Zakwan Mansuri

AI Automation Consultant | ML Engineer | Agentic Systems | Remote-First

Ahmedabad, India • +91 97121 21230 • mansurizak@gmail.com

Portfolio: zakwanai.github.io ·LinkedIn: linkedin.com/in/zakwanai

PROFESSIONAL SUMMARY

AI Automation Consultant with 6+ years of experience designing and deploying intelligent systems for clients across India, Australia, and international markets. Specialised in multi-agent AI workflows, LLM-powered automation, RAG architecture, and end-to-end ML pipelines — from business problem definition through production deployment. Proven track record of reducing operational costs, automating complex workflows, and delivering measurable ROI for enterprise and SMB clients globally. Available remotely for consulting engagements, contracts, and long-term partnerships worldwide.

CORE SKILLS

Agentic AI & LLMs

LangChain, LangGraph, CrewAI, SmolAgents, Multi-Agent Systems, RAG Architecture, Prompt Engineering, Fine-tuning, LiteLLM, LlamaIndex, OpenAI, Gemini, Ollama

ML & Data Science

Scikit-learn, TensorFlow, PyTorch, Keras, XGBoost, Random Forest, NLP, CNNs, RNNs, LSTM, Anomaly Detection, Forecasting, Clustering

Languages & Data

Python, SQL, R, Pandas, NumPy, Matplotlib, EDA, Feature Engineering, Data Cleaning, Statistics, Linear Algebra

Vector & Search

ChromaDB, Pinecone, FAISS, pgvector, Semantic Search, Vector Indexing & Optimisation

Backend & Infrastructure

FastAPI, Flask, MySQL, PostgreSQL, REST APIs, Docker, AWS, GitHub Actions (CI/CD), Multi-tenant Schema Design

Automation & Tooling

n8n, Zapier, Streamlit, Tableau, Power BI, OCR Pipelines, SSE Streaming, Workflow Orchestration

PROFESSIONAL EXPERIENCE

AI Automation Consultant — Freelance | **Independent — Global Remote** Apr 2020 – Present | India & International

Selected Engagements:

Regulatory AI & Compliance Automation — LLM Fine-tuning • Agentic Workflows • Multi-tenant PostgreSQL • vLLM

- Fine-tuned a large language model (30B parameters) on domain-specific regulatory and financial data to power an AI-driven compliance checking system across complex multi-tenant enterprise datasets.
- Engineered agentic reasoning loops with tool-calling for real-time document retrieval, automated compliance flagging, and structured audit reporting — reducing manual review time by 70%.
- Deployed model on private GPU infrastructure with optimised inference (vLLM, Llama.cpp, Ollama) and unified model orchestration via LiteLLM for seamless provider switching.

Conversational Analytics Agent — FastAPI • SSE Streaming • PostgreSQL JSONB • Agentic AI

- Built a production-grade conversational AI assistant embedded in a financial ERP dashboard — featuring multi-turn memory, real-time SSE streaming, and per-user conversation history in PostgreSQL JSONB.
- Enabled non-technical users (C-suite, finance teams) to query complex financial datasets in plain English — handling concurrent requests and reducing analyst turnaround time by 85%.

Automated Invoice & Document Processing Pipeline — OCR • STT/TTS • MySQL • Python

- Designed and deployed a multimodal automation pipeline integrating invoice OCR, bank statement parsing, and an IVR calling system (STT/TTS) — mapping extracted data directly into enterprise databases.

- Eliminated manual data entry for an enterprise accounts team — saving 25+ hours/week and reducing processing errors to near zero.

NL2SQL Executive Intelligence Engine — Multi-Model LLM Pipeline • ChromaDB • MySQL • FastAPI

- Deployed a natural language to SQL system enabling leadership to query a 200+ table production database via plain English — built a cost-optimised multi-model pipeline routing queries by complexity (Gemini 1.5 Flash, Phi-3 Mini, Gemma 3).
- Built semantic search layer using ChromaDB and MiniLM embeddings for schema-aware query orchestration — reduced internal data-request turnaround by 90%.

Demand Forecasting & Resource Optimisation — XGBoost • Random Forest • Geospatial Feature Engineering

- Built predictive models forecasting peak-hour demand and regional supply shortages for a ride-sharing platform — improved resource allocation efficiency by 30% across operations.
- Engineered geospatial and temporal features at scale; deployed models into production with automated retraining pipelines.

Multi-Agent Financial Research System — CrewAI • LangChain • FastAPI • Streamlit

- Architected a multi-agent system autonomously generating equity research reports — live data from Yahoo Finance and Tavily, sentiment analysis across ASX and global tickers, flexible LLM orchestration supporting 100+ providers.
- Designed 4 specialised agents (research, analysis, sentiment, synthesis) with zero-config model swapping via environment variables — production-deployed with Streamlit dashboard.

Intelligent Recommendation & Matching Systems — NLP • Unsupervised Clustering • Semantic Similarity

- Delivered multiple recommendation engines for clients across talent platforms and e-commerce — NLP-based semantic similarity, unsupervised clustering, and content-based filtering; average 20–30% improvement in match accuracy.

MS/PhD Research Consulting — Statistical Modelling • ML Implementation • Python

- Delivered thesis-level data analysis, statistical modelling, and ML implementation for graduate researchers across India and internationally — long-term client relationships built on consistent quality and reliability.

OPEN SOURCE & NOTABLE PROJECTS

AI-FAQs — Open Source FAQ Generation Tool — LLMs • RAG • Python

- Developed and open-sourced an LLM-powered FAQ generation tool that automatically creates structured FAQs from business data — integrated as a RAG knowledge base powering in-product chatbots; solo developer and maintainer.

MCQ Generation System — LangChain LCEL • OpenAI • Gemini • Ollama

- Built a production-ready MCQ generator supporting multiple LLM providers (GPT-4, Gemini, local Ollama) with dynamic model config via UI — processes PDF and text inputs, exports structured MCQs as CSV with automatic grammar and complexity evaluation.

EDUCATION

Bachelor of Engineering — Information Technology

Government Engineering College, Gandhinagar, India | 2013–2017 | CGPA: 7.08/10

Continuously upskilling: LangGraph, Advanced Agentic Systems, GPU Model Serving (vLLM, Ollama), LlamaIndex, Computer Vision, Production LLM Infrastructure

ADDITIONAL

Availability: Immediately available for remote consulting, contract engagements, and global opportunities

Engagement Models: Hourly consulting, fixed-price projects, monthly retainers, or long-term embedded consulting — flexible to client needs

Community: Google Meetups, Data Science Bootcamps, AI community events (Ahmedabad); Microsoft Imagine Digital India Summit, Gandhinagar

Working Style: Self-directed and outcome-focused — experienced in async remote collaboration, clear client communication, and delivering complex AI systems independently from scoping through production deployment