

FAIZ KHAN

AI ENGINEER | LLM & AGENTIC SYSTEMS | FULL-STACK

India | +91 94924 57193 | faiz.itguy@gmail.com | linkedin.com/in/faizitguy

PROFESSIONAL SUMMARY

Full-Stack AI Engineer with 6+ years building LLM-powered, agentic systems using LangChain, LangGraph, RAG, and MCP over Claude (AWS Bedrock) and OpenAI. Proven impact in agent guardrails, LLM observability, and cost optimization. AWS Certified AI Practitioner, strong in Python (FastAPI), Node.js, and React, shipping well-tested AI systems on AWS and Azure for enterprise clients in fintech, publishing, and telecom.

CORE SKILLS

Generative AI & LLM: LangChain, LangGraph, Agentic AI, Multi-Agent Systems, RAG, Model Context Protocol (MCP), Prompt Engineering, LLM Observability (Langfuse), Agent Guardrails, AI Safety & Evaluation, LLM Cost Optimization, AWS Bedrock (Claude), OpenAI API

Vector DBs & Data: Pinecone, Qdrant, PostgreSQL, MongoDB, MySQL, Redis

Languages: Python, TypeScript, JavaScript, SQL, Java

Backend: FastAPI, Node.js, Express.js, Flask, REST APIs, Microservices, Server-Sent Events (SSE), WebSockets, Pydantic, async/await

Frontend: React, Next.js, Redux, Zustand, TanStack Query, TypeScript, Vite, Tailwind CSS, Material UI

Cloud & DevOps: AWS (Bedrock, SageMaker, Lambda, ECS, Step Functions, S3, EC2, IAM, VPC), Azure (App Services, Functions, AI Services, DevOps), Docker, Kubernetes, Terraform, GitHub Actions, Jenkins, CI/CD

Testing & Tooling: pytest, Bruno, Langfuse, Git, Postman, JIRA

Collaboration & Leadership: Technical Mentorship, Cross-functional Collaboration, Code Review, Agile / Scrum, Stakeholder Communication

PROFESSIONAL EXPERIENCE

Edstem Technologies Pvt. Ltd — Senior AI Engineer

May 2021 – Present

Remote, India

- Architected AI-enabled applications (Python, FastAPI) for enterprise clients, improving load time 40% and reducing latency 30%.
- Cut deployment time 60% with CI/CD pipelines (Jenkins, GitHub Actions); lowered production issues 35% through code reviews and continuous testing.
- Mentored 50+ engineers in Python, Node.js, and React, accelerating delivery 50% and standardizing team workflows.

KEY CLIENT PROJECTS

Vericast Marketing Strategist Agent ("vAI")

Client: Vericast (USA) | Backend AI Engineer

- Engineered a conversational AI system on a LangGraph agentic loop over Claude (AWS Bedrock), orchestrating 30+ LangChain tools and subagents with PostgreSQL persistence and real-time SSE streaming to a React 18 frontend.
- Designed a real-time **agent-loop guardrails framework** (HookedToolNode + HookRunner) that **eliminated a recurring ~\$946/two-week LLM token-waste problem** and 1M+ token context blowouts; shipped ~700 LOC with 86 tests and zero regressions.
- Implemented **self-hosted Langfuse LLM observability** via a custom REST callback handler (~500 LOC) with per-request cost tracking, latency monitoring, and PII redaction, keeping 100% of financial data on-prem.
- Refactored the web analyzer into 1 of 4 internal LangGraph subagents, removing a cross-service network hop and adding a "Quick Analyze" fast-path that fixed timeouts on heavy SPAs.
- Hardened a sandboxed Python code-executor, closing 2 vulnerability classes (sandbox-escape, path-traversal) with a Bruno security test suite.

- Authored the backend test suite (1,500+ tests) across agent orchestration, tools, database, and auth — the platform's regression safety net.

Tech: Python, FastAPI, LangGraph, LangChain, Claude (Bedrock), PostgreSQL, Redis, React 18, Docker, Kubernetes, Langfuse, pytest

Imperium — Trust Management Platform

Client: PT Trustees Ltd (UK) | Full-Stack AI Engineer

- Integrated LLM-powered assistants (OpenAI, LangChain, MCP) into internal tools and built a RAG knowledge bot (Pinecone) that boosted documentation retrieval 45%.
- Shipped full-stack solutions (Node.js, React) for a trust-management platform handling £6B+ in trust funds annually.
- Managed AWS infrastructure at 99.99% uptime; automated CI/CD with Terraform and GitHub Actions, shortening release cycles 40%.

Tech: Node.js, React, OpenAI, LangChain, MCP, Pinecone, AWS, Terraform, GitHub Actions

Rights & Permissions Collector

Client: Elsevier (UK) | Full-Stack Engineer

- Built a distributed permissions system (Node.js/Express microservices, React/TypeScript) on Azure with Application Insights, cutting API response time 40% through SQL and MongoDB optimization.

Tech: Node.js, Express.js, React, TypeScript, Azure, Docker, SQL, MongoDB

Telstra.com.au — Checkout & Payments

Client: Telstra Group Limited (Australia) | Senior Backend Developer

- Enhanced backend checkout and payment flows (Node.js microservices) for a telecom platform (~A\$23B FY24 revenue), securing billing and reducing transaction failures across markets.
- Diagnosed and resolved checkout defects across 3 environments (CI, dev, staging), stabilizing the release pipeline.

Tech: Node.js, Express.js, SQL, Docker, Bamboo CI/CD

More Retail Limited — Management Trainee

Jun 2019 – Aug 2020

Proddutur, India

- Oversaw inventory, purchase orders, and reporting in SAP Retail and SIM systems, supporting demand forecasting and stock optimization.

CERTIFICATIONS

- **AWS Certified AI Practitioner** — Amazon Web Services ([verify](#))
- **Agentic AI** — DeepLearning.AI (Andrew Ng); building agents that plan, use tools, and self-improve via reflection and multi-agent workflows ([verify](#))

EDUCATION

Full Stack Web Development — Masai School

Aug 2020 – Apr 2021 | Online, India | Data Structures, Agile Practices

Master of Business Administration (Marketing) — KSRM College of Management Studies

May 2017 – Apr 2019 | Kadapa, India

B.Sc. (Mathematics, Statistics & Computer Science) — Sri Venkateswara Degree College

May 2014 – Apr 2017 | Kadapa, India

AWARDS & ACHIEVEMENTS

- Received the Endeavor Star Award for exceptional commitment and value creation.
- Delivered technical lectures at internal company events, driving cross-team knowledge sharing.