Akshath Tiwari – AI Engineer

tiwariakshath@gmail.com | 8185983094 | github.com/akshathtiwari | linkedin.com/in/akshathtiwari/ | Hyderabad, India

Summary

Developer with 2+ years of hands-on experience architecting and building backend systems for web applications. Proficient in Python, Machine Learning, Generative AI, with experience in designing scalable inference pipelines and RESTful APIs using Fast API. Proven track record of integrating LLM capabilities in production environments, optimizing model performance, and delivering high-throughput data retrieval solutions that drive actionable insights and improve system efficiency.

Education

Manipal University Jaipur – B.Tech **Jaipur**

July 2019 – July 2023

Relevant Coursework: Data Structures and Algorithms, Machine Learning, Statistics & Mathematics, Generative AI, DevOps, Prompt Engineering, LLMs (Large Language Models)
Core competencies: Backend Engineering**,** Solution Architect**,** Generative AI, NLP, Machine Learning
Proffesional Certification: Google Cloud Professional Machine Learning Engineer Certification, Medallia Technical Developer

Skills

Programming Languages: Python, JavaScript, C++
Developer Tools: Dbeaver, Postman, Git, Docker, Kubernetes, ArgoCD
Database: Postgres, Qdrant, Milvus, Pinecone
Technology: Fast API, AutoGen, SQL Alchemy, LangChain, LangGraph, PyTorch
Project Management: Jira, Agile, MS Office
Soft skills: Leadership, Effective communication, collaborator, crisis management, problem solving.
Others: Understanding of HTML, CSS, UX and Figma Design

Work Experience

Deloitte India **Hyderabad**
AI Engineer *Aug 2023 – Present*

- Developed "Work Analyzer", web-based application which helps organization to explore automation opportunity for the repetitive tasks performed by their workforce, estimate Full-Time Equivalent reduction opportunities, and quantify cost savings.
- Implemented a **RAG pipeline** supporting PDF, image, and document queries using **LangChain** with **Tesseract OCR**. Enabled multiple LLM inference backends (**gpt-4o-mini, Mixtral and Llama3**) and leveraged sentence-transformers/all-MiniLM-L6-v2 from Hugging Face for enhanced text embeddings. Developed **CI CD pipelines**, docker files and built image containerized solution which was deployed on VM on AWS.
- Developed Jira Agent using agentic workflows which help product owners create Jira stories it reduced 2 days of human effort per sprint.
- Trained and deployed **regression model** on historical unplanned leaves. Predicted values were taken as input signal for supply-demand prediction of workforce. Model helped the Production team (Manufacturing) to optimize their roosters and reduce the idle machine time in range of 10-25 %
- Developed scalable RESTful APIs to support AI analytics and inference workflows. Ensured robust integration between microservices and Retool-based UIs. **Reduced dataframe write operations by ~76%** by optimizing code using external libraries and tools (Polaris, Numba and Mojo and copy method by postgres)
- Modeled and optimized relational schemas in PostgreSQL. Implemented security constraints to **safeguard against SQL injection, HTML injection**, and other data vulnerabilities. Led database migrations and post-audit security enhancements.
- DevOps & Release Management: Streamlined CI/CD workflows by resolving container build issues. **Managed production Kubernetes pods using ArgoCD and Kubectl**, containerized services via Docker, and configured multi-environment deployment pipelines.
- Facilitated product workshops with HR and leadership teams to gather feedback, refine AI features, and iteratively improve **LLM prompting** logic to align with real-world decision-making workflows.
- Frontend Development: Developed multiple custom JavaScript logic within Retool to power interactive dashboards, administrative controls, and end-user interfaces for multiple modules across the platform.

Personal Projects

Autonomous Agents for Bank Ecosystem using Autogen Core (Jan – present)

- Implemented a multi-agent conversational banking chatbot in Python using **AutoGen Core** and **FastAPI**.
- Developed each agent using only **low-level classes** from Autogen Core and developed feature like communication, tool call, handoff, delegate for each agent by implementing **OOPs** concept.
- Made sure the agents are balanced between deterministic and generative flow.

- Adopted a multi-agent design pattern (similar to a **pub-sub/handoff pattern**) in which a Domain Classifier Agent parses user intent and hands off tasks to specialized domain agents (Retail, Corporate, etc.).
- Structured real-time interactions via **WebSocket integration**, returning JSON "agent_response" messages; employed **pydantic** for input validation, CSV-based data for user accounts, balances, and payment ledgers.
- Ensured extensibility by separating each agent's domain logic and conversation flow, enabling sub-agent transitions (e.g., "make payment," "check balance") with minimal code overlap.
- Focused on **agentic design principles**, enabling each agent to handle or delegate tasks seamlessly, thereby supporting future domain expansions (like Investment support, or wealth management).

Database AI Agent – April 2024

- Built a GenAI-powered agent leveraging vanna (open-source) and Gemini 1.5 for SQL database interaction which allows user to query any Database using connectors and create charts and graphs for each query.
- Performed Reinforcement Learning from Human Feedback (**RLHF**) on **GCP Model Garden via Vertex AI pipelines**, continuously improving the model's natural language querying capability.

RAG Application for all files with OCR (Jun 2024 – Jul 2024)

- Implemented a RAG pipeline supporting PDF, image, and document queries using LangChain with Tesseract OCR.
- Enabled multiple LLM inference backends (e.g., gpt-4o-mini, Mixtral and Llama3) and leveraged sentence-transformers/all-MiniLM-L6-v2 from Hugging Face for enhanced text embeddings.
- Developed CI CD pipelines, docker files and built image containerized solution on ec2 instance on AWS

Miscellaneous

- Built a RAG application using a **llama-3** inference endpoint on Groq.
- Integrated the nomic-embed-text embedding model from Ollama within **LangFlow** to facilitate efficient semantic search and context retrieval. **Vector Store – Croma DB**
- Contributed to a 3D avatar-based chatbot featuring text-to-speech integration.
- Fine-tuned voice models using Eleven Labs and Google **Text-to-Speech AI TTS solutions**, enhancing user interactivity and immersion.
- Designed on-premises and cloud-based solution architecture for a production-grade healthcare voice-bot, Autonomous Banking chatbot on AWS
- Scoped compute and storage requirements, infrastructure deployments, and demoed a functioning POC to client stakeholders.