Shahzaib Khan

hello@shahzaib.ai | +92-370-4763017 | linkedin.com/in/shahzaib17 | github.com/Shahzaib0217 | shahzaib.ai

Professional Summary

- Full-Stack AI Engineer with 3+ years of experience building production AI systems: RAG pipelines, multi-agent workflows, and voice AI agents, across hospitality, meetings & productivity, legal, governance, and SaaS domains.
- Founding engineer at multiple AI startups, including a Y-Combinator funded company. Shipped MVPs, scaled products to 300+ users, and led engineering teams.
- Strong background in AI agent architecture, MCP integrations, and end-to-end product delivery. Co-authored a research paper on LLM-assisted testing for smart contracts.

Education

National University of Computer and Emerging Sciences (Fast-NUCES) Sept 2020 – June 2024

- BS in Computer Science

Skills Summary

Programming Languages: Python, JavaScript, TypeScript, C++, SQL

Web Development: FastAPI, Node.js, React.js, Next.js, gRPC, Docker, AWS (S3, EC2, Bedrock), GCP (Cloud Run, E2)

Databases/Vector Databases: PostgreSQL, MySQL, MongoDB, Supabase, Redis, Pinecone, pgvector, Weaviate

Gen AI: AI-Agents, RAG, MCP, LangChain, LlamaIndex, Pydantic AI, Fine-tuning, Voice AI (LiveKit, VAPI, Retell)

Tools: Git, GitHub Actions, Nginx, n8n, Twilio, Telnyx

Work Experience

AI Product Engineer & Team lead, DreamDesk AI – Remote, UK Jan 2026 – Present

- Led engineering team and shipped POC within 4 weeks. Designed end-to-end product architecture and UX for a hospitality AI platform. Collaborated directly with the product owner to scope features and drive R&D.
- Built multi-channel AI agents handling phone calls, emails, web chat, WhatsApp, and Booking.com messages, automating end-to-end guest communication for hotels.
- Integrated Mews & Guestline PMS platforms for real-time reservation sync. Built RAG pipeline, ingesting hotel knowledge bases.
- Tech Stack: FastAPI, Next.js, GCP (Cloud Run & Compute Engine), Pydantic AI, Pinecone, Postgres, GitHub Actions

FullStack AI Engineer, Simplora AI – Remote, USA (Part-time) July 2025 – Present

- Founding AI Engineer. Shipped MVP in 2 months, scaled platform from launch to 300+ active users.
- Built agentic workflows using LangChain to automate the complete meeting lifecycle. Integrated MCP servers to connect agents with CRMs and platforms like ClickUp, Slack, Google Drive and Notion.
- Integrated third-party APIs like Apollo. Optimized DB queries, reducing response latency from 60s to under 2s.
- Tech Stack: Node.js, Next.js, LangChain, MCP, Postgres, Digital Ocean, GitHub Actions

FullStack AI Engineer, AI Gov Sandbox – Remote, USA April 2025 – Nov 2025

- Founding Fullstack AI Engineer. Built a multi-agentic system for automated evaluation of AI systems against EU AI Act, US state-level AI regulations, and global compliance frameworks.
- Integrated MIT AI Risk Repository, IBM AI Risk Atlas Nexus, and Trusted AI's ART for security and risk evaluation. Dockerized each agent as an independent microservice.
- Tech Stack: Python, FastAPI, LangGraph, Docker, MongoDB, AWS EC2 & S3, GitHub Actions.

FullStack AI Engineer, Foomotion LLC – Lahore, Punjab Dec 2024 – April 2025

- Built LLM-powered RAG and multi-agentic systems using AWS Bedrock, LangGraph, LangChain, and LlamaIndex. Developed LiveKit voice agents and n8n automation workflows.
- Worked directly with **YC-funded startup** (Pearson Labs), scoped requirements, led daily client syncs, and independently owned AI development end-to-end.
- Tech Stack: FastAPI, Next.js, LangGraph, LangChain, LlamaIndex, LiveKit, n8n, Postgres. AWS (EC2, S3, Bedrock)

AI Engineer, Cyber Evangelists – Lahore, Punjab June 2024 – Dec 2024

- Developed a real-time RAG voice calling agent using Node.js, integrating TTS, STT, and voice cloning for automated customer conversations.
- Built the MVP for an Attack Surface Management (ASM) platform using FastAPI and gRPC microservices. Dockerized each service with PostgreSQL and Redis for storage and caching.

Full Stack AI Engineer, Freelance – Remote (Part-time / Contract based) June 2023 – Feb 2025

- Built MVPs for AI-powered products and an ERP system, spanning RAG, Voice AI agents, and multi-agentic systems.
- Tech Stack: FastAPI, Next.js, React, LangGraph, LangChain, LlamaIndex, VAPI, LiveKit

Software Engineer, Intern, DevDen – Faisalabad, Punjab June 2023 – Aug 2023

- Built an admin dashboard for a POS system with React.js, featuring user management and sales analytics.

Projects

Pearson Labs - Chat with Legal Documents (Y-Combinator Funded) App Link

- Built an advanced RAG pipeline using LangChain, LlamaIndex, and PyMuPDF for document processing. Implemented multi-tenancy in Pinecone for scalable vector storage and query expansion techniques to improve retrieval accuracy. Added agentic behavior via tool calling.
- Integrated the pipeline with FastAPI and AWS S3, built APIs and connected them with a Next.js frontend.
- Tech Stack: Pinecone, FastAPI, AWS S3, Next.js, LangChain, LlamaIndex, PyMuPDF, Pydantic

Voice Calling Agent using LLMs App Link

- Built a real-time voice calling agent using WebSocket and Node.js with STT, TTS, voice cloning, and intent/emotion detection. Reduced response latency and cost by 20%. Deployed with Nginx on VPS.
- Tools: Node.js, Deepgram, Play.ht, LiveKit, Twilio

Attack Surface Management Software (MVP) App Link

- Built microservices using FastAPI and gRPC to run cybersecurity tools independently. Dockerized each service with PostgreSQL and Redis for storage and caching. Developed a RAG agent using Llama-3.1 to query real-time data from the tools.
- Tools: Python, FastAPI, gRPC, Docker, LangChain, Postgres

Research Project

LLM-assisted High-Quality Property Generation for Solidity Smart Contracts (Under review at STVR)

- As the **1st author** of the paper, I worked remotely with Dr. Hassan (DFKI Bremen, Germany) and Dr. Jalil (UAEU).
- Proposed an LLM-assisted approach to generate high-quality test invariants for property-based testing of Solidity smart contracts. Built a pipeline that generates invariants, then uses an intelligent agent to automatically resolve compilation errors in the output. Evaluated multiple prompt engineering techniques and tuned hyperparameters to reduce hallucinations and improve invariant reliability.
- Benchmarked LLM-generated properties against human-written ones using mutation testing. The LLM-assisted approach achieved a 25.99% mutation score vs 31.75% manual, demonstrating that LLMs can significantly reduce human effort in test invariant generation while approaching manual quality.
- Tools: Python, LangChain, LLMs, Azure-ML, Fuzzing, Property-based testing, Mutation testing, LaTeX

Certifications & Hackathons

Data Analyst with Python (36 hours course) | Certificate Data Camp
LangChain Master class (Python) | Certificate Udemy

International Hackathons

IBM WatsonX Assistant Hackathon - lablab.ai November 2024

- Our HR Bot boosts employee engagement and streamlines HR by analysing satisfaction through daily feedback. Link!