Muhammad Jehanzaib

Senior ML / Full-Stack AI Engineer & AI Architect

Production AI projects across FinTech & TaxTech, PropTech, E-Commerce, HealthTech, Logistics, Sales, Telecom, EdTech, HR Operations, and AI Infrastructure — each with the problem, approach, outcome, and stack. Grounded in agentic LLMs, MCP, RAG, computer vision, and MLOps.

mjehanzaib99@gmail.com · linkedin.com/in/muhammad-jehanzaib-4036b5195 · github.com/mjehanzaib999 · Resume: mjehanzaib-ai-resume.up.railway.app

LIVE deployed & clickable · DEMO video walkthrough · FLAGSHIP lead production work

Contents

32 projects in 11 domains. Click a domain to jump to it.

01 FinTech / TaxTech — 4 projects

02 PropTech / Real Estate — 4 projects

03 E-Commerce & Retail Intelligence — 4 projects

04 Logistics & Operations — 1 project

05 HealthTech & Elderly Care — 12 projects

06 AI Sales & Outreach — 2 projects

07 Conversational AI / Telecom — 1 project

08 EdTech / LMS — 1 project

09 Content / SEO — 1 project

10 AI Infrastructure / Open Source — 1 project

11 HR Operations / Agent Tooling (MCP) — 1 project

01 · FinTech / TaxTech

4 PROJECTS

Contract Analysis Engine — IRC §41 R&D Tax Credit FLAGSHIP DEMO

Agentic platform that determines whether research performed under a client's contracts qualifies for IRC §41 R&D credits — automating the §41(d)(4)(H) "funded research" analysis that normally takes hours of manual legal review per contract. A **multi-agentic LangGraph pipeline** (document triage, rights-gate, contingency/financial-risk, jurisdiction & state-conformity, web search, adversarial self-review, IRS counter-argument simulation) yields a qualified / not-qualified determination with confidence scoring, evidence citations, and per-prong contract references. Bulk upload with native + vision-OCR text extraction, deterministic file bundling, full batch orchestration with cancellation, and an agentic chat that feeds a correction-driven re-run — grounded by a case-law & state-conformity knowledge base. Built at Ventive for a U.S. tax-advisory firm.

STACK Claude Opus (via OpenRouter) · LangGraph · FastAPI · Celery + Redis · PostgreSQL (SQLAlchemy 2.0 async) · React 19 + TypeScript + Vite (TanStack, Zustand) · Gemini Vision OCR

LINKS Watch demos: 1 | 2 | 3 | 4 | Descript

Tax Form Extraction & Reconciliation Engine FLAGSHIP DEMO

Form-aware extraction pipeline for payroll/wage documents (W-2, W-3, census) that routes each page chunk to a dedicated per-form handler, then verifies with **dual-model consensus** (Gemini 2.5 Flash + Grok 4.1 Fast) and a Claude verify pass, with an AWS Textract path for scanned docs. Handles the messy realities of real packets — multi-employer batches, multi-page/2-up/4-up forms, and per-state W-2 splits. An automated **W-2 ↔ W-3 reconciliation engine** keyed per (EIN, tax year) cross-checks employee count, federal Box 1, and state Box 16 wages with cents-level tolerance, emitting PASS / FAIL / PARTIAL / UNVERIFIABLE verdicts plus a roster and human-readable diagnostics; structures extracted entities into a Neo4j knowledge graph. Built at Ventive.

STACK Python · AWS Textract · Gemini 2.5 Flash + Grok 4.1 Fast (dual-model via OpenRouter) · Claude Opus & Sonnet · Pydantic · FastAPI · Neo4j · Next.js

LINKS Watch demo ↗

Tax Assistant — AI Tax Strategy Optimisation DEMO

A GPT-4.1 + LangGraph agentic system that takes a tax scenario, asks structured clarifying questions, then generates optimised strategies with simulated returns for each and a comparative trade-off analysis — integrated with tax-filing APIs for execution.

STACK OpenAI GPT-4.1 · LangGraph · Python · Streamlit · Tax Filing APIs

LINKS Watch demo ↗

AI-Enhanced Financial Analysis Tool

Bank-statement analysis pipeline with OCR + LLM extraction, ML transaction categorisation, anomaly detection, and advice generation, visualised through FastAPI and Plotly.

