawal1311@gmail.com

+91-8871928567
linkedin.com/in/sarthakagrawal11
iamme1311.github.io/Portfolio

SUMMARY:

AI/ML engineer focused on designing agentic AI systems that autonomously diagnose, reason, and act in production environments. Experienced with tool-calling LLM agents, retrieval pipelines, structured reasoning loops, evaluation frameworks, and backend infrastructure using Python, FastAPI, LlamaIndex, and LangChain.

PROFESSIONAL SKILLS:

GenAI & Agentic Systems:

• LLM APIs (OpenAI, Gemini, Llama), prompt engineering, vector search, embeddings.

• Tool-calling agents, secure tool execution, structured reasoning, validation loops.

• RAG pipelines, knowledge curation, evaluator functions, guardrails & safety.

Cloud & DevOps:

•AWS (Lambda, S3, ECS, EKS, RDS, DynamoDB, CFT, CloudWatch), Docker, CI/CD, GitHub Actions.

ML Engineering:

• Model evaluation, monitoring, dataset curation, benchmark design.

• Latency optimization, caching, batching, model routing.

Backend Engineering:

• Python, FastAPI, Async programming, PostgreSQL, SQLAlchemy.

• System design, performance optimization, API security & RBAC.

WORK EXPERIENCE:

AI Engineer

RippleLinks, Bengaluru, India - June 2024 - Present

- Designed and deployed **agentic AI systems** enabling autonomous task execution (tool calling, routing, and multi-step reasoning) to replace manual workflows across operations.
- Designed backend services using **FastAPI + PostgreSQL + async workers**, enabling scalable orchestration of agents, tool APIs, and long-running tasks.
- Built **secure REST APIs** with OAuth2/JWT authentication, RBAC authorization, and request guardrails to ensure safe AI and data access.
- Implemented **agent action pipelines** with queue-based background execution, context caching, and self-correction loops, improving task throughput and reliability.
- Operationalized multiple AI providers (OpenAI / Ollama) in a **cost-optimized inference layer** with fallback, routing, and model selection policies.
- Delivered full **observability** via structured logging, traces, and performance metrics (OpenTelemetry + Grafana) resulting in faster debugging and reduced downtime.
- Containerized microservices using **Docker** and enabled **zero-downtime CI/CD deployments** with GitHub Actions, improving release frequency and stability.
- Optimized infra footprint and performance via async I/O, connection pooling, and request batching, reducing latency and compute costs.

Data Science Intern

Innomatics Research Labs, Hyderabad, India - January 2024 - April 2024

- Developed RAG pipeline using LangChain + Gemini with retrieval, prompt engineering, and validation loops.
- Built a semantic subtitle search engine using NLP + embeddings.
- Conducted ML experiments on 8.5k+ reviews dataset improving classification accuracy.

EDUCATION:

Masters in Computer Applications - AI & ML - January 2022 - January 2024

Greater Noida, India

LANGUAGES:

English, Hindi, Tamil