ADITYA

Generative AI / Agentic AI Engineer

aditya78332412@gmail.com | +91-7833934588 | Noida, India | linkedin.com/in/adityasanjay1999 | github.com/aditya2425

aditya2425.github.io/Portfolio

PROFESSIONAL SUMMARY

Generative AI / Agentic AI Engineer with 3+ years at Accenture, delivering production AI systems for Fortune 10/100 clients across healthcare, aviation, and entertainment. Designed a multi-agent failure triage system cutting resolution time from 1–2 days to under 10 minutes; architected a real-time healthcare call assistant processing thousands of live utterances daily with sub-second latency. Deep expertise in RAG pipelines, multi-agent orchestration, LLM fine-tuning, AI safety guardrails, and end-to-end GenAI system ownership — from data ingestion and retrieval to production deployment and monitoring. Adept at collaborating with cross-functional teams across engineering, QA, and product to align GenAI solutions with enterprise workflows.

TECHNICAL SKILLS

GenAI & LLMs: OpenAI GPT-4o, Azure OpenAI Service, Anthropic Claude, Mistral 7B, Hugging Face Transformers, Prompt Engineering (CoT, Few-Shot, ReAct), RAG, Multi-Agent Systems, Agentic AI, LLM Fine-Tuning (QLoRA, LoRA, PEFT, RLHF, SFTTrainer), Multimodal RAG, Embeddings

Agentic AI & Orchestration: LangChain, LangGraph, CrewAI, Semantic Kernel, Azure AI Foundry, Multi-Agent Orchestration, Tool Calling, Function Calling, MCP Servers (Model Context Protocol), ReAct Pattern, Human-in-the-Loop

RAG & Vector Search: Azure AI Search, Azure Cognitive Search, Pinecone, ChromaDB, BM25, Hybrid Search (Keyword + Vector), Cohere Reranking, Vector Embeddings, Semantic Search, Reciprocal Rank Fusion (RRF)

AI Safety, Evaluation & MLOps: LLM Evaluation (RAGAS-style, BERTScore, BLEU, ROUGE, LLM-as-Judge), Guardrails, Content Filtering, PII Detection, Prompt Injection Prevention, Model Monitoring, Drift Detection, A/B Testing, LLMOps, GitHub Actions CI/CD

Cloud & Infrastructure: Microsoft Azure (AI Foundry, OpenAI Service, Functions, App Service, Event Hub, Logic Apps, Blob Storage, Key Vault), Docker, Docker Compose, FastAPI, Health Checks, Security Scanning, Azure DevOps

Backend & Data: Python, SQL, Flask, Cosmos DB, Azure SQL, SQLite, REST APIs, OpenAI API, Streamlit, Gradio, Pydantic, Git, pytest, Event-Driven Architecture, Microservices

PROFESSIONAL EXPERIENCE

Software Engineer – Generative AI Jul 2024 – Present | Gurugram,IN
Accenture Solutions Pvt Ltd

- Designed and deployed a **multi-agent AI system** for a global airline technology provider using Azure AI Foundry and Semantic Kernel — automating failure triage, test generation, and regression scoping, reducing resolution time from **1–2 days to under 10 minutes**
- Built Hybrid **RAG pipelines** with deduplication and human-in-the-loop workflows, serving 50+ internal users with **90%+** accuracy and sub-60s response, reducing L2 escalations by **~35%**
- Engineered **agentic test case generation** with regression impact prediction and risk scoring (1–5), processing **500+** daily test scenarios via automated data ingestion into Cosmos DB
- Collaborated with cross-functional teams across engineering, QA, and product to align GenAI solutions with enterprise workflows and delivery milestones

Associate Software Engineer Jan 2023 – Jun 2024 | Gurugram,IN
Accenture Solutions Pvt Ltd

- Architected a **real-time AI assistant** for a Fortune 10 healthcare company processing **5,000+ live call utterances** daily using Flask, Azure OpenAI, and Cognitive Search with sub-second latency, supporting 200+ agents
- Built production APIs for NLP identity verification, **real-time intent detection (CLU + GPT fallback), confidence-scored auto-suggestions, and post-call summarization** — handling 10K+ API calls/day with **99.5%** uptime
- Designed **event-driven pipeline** (Event Hub → Functions → Cosmos DB → Flask → React) with multi-threaded execution and feedback loops, reducing end-to-end processing latency by **40%**
- Delivered a large-scale GenAI application for a Fortune 100 entertainment company using GPT-4o — automating document generation, competency mapping, and interview guides, reducing manual effort by **~70–80% across 15+ departments**
- Engineered specialized prompt workflows with context injection and inclusive language detection across 8 bias categories, achieving **95%+** compliance on document validation

PROJECTS

DocuMind — Production RAG Pipeline [GitHub]
Python, OpenAI, ChromaDB, Pinecone, FastAPI, Docker
- Built production **RAG pipeline** with 3 chunking strategies, **hybrid search (dense + BM25)**, Cohere reranking, and multi-LLM cited answer generation
- Developed RAGAS-style evaluation with 6 preset experiment configurations and interactive results dashboard

AgentForge — Multi-Agent Research System [GitHub]
Python, LangGraph, CrewAI, Tavily, ChromaDB, Streamlit
- Designed **5-agent** research pipeline with reflection loops in both LangGraph and CrewAI to benchmark framework trade-offs
- Integrated guardrails for **prompt injection, PII masking, hallucination scoring, and loop prevention**

ModelForge — LLM Fine-Tuning Pipeline [GitHub]
Python, HuggingFace, QLoRA, PEFT, BERTScore, Streamlit
- Fine-tuned Mistral 7B using QLoRA on 500+ code review examples with SHA256 and MinHash deduplication
- Built evaluation dashboard comparing 4 models across **BLEU, ROUGE, BERTScore, and LLM-as-Judge** with cost breakeven analysis

GuardianAI — LLM Safety & Observability Platform [GitHub]
Python, SQLite, Streamlit, Plotly
- Built **decorator-based** LLM tracing capturing latency, tokens, cost, and 4 quality metrics across 9 model pricing tiers
- Implemented drift detection, configurable alert rules, LRU cache, and A/B testing framework

ShipAI — ML Deployment & DevOps Pipeline [GitHub]
Python, Docker, GitHub Actions, FastAPI, CI/CD
- Created CI/CD pipeline with GitHub Actions, multi-stage Docker builds, and security scanning for dependencies and Dockerfiles
- Designed 3-tier health checks and deployment runbooks with step-by-step rollback procedures

EDUCATION

B.Tech in Computer Science 2018 – 2022 | Noida,UP
Amity University, Noida

ACHIEVEMENTS

Sparkling Star Award 2024
Recognized for building multi-agent GenAI system adopted across enterprise workflows — Accenture & Avanade

ACE Award (Team) 2023
Awarded for building and deploying a real-time healthcare AI assistant handling 5,000+ daily utterances with sub-second latency and 99.5% uptime

CERTIFICATIONS

- Microsoft Azure AI Engineer (AI-102)
- Microsoft Azure Developer (AZ-204)
- Microsoft Azure AI Fundamentals (AI-900)
- Microsoft Azure Fundamentals (AZ-900)
- AWS Cloud Practitioner
- MongoDB Certified