Shaikh Faizan Ahmed

+91 8989422489
faizan.iitjammu2023@gmail.com
Hyderabad, India
Github
LinkedIn
Portfolio

Generative AI Engineer | Software Developer | LLM & Agentic AI Systems | Full Stack Developer | RAG Pipelines | AI Voice | Automation | AWS Cloud Infrastructure | Backend & Observability Engineering

PROFILE SUMMARY

- **Generative AI Engineer** and **Backend Systems Developer** with 2.4+ years of experience building scalable production-grade AI platforms at Amazon and Vibrium AI, specializing in LLM applications**,** RAG pipelines**,** and agentic AI systems.
- Engineered **real-time AI voice automation infrastructure** handling **1L+ daily call sessions per client**, leveraging AWS Bedrock**,** OpenSearch vector search**,** Titan embeddings**,** Aurora RDS**,** Lambda**,** and event-driven architectures**.**
- Developed advanced **multi-agent orchestration frameworks** with **agent-to-agent** and agent-to-human transfer**,** custom function tooling**,** contextual memory preservation, and production-ready observability pipelines **for** latency**,** TTFT**,** token usage**, and** analytics tracking**.**
- Delivered high-impact **backend engineering** and **observability systems** at Amazon, including **ETL-driven diagnostics systems**, **performance regression analysis platforms**, and large-scale logging optimizations improving debugging efficiency**,** system reliability**, and** troubleshooting speed**.**
- M.Tech graduate from **IIT Jammu** with strong expertise across **Python**, **AWS Cloud**, **FastAPI**, **LangChain**, **OpenAI APIs**, **distributed systems**, and **AI infrastructure engineering**, with proven experience in client-facing production environments **and** high-scale AI deployments**.**

WORK EXPERIENCE

Full Stack Developer | Vibrium AI | Oct 2025 – Apr 2026 | Hyderabad

- Designed and implemented a production-grade observability framework for AI voice agents using AWS Aurora RDS, Lambda, and SQS, enabling real-time tracking and analytics of **1L+ daily call sessions per client**.
- Built and optimized Retrieval-Augmented Generation (RAG) pipelines for voice-based GenAI systems using AWS Knowledge Bases, Titan embeddings, and OpenSearch vector search, **achieving ≤ 40 ms retrieval latency for real-time conversational responses**.
- Engineered a **UI-driven agent orchestration framework** enabling dynamic agent-to-agent and agent-to-human transfers, supporting unbounded chaining with stateful context preservation across live sessions, successfully deployed for 2 production clients.
- Developed scalable **backend services for LLM-powered voice automation systems**, integrating vector retrieval, conversation orchestration, and cloud-native infrastructure for real-time AI applications.
- Built event-driven data pipelines processing **1L+ daily call events**, transcripts, and analytics records, enabling structured monitoring, performance insights, and post-call intelligence generation.
- Implemented a **dynamic function tool framework** allowing agents to execute external actions such as API calls, workflow triggers, and phone transfers, fully configurable through a platform UI.
- Led cloud modernization by upgrading a client on-premise deployment to **AWS-based infrastructure**, implementing Aurora RDS and supporting cloud services, significantly improving scalability, reliability, and operational efficiency.

CORE COMPETENCIES

- Generative AI & LLM Application Development
- Agentic AI Systems & Multi-Agent Orchestration
- Retrieval-Augmented Generation (RAG) Architecture
- AI Voice Automation & Real-Time Systems
- Back-end Systems Design & Development
- Cloud-Native Architecture (AWS & GCP)
- Event-Driven & Distributed System Design
- Observability, Monitoring & Performance Engineering
- API Design, Integration & Microservices Architecture
- Production System Debugging & Optimization

TECHNICAL SKILLS

Programming Languages: Python, Java, JavaScript, C, C++, HTML
Backend & Frameworks: FastAPI, Flask, Node.js, NumPy, Pandas, TensorFlow, PyTorch
Databases & Storage: MySQL, DynamoDB, Redis, AWS S3, AWS Aurora RDS
Cloud & DevOps: AWS (EC2, S3, Lambda, Bedrock, Glue, Athena, CloudWatch, OpenSearch), GCP, Docker, Linux, Jenkins, Git, Gerrit
GenAI / AI Stack: LLMs, RAG, LangChain, HuggingFace, OpenAI API, Prompt Engineering, Agentic AI, Vector Databases, Embeddings (Titan, OpenAI)
Data & Observability: ETL Pipelines, Kibana, Log Analytics, Performance Monitoring, Distributed Logging Systems
CS Fundamentals: Data Structures & Algorithms, OOP, System Design, RDBMS
Professional Skills: System Design, Debugging, Performance Optimization, Code Review, Problem Solving, Leadership, Client Communication

EDUCATION

Master of Technology in Computer Science and Engineering | 2023
Indian Institute of Technology Jammu, Jammu

Bachelor of Technology in Computer Science and Engineering | 2019
Jaipur National University, Jaipur

Software Development Engineer | Amazon | Jan 2024 – Jul 2025 | Bengaluru

- Rebuilt a mission-critical internal log analysis platform, **restoring 100% core functionality**, improving system stability, and successfully deploying it to production environments.
- Engineered and deployed an ETL-based Low Memory Killer (LMK) reporting system, utilizing custom parsing logic and a dynamic JavaScript UI with AWS Glue, Athena, and S3, **improving log traceability by 3x**.
- Led a complete UI modernization of LogParrot, significantly reducing internal troubleshooting time by approximately 25% and improving adoption across engineering teams.
- Enhanced PerfTracker observability platform, **fixing 10+ legacy issues**, introducing OOM scoring, memory KPIs, and configurable measurement intervals for CPU, memory, perfstats, eMMC, and job scheduler metrics.
- **Reduced logging noise by approximately 60%** and improved **debugging efficiency by 30%** through performance and observability enhancements in PerfTracker logging architecture.
- Delivered **critical Fire OS stability improvements**, including configurable ANR timeouts, resolution of 2 major crash bugs in FOS 6, and implementation of granular tombstone diagnostics for crash analysis.

PROJECT

Agentic Voice AI System | Python, LiveKit, OpenAI, Cartesia TTS, Google Calendar API | Link: ByteonAI

- Built and deployed a **production-grade multi-agent voice AI system** on **LiveKit Cloud** with intelligent routing, real-time **STT/TTS**, and function-calling capabilities for **appointment booking and contact management**.
- Integrated **Silero VAD, Cartesia Sonic-3 TTS, and Google Calendar API**, enabling **sub-100ms response latency**, automated scheduling, conflict detection, and email confirmations.
- Designed a **web-integrated conversational voice bot** capable of handling **concurrent sessions with persistent context management**.
- Deployed as a scalable system powering real-time voice interactions.

MyTripMate: Multi-Agent AI Travel Planner | Python, Google ADK, Vertex AI, GCP | Github Repo: MyTripMate

- Designed a **multi-agent travel planning system** using **Google ADK**, featuring a root orchestration agent and specialized sub-agents for brainstorming, logistics, attractions, accommodation, booking, and monitoring.
- Built an **adaptive itinerary engine** capable of dynamically adjusting plans based on **weather changes, delays, and budget constraints**, providing optimized alternatives in real time.
- Implemented a **collaborative agent framework** enabling structured reasoning and task decomposition across multiple AI agents.
- Developed as part of a **Google GenAI Hackathon submission**.

Insurance Document RAG Chatbot | Python, LangChain, Gradio, HuggingFace, Groq | | Github Repo: LLM_RAG_Agent

- Built a **Retrieval-Augmented Generation (RAG) chatbot** for multi-format insurance documents with **90%+ retrieval accuracy** using **FAISS vector search and optimized chunking strategies**.
- Integrated **LangChain, Groq LLM API, and HuggingFace embeddings** to deliver fast conversational responses under strict token limits (100–300 tokens) with memory-aware context handling.
- Designed a **real-time document QA system** with sub-second response latency and semantic search capabilities.

Real-World Motion Blur Video Super Resolution | Python, PyTorch, NumPy

- Proposed and implemented a deep learning model for video super-resolution, enhancing low-resolution frames into high-resolution outputs.
- Addressed real-world degradations including motion blur, noise, downsampling artifacts, and compression distortion.
- Improved visual fidelity using advanced reconstruction techniques for temporal frame enhancement.

GAN-based Single Image Super Resolution | Python, PyTorch, CelebA Dataset

- Designed and trained a **SRGAN-based architecture** to enhance image resolution up to **4× upscaling**.
- Applied **perceptual loss optimization and adversarial training** to improve visual realism and fine-grained detail restoration.
- Tuned hyperparameters to improve model stability and output quality across diverse image samples.

CERTIFICATIONS

- **Google Cloud Skills Boost** – Deploy Multi-Agent Systems with ADK and Agent Engine
- **IBM (Coursera)** – Develop Generative AI Applications: Get Started
- **IBM (Coursera)** – Build RAG Applications: Get Started
- **DeepLearning.AI (Coursera)** – Machine Learning Specialization
- **Google (Coursera)** – Google Data Analytics Professional Certificate
- **HPE** – Big Data Programming and Development (Certificate No. HPE/CoC/ET/1807-03153)