Shad Sheikh

Full-Stack AI Engineer
+91-9039362991 - Portfolio - shad07sheikh@gmail.com - linkedin - github

Professional Summary

Won the best engineer award in 2024 for outstanding contributions. Solutions-driven professional with over three years of experience developing and deploying generative AI solutions for real-life challenges, mainly in the legal and health industries. Skilled in retrieval-augmented generation (RAG), model fine-tuning (LoRA/QLoRA), LLMs, and MLOps, integrating full-stack development (React, FastAPI, SQL, and NoSQL) with cloud-native deployments on AWS and Azure. Proficient in gathering requirements, communicating with cross-functional teams, streamlining processes, and solving problems.

TECHNICAL SKILLS

Programming Languages: *JavaScript (ES6+), Python, C++*
Libraries and Tools: *NumPy, Pandas, Matplotlib, scikit-learn, PyTorch, LLM, Generative AI, fine-tuning (LoRA, QLoRA), TensorFlow, Hugging Face, LangChain, LangGraph, MLflow, Spark, REST API, ReactJS, FastAPI, Postman, CI/CD (GitHub Actions, Jenkins), Kubernetes, Docker, Terraform, Kafka, Git, Jira, Confluence*
Database and Concepts: VectorDB (Pinecone, Chroma, Faiss), MySQL, PostgreSQL, MongoDB, Retrieval-Augmented Generation (RAG), NLP, Deep Learning, AI models (GPT, Claude, LLaMA), Azure (OpenAI, App Services, Functions, Storage, Data Lake), AWS (Sagemaker, EC2, S3, Lambda, VPC, IAM, DynamoDB), Databricks, Hadoop, Full-Stack Development, Agile Methodology, Computer Science, Software Development, Data Structures and Algorithms, Software Engineering, Testing, Cloud Computing, System Design, Code Reviews, Object-Oriented Programming, Software Development Life Cycle

WORK EXPERIENCE

Software Engineer-II

ConsultAdd Inc, Pune, Maharashtra, India — *Jan 2023 – Present*

US Based Leading Legal Client

- **Collaborated** with **stakeholders** to identify challenges in legal document summarization and designed an **AI-driven solution** to improve **accuracy and relevance**.
- Developed efficient document **chunking** and **parsing** pipelines using **NLP** and **LLM embeddings** for semantic representation of lengthy legal texts.
- Implemented Retrieval-Augmented Generation **(RAG) architecture** by indexing document **embeddings** into **Pinecone** vector database for fast, **contextual search**.
- Designed **preprocessing** pipelines for both **OpenAI GPT-4** APIs and utilized PyTorch to **fine-tuned** models **(LoRA/QLoRA)**, deployed on AWS SageMaker. Supported scalable, **high-volume** workloads across multiple client environments.
- Leveraged **LangChain** to architect and **orchestrate** complex Retrieval-Augmented Generation (RAG) workflows, integrating document **chunking**, embedding generation, **Pinecone** vector search, and fine-tuned GPT-4 LLMs to deliver accurate, **context-aware** legal document summarizations, **improving efficiency** and **reducing hallucination** errors by 15%.
- Established CI/CD practices integrating **testing** and **deployment** automation using containerized **microservices** (**Docker, Kubernetes**) and **CI/CD** tools (Jenkins, Git, Jira), reducing manual errors by 30%.
- Achieved a 25% increase in summarization **accuracy**, 15% reduction in **hallucination** errors, and **40%** decrease in response times, significantly **boosting** client **productivity**.
- Integrated React-based user interfaces and FastAPI **microservices** to enable **seamless** document upload, query processing, and **real-time** summary delivery, **enhancing** user experience and **workflow efficiency**.

U.S.-Based Health Care Cleint

- Designed and developed a full-stack **medallion architecture** (Bronze, Silver, Gold, Platinum) using Azure Data Lake, Databricks, and Synapse Analytics to standardize multi-stage **ETL processes** and streamline advanced **AI/ML workflows**.
- Managed **20+ PMS** systems data from on-prem and **cloud environments** into Azure Data Lake. Ensured continuous schema **evolution** and zero data loss.
- Built **ETL pipelines** with Azure Data Factory and Azure Functions using **event-based** and **schedule-based triggers**, optimizing **throughput** and reducing **latency** by 35%.

- Engineered ingestion modules using **Auto Loader** with schema evolution, rescue handling, and **parallel execution**, supporting **real-time** analytics and **AI retraining** workflows.
- Refactored legacy code in **Databricks** notebooks and **Python** to enhance **maintainability**, **pipeline** accuracy, and runtime efficiency, cutting **execution** time by 30%.
- Applied NumPy, Pandas, Matplotlib, and scikit-learn for data cleansing, **validation**, and **visualization** of **TB-scale** healthcare datasets, **accelerating** feature engineering cycles.
- Delivered a unified, scalable data lake and optimized **ML pipelines** that enhanced the client's **AI-driven clinical decision** support and operational **analytics capabilities.**
- Collaborated with **cross-functional** teams to troubleshoot minimal-documentation **legacy migrations**, enabling successful delivery and superior **client satisfaction**.

Internal Product

- Led full **agile** lifecycle development from gathering **requirements** to delivering the final **product**.
- Developed **large-scale** products in **React and FastAPI** to **analyze system** data and visualize system status on dashboards, optimizing **real-time** monitoring.
- Designed and deployed **microservices-based architectures** in FastAPI to power **real-time** dashboards and integrations.
- Programmed a **Python** app to **automate** data collection from systems and send it to an API, reducing manual effort by **90%**.

Software Engineer

Metafic co, Indore, MP, India *Jul 2022 - Dec 2022*

- Built **responsive** front-end from **scratch** using React, ensuring modern **UI/UX** standards, while collaborating with **cross-functional** teams and maintaining quality through **code reviews**, **unit testing** (Jest), and performance troubleshooting.

EDUCATION

Acropolis Institute of Technology and Research **MP, India**

Bachelor of Technology in Information and Technology *CGPA - 8.72/10*

PROJECTS

- **3D Portfolio**, Designed and developed a personal 3D portfolio website to showcase skills, projects, and **interactive 3D** elements, creating an **engaging** user experience. Utilized **advanced** web technologies, including JavaScript, **Three.js**, and **WebGL**, to implement **dynamic** 3D animations and models. link

CERTIFICATIONS

- **Meta** Certified Full-Stack Software Engineer Link
- **Amazon** Web Service (AWS) Certified Solution Architect Link
- **Google** Certified IT Support Link