STACK Azure OCR · GPT · Pandas · NLTK · scikit-learn · PyCaret · FastAPI · Plotly

02 · PropTech / Real Estate

4 PROJECTS

PropertyAI CRM — AI Property Management Platform LIVE DEMO

An AI tenant-communication layer on top of Buildium that autonomously handles tenant SMS — rent follow-ups, maintenance intake (auto-creates work orders), lease-renewal outreach with PDF generation, and smart escalation — with per-tenant conversation memory. Includes a grounded Portfolio Copilot, a Buildium-mirrored dashboard, a WhatsApp-style inbox, role-based admin, and dark mode.

STACK React 18 · TypeScript · Vite · FastAPI · PostgreSQL · OpenRouter (Claude Sonnet 4.5) · Vonage SMS · Buildium REST API · APScheduler · ReportLab · JWT

LINKS propertyai.kognio.ai ↗ · Watch demo ↗

Real-Estate CRM Intelligence Layer LIVE

AI layer connecting four platforms that do not talk to each other for a high-volume real-estate team: Gmail, Follow Up Boss (people), Monday.com (deals), and Slack. Reads the agent's inbox in real time through Gmail Pub/Sub push, runs Claude over every email to classify it, identify each person and their role (buyer, seller, agent, vendor, co-owner), extract purchase-agreement terms from attached PDFs, and keep contact tags, deal stages, and activity notes in sync across the CRM and the deal board, with an **SLA state machine** that escalates missed VIP messages. **Thin fast-ack webhook edges, fat Celery workers**: every LLM call and platform write runs as a task, with nightly reconciliation and an LLM-judged property matcher behind a confidence threshold. All client-specific behavior lives in one policy file, so the same codebase deploys to the next agent by reconfiguring. Validated by **month-by-month dry-run replays** of 500 to 800 real emails per month with per-email verdicts before any live write. ~1,150 tests.

STACK Python 3.12 · FastAPI · Celery + Beat · Redis · PostgreSQL · SQLAlchemy 2.0 · Alembic · Claude Sonnet 5.5 (Anthropic SDK) · Gmail API (Pub/Sub) · Follow Up Boss REST · Monday.com GraphQL · Slack Bolt · Railway

Deal Analyzer (DealScope) — Investment Underwriting LIVE

Underwrites any property deal with a deterministic engine (cap rate, cash-on-cash, NOI, comps) plus an AI Underwriting Agent that streams a multi-strategy fit assessment (buy-and-hold / BRRRR / flip / wholesale) and a Negotiation Strategist that produces an offer ladder and scripts. Kanban pipeline, portfolio dashboard, side-by-side compare, CSV import.

STACK Next.js (App Router) · React · TypeScript · shadcn/ui · FastAPI · SQLAlchemy (async) · OpenRouter (Claude)

LINKS analyzer.kognio.ai ↗

Forensic Lease Audit / Desktop True-Up Assistant DEMO

Agentic system that ingests leases, amendments, prior true-ups, and current invoices; extracts and interprets clauses to validate line items; cross-checks against external data (weather, utility benchmarks — e.g. snow-removal fees vs. actual snowfall); performs year-over-year anomaly detection; and gives clause-cited justification for every decision, with a conversational agent to question and override findings. **Automated ~90% of the manual audit process.**

STACK React · LangGraph · OpenAI · Python · FAISS · External Data APIs

LINKS Watch demo ↗

03 · E-Commerce & Retail Intelligence 4 PROJECTS

Commerce IQ — Commerce Intelligence Platform LIVE DEMO

A founder-facing decision-support platform for Amazon-led D2C brands: ingests commerce data into a canonical warehouse, applies deterministic scoring plus AI workflows, and surfaces "what changed yesterday, what to worry about, what to do" via dashboard, a tool-calling chat agent, real-time alerts, and a daily Telegram digest. Lead-time-aware reorder recommendations, a True-Profit calculator, Buy-Box monitoring, and returns/refunds tracking.

STACK React 18 · TypeScript · Vite · Tailwind · TanStack Query · FastAPI · SQLAlchemy 2.0 (async) · PostgreSQL 16 · Redis 7 · OpenRouter (multi-model) · APScheduler · python-telegram-bot

LINKS commerce.kognio.ai ↗ · Watch demo ↗

Digital Karen — Order Lifecycle Tracker

A multi-agent system that classifies and parses retailer emails (order confirmation, shipping, returns), extracts Order ID / Tracking Number / SKU, and presents a unified end-to-end order-tracking view.

STACK LangGraph · OpenAI · Supabase · Docker

Ecommerce AI — Buyer/Seller Matchmaking

A multi-agent solution that analyses buyer-seller conversations for intent, interest, and temporal patterns to enable intelligent real-time matchmaking; containerised and deployed on AWS.

STACK LangGraph · OpenAI · Docker · Firebase · MongoDB · AWS

Smart Scan AI — Retail Product Extraction

An OCR + LLM pipeline that extracts and structures product details from retail labels (including Icelandic text) in real time to power personalised recommendations and dynamic retail updates.

STACK Azure OCR · GPT-4 · Pandas · SpaCy · FastAPI · Flask · Streamlit

04 · Logistics & Operations

1 PROJECT

MoveOps — AI Operations Platform for Movers LIVE

An AI-driven operations platform for moving companies: a lead pipeline with an event-triggered automation engine, an AI Move Estimator that streams reasoning and returns structured inventory/volume/crew/price, lead-intelligence scoring, jobs & scheduling, invoicing, and a public quote flow. Full light/dark theming and role-based access.

STACK React · TypeScript · Vite · Tailwind · FastAPI · SQLAlchemy (async) · PostgreSQL · OpenRouter (Claude Sonnet 4.5)

LINKS moveops.kognio.ai ↗

05 · HealthTech & Elderly Care

12 PROJECTS

Biologics Intelligence: Citation-Verified Research Assistant

FLAGSHIP LIVE DEMO

The problem. A consultancy in the spinal biologics market answered client questions such as "which of these 49 products are cleared as 510(k), PMA, or 361 HCT/P, and since when?" by reading FDA records, trial registries, journals, and SEC filings by hand. Days per question, and no way to show where each fact came from.

What I built, as sole engineer from first commit to production handover. A research assistant that answers such questions in **about 2 minutes for about $1 of model spend**, with every sentence tied to a passage the reviewer can open. The **ingestion pipeline** keeps a library current from five public sources (FDA 510(k)/PMA/openFDA, ClinicalTrials.gov, PubMed and PMC, SEC EDGAR, DailyMed) plus uploads: PDF parsing with GROBID, chunking, embedding, and product and company **entity resolution** across **33,800 documents, 668,000 passages, and 711 tracked products**. **Retrieval** runs three lanes (vector, keyword, and sentence windows that surface one decisive sentence buried in boilerplate) merged by rank fusion over a **pgvectorscale DiskANN filtered index**. A **five-stage LangGraph workflow** then plans the searches, has a **judge model read every candidate passage** and keep only what supports an answer, writes, and **checks each sentence against its citation** before the reader sees it, refusing when the corpus does not hold the answer.

Results. A hard question went from **6 minutes and $4 to 2 minutes and $1** after retrieval and judge optimisation (filtered index, judging during retrieval, a small judge model on a fast host with fallback). On the client acceptance set, **every regulatory number in the answers was verified correct** against the FDA record. Each answer ships with a live progress view, a source panel per citation, an audit trail of what was searched and kept, and its cost. Delivered with an admin console (product registry, source schedules, user approval, per-stage model and reasoning settings, cost reporting), scheduled backups, and a full operations runbook, administrator guide, illustrated user guide, and cost sheet for handover.

STACK Python · FastAPI · LangGraph · PostgreSQL 16 + pgvector + pgvectorscale · SQLAlchemy 2.0 (async) · Alembic · Celery + Redis · Next.js 15 · TypeScript · Tailwind v4 · OpenRouter (GPT-5.5, gpt-oss-20b on Groq) · GROBID · Docker · Railway · GitHub Actions

LINKS Watch demo ↗

Cure Health — Unified Digital Health Platform

LIVE

A multi-portal platform connecting the whole care loop across six role-scoped dashboards — booking → visit → e-prescription → pharmacy dispensing → insurance adjudication → financing — with AI clinical support (symptom triage, a streaming Care Navigator, clinical-note drafting via a tool-calling agent) and in-browser WebRTC telehealth. 46 API endpoints, 24 role-scoped screens.

STACK Next.js 16 · React 19 · TypeScript · shadcn/ui · FastAPI · async SQLAlchemy · Alembic · PostgreSQL · WebRTC · OpenRouter (Claude)

LINKS curehealth.kognio.ai ↗

Gait Analysis — CV for Rehab & Athletics DEMO

Depth-camera keypoint detection and pose estimation for real-time gait analysis across 15 clinical parameters (step length, knee-flexion, cadence), with historical comparison and personalised suggestions for rehab professionals and athletic trainers.

STACK OpenCV · TensorRT · ROS · MediaPipe · PyTorch · scikit-learn · Azure Kinect / Orbbec · FastAPI · Plotly · ReportLab

LINKS Watch demo ↗

Exercise Module — Rep Counting & Form Feedback DEMO

A lightweight exercise-recognition and rep-counting system from video with real-time form feedback and personalised workout recommendations, deployable on ARM64 edge devices.

STACK OpenCV · TensorFlow Lite · MediaPipe · PyTorch · TensorRT · Flask · Streamlit

LINKS Watch demo ↗

ADL / Transition Detection for Dementia Care DEMO

Real-time recognition of Activities of Daily Living and risky transitions (sit-to-stand, lying-to-sit) to flag unsafe movements and trigger timely interventions in elderly-care facilities and home monitoring — reducing fall risk and improving caregiver response.

STACK PyTorch · MMAction2 · MMDetection · MMCV · OpenCV · Flask · Plotly

LINKS Watch demo ↗

Risky-Behaviour Detection in Post-Op Elderly DEMO

Pose-estimation detection of risky movements (improper knee angle, hip flexion) for hip/knee-surgery patients, integrated with healthcare systems for predictive fall-risk forecasting, personalised prevention plans, and real-time alerts.

STACK TensorFlow · TFLite · ONNX · TensorRT · MediaPipe · Flask

LINKS Watch demo ↗

Agitation Detection in Dementia Patients

A PoseC3D-based system that identifies agitation-related behaviours with real-time monitoring, historical trend analysis, and personalised care suggestions — reducing caregiver burden and response time.

STACK PyTorch · MMAction2 · MMPose · MMDetection · OpenCV · Flask · ROS

Real-Time Facial Recognition on Edge

DeepFace deployed directly on Jetson Xavier for on-device facial recognition, mapping patient activities to identities for secure, low-latency, individualised behaviour tracking in care facilities.

STACK DeepFace · Jetpack · Jetson Xavier · OpenCV · TensorRT · Flask · ROS

Scene Description for Elderly Care

A BLIP-based scene-description module that captions images from care environments in real time to support monitoring and documentation; Dockerised and deployed on AWS.

STACK BLIP · FastAPI · Hugging Face · PyTorch · Docker · AWS

Multi-Modal Empathetic Healthcare Chatbot

A planner-based chatbot that integrates vital signs, medical history, and real-time voice + facial emotion recognition, with patient-persona profiling to generate empathetic, personalised responses.

STACK Opencha · OpenAI · LangChain · DeepFace · Voice/Facial Emotion Recognition · SQLite

Physio Facilitator — Patient Record Summarisation

An LLM-powered chatbot that generates concise, relevant summaries of patient medical histories for clinicians, with modular integration into existing patient-data sources.

STACK Opencha · OpenAI API · LangChain · RAG

Atheris — Medical Data Insights & Visualisation

Scraped large medical datasets and built ETL pipelines for clinical-trial extraction, expert identification, and social-media sentiment analysis on drug trials/conferences, visualised in Power BI via a full-stack app with auth on AWS.

STACK Web Scraping · ETL · OpenAI GPT · Topic Modeling · Power BI · LinkedIn/Twitter APIs · AWS

06 · AI Sales & Outreach

2 PROJECTS

Speed-to-Lead — Multi-Channel AI Intake & Booking Engine DEMO

One lead-qualification engine behind three doorways — website chat (GoHighLevel Conversation AI), a web and phone voice agent (Vapi), and web forms — that answers within seconds, any hour. Each lead is **scored by Claude against rules the business owns in Airtable** (weights, threshold, disqualifiers — change a cell, behaviour changes, no deploy), enriched from the lead's own website, and the assistant books a discovery call on the live calendar mid-conversation. Contact, scored opportunity, note and follow-up task land in whichever CRM the client already uses — **GoHighLevel or HubSpot, switched by one field**. Engineered for production: idempotent intake with a replay guard, a dead-letter path where every failure answers 500 rather than a silent 200, per-contact opportunity upsert, an outbound Google Maps prospecting lane, a nightly metrics roll-up, and a client-facing Lead Desk dashboard. Chat-to-booked-call in about a minute; form-to-CRM in ~20 seconds.

STACK n8n · Claude Haiku 4.5 (OpenRouter) · Vapi (Deepgram STT, Clara voice) · GoHighLevel API · HubSpot API · Airtable + Interface Designer · Apify · Cloudflare Tunnel · Docker · PostgreSQL

LINKS Watch demo ↗

AI-Driven Personalised Sales Outreach

An outreach pipeline that uses LinkedIn data (via ProxyCurl), generates personalised emails with GPT (content in S3), and tracks emails and history in DynamoDB.

STACK Python · GPT (LLMs) · FastAPI · AWS S3 · AWS DynamoDB · ProxyCurl

07 · Conversational AI / Telecom

1 PROJECT

Omvi William — Multi-Tenant Agentic AI SaaS with Real-Time Voice FLAGSHIP DEMO

Designed and built end to end a white-labelled, multi-tenant AI SaaS for a telecom operator (World Mobile). Users talk or type to "William," which answers from the company knowledge base, acts in email and calendar, remembers earlier conversations, and turns a voice call into a project that a team of AI agents completes live. **Architected six services around one agent brain**: a FastAPI gateway that every chat and voice request passes through (Redis sessions and history, SSE streaming, LangGraph onboarding flows with human-in-the-loop pauses); a **standalone LangGraph graph server** so the agent ships independently of the API; a **LiveKit voice worker**; a **pgvector RAG service** with neighbouring-passage context; a **multi-agent orchestrator**; and a Next.js studio.

Agentic core: the LangGraph agent classifies every message and routes it to dedicated **email and calendar orchestrators**, knowledge-base search, **cross-session memory recall** with safeguards against fabricated recall, or a tool-calling responder (knowledge, web search, time, task creation), streaming tokens as they are generated, including the RAG answer stream relayed through the graph. Tasks extracted from each call run through **five specialised agents** (research, plan, build, review, write-up) with live WebSocket progress, per-agent output and token cost, instruction-driven re-runs, and downloadable reports. **Voice:** built the real-time voice agent (LiveKit Agents, Deepgram speech-to-text, Cartesia text-to-speech, Silero VAD, noise cancellation) on the same brain as chat, **patching the LangChain adapter so the graph controls what is**

spoken; clear product questions go straight to retrieval behind an instant spoken filler, warm workers let the agent join a call in about 10 s, and each call's transcript becomes a titled session of linked tasks.

Knowledge and models: pgvector RAG with neighbouring-passage context and **multilingual retrieval**; every model routed through OpenRouter with automatic fallbacks and per-provider token and latency tracking; token streaming fixed for LangGraph on Python 3.10, cutting the first chat reply from **35 s to under 10 s**. **Platform:** designed and implemented the **multi-tenancy model** (organization-scoped state, role-to-scope RBAC, scope-filtered tool binding) and authored the architecture plans the team adopted for serving many clients from one codebase, so a new client is configuration rather than a fork. Shipped to GCP through GitHub Actions and later as **8 services on Railway** with auto-deploy from main and Playwright-verified releases.

STACK Python · LangGraph · LangChain · FastAPI · LiveKit Agents · Deepgram · Cartesia · Silero · PostgreSQL + pgvector · Redis · OpenRouter · WebSockets · SSE · Next.js 16 · TypeScript · Clerk · Docker · Railway · Playwright · GCP · GitHub Actions

LINKS Watch demo ↗

08 · EdTech / LMS

1 PROJECT

AI-Infused E-Learning Platform DEMO

An end-to-end e-learning app with AI agents for content creation, automated marking, exam creation and feedback, and a content-aware chatbot; plus user management, AI games, quizzes, a practice trainer, and an admin dashboard.

STACK LangChain · OpenAI · Supabase · Vercel · Next.js (full-stack)

LINKS Watch demo ↗

09 · Content / SEO

1 PROJECT

AI SEO Blog Generator DEMO

A full-stack app that turns a single keyword (plus tone and length) into a publish-ready SEO blog post — meta descriptions, title suggestions, keyword analysis, structured outline — in under 60 seconds via a 4-step AI pipeline, with a content library, inline editing, DOCX export, and Google SERP preview.

STACK Next.js 15 · React 19 · TypeScript · Tailwind · FastAPI · PostgreSQL · SQLAlchemy · Alembic · Claude Sonnet (OpenRouter) · Docker

LINKS Watch demo ↗

10 · AI Infrastructure / