

ARSLAN IQBAL

AI Software Engineer | Full Stack | LLM & Agent Systems
arslaniqbalmgt@gmail.com • +923164547382 • linkedin.com/in/arslan-iqbal-ai

PROFESSIONAL SUMMARY

AI Software Engineer with 5+ years of experience building and shipping production-ready full-stack web applications integrated with AI and LLM capabilities. Proven track record designing AI agent workflows, RAG systems, and scalable APIs on Azure and AWS. Deep hands-on expertise with Anthropic Claude, OpenAI, LangChain, Supabase, and MCP — with a strong focus on system reliability, security, and iterative product delivery. Codes daily with Claude Code, Cursor, and GitHub Copilot. Experienced leading teams, translating product requirements into working technical solutions, and owning features end-to-end from concept to launch.

TECHNICAL SKILLS

AI / LLMs	Anthropic Claude, OpenAI, Gemini, LangChain, MCP, ADK, AI Agents, Embeddings, RAG
Full Stack	React.js, Next.js, TypeScript, TailwindCSS, Node.js, FastAPI, Django, REST APIs
Backend / DB	Supabase, PostgreSQL, MongoDB, MySQL, Redis
Cloud / DevOps	Azure, AWS, Docker, GitHub Actions, RunPod, CI/CD
AI Dev Tools	Claude Code, Cursor, GitHub Copilot
Security	Authentication, Permissions, Data Protection, API Security

PROFESSIONAL EXPERIENCE

AI Software Engineer | Zapta Technologies 2025 – 2026

- Built and shipped LLM-powered features into full-stack production applications using Anthropic Claude, OpenAI, and Gemini APIs
- Developed multi-agent AI workflows using ADK for intelligent task automation with tool-use and memory
- Designed and implemented AI guardrails to handle hallucinations, edge cases, and failure scenarios in production
- Deployed AI services on Azure and AWS with automated CI/CD pipelines via GitHub Actions — ensuring reliability and scalability
- Led client meetings, gathered product requirements, designed AI solution architectures, and owned end-to-end delivery
- Built product interfaces with React.js and Next.js — shipping features rapidly and iterating based on feedback

AI Engineer / Full Stack Lead | KCube AI 2024 – 2025

- Designed and deployed AI agent systems with tool-use and memory using Azure OpenAI, LangChain, MCP, and Grok
- Built Supabase-backed and PostgreSQL data architectures supporting scalable API design and efficient data retrieval
- Implemented RAG pipelines with Embeddings and vector search for large-scale document knowledge retrieval
- Managed MLOps pipelines on Azure and AWS; deployed GPU inference workloads on RunPod for scalable AI model serving
- Led cross-functional engineering team — system design, sprint planning, code reviews, and direct client communication
- Scaled product features from MVP through production — iterating based on real-world usage and performance data

Software Engineer | Ripeseed 2022 – 2024

- Integrated OpenAI APIs into production applications for AI-powered features — built with reliability and performance in mind
- Built complex frontend components in React.js and Next.js (TypeScript) with strong attention to UX quality
- Grew from team member to team lead — taking ownership of system architecture, database design, and delivery accountability

Software Engineer | JAX Agency 2019 – 2022

- Developed data extraction pipelines using Python and Scrapy for web scraping at scale
- Built REST APIs in Node.js to integrate and serve scraped data into web platforms

PROJECTS BUILT

■ X (Twitter) Account Automation

Problem:

Automate X accounts to grow engagement — AI generates contextually relevant comments, with 30 different profiles posting human-like responses to boost account visibility.

Solution:

Used Playwright to simulate human browsing behaviour, saving and injecting cookies into Chromium sessions for each profile. Assigned static IPs per profile so each is treated as a separate machine. Groq powers the comment generation with fast, creative responses. Cron jobs schedule posting activity at natural intervals.

Tech Stack: Node.js, TypeScript, Groq, Playwright, Azure VM, Cron Jobs

■ Core Memory — AI-Powered Personal Memory Platform

Problem:

A SaaS app where users store personal memories and an AI voice agent narrates their life story in the third person, generating an autobiography summary after each conversation. Users can share memories with friends and family, who can then converse with the shared person's memories in their cloned voice.

Solution:

Integrated ElevenLabs for voice cloning and the conversational voice agent. Built a React frontend embedding the agent via its agent ID. FastAPI handles all backend logic including memory storage, sharing flows, and session management. Deployed on Azure Web Apps.

Tech Stack: React.js, FastAPI, ElevenLabs API, PostgreSQL, Azure Web Apps

■ AI Voice Takeaway Ordering System

Problem:

A takeaway restaurant needed a voice agent to handle phone orders and log them to Google Sheets. Initial solution via VAPI cost ~\$1,000/month — client needed it under \$200.

Solution:

First iteration: replaced VAPI with Ultravox integrated via a FastAPI/Python script through Twilio for inbound calls — reducing cost to ~\$500/month. To cut further, built a custom pipeline: Deepgram transcribes speech in real-time, OpenAI generates the response, and Azure AI Speech synthesises the reply — bringing cost to ~\$200/month. Latency was the critical challenge; after extensive research adopted Pipecat, which orchestrates the full real-time audio pipeline and resolved the latency issues.

Tech Stack: FastAPI, Twilio, Pipecat, Deepgram, OpenAI, Azure AI Speech

■ MindDump — AI Advanced Note-Taking App

Problem:

Users need to quickly capture raw thoughts and brain dumps without friction, then retrieve them in structured, connected form through natural language queries.

Solution:

Built semantic search over user notes using embeddings and vector databases, with an LLM layer to organise and surface related thoughts intelligently.

Tech Stack: LLMs (Claude/OpenAI), Embeddings, Semantic Search, Next.js, FastAPI, Supabase

■ AI Space — Advanced Document Intelligence Platform

Problem:

Knowledge workers and researchers needed to have deep, context-aware AI conversations with large document collections at SaaS scale.

Solution:

Built an advanced RAG retrieval pipeline with vector databases for high accuracy over large corpora. Designed for teams requiring reliable document Q&A at scale.

Tech Stack: LangChain, RAG, Vector Databases, Embeddings, FastAPI, React.js, Azure OpenAI

EDUCATION & CERTIFICATION

Bachelor of Computer Science | University of Agriculture, Faisalabad 2017 – 2021

Machine Learning with Python — Coursera | Verified Certificate

CONTINUOUS LEARNING & SELF-RESEARCH

- Runs and experiments with open-source LLMs locally on a personal RTX 4060 Ti GPU — model loading, inference optimization, and performance benchmarking
- Researches and deploys models on RunPod GPU cloud — from setup through to serving live inference endpoints

- Actively evaluates frontier models (DeepSeek, Qwen) via OpenRouter — staying current with global model benchmarks and cost-performance tradeoffs
- Tests emerging AI tools as they release — assessing real-world applicability and product viability ahead of market adoption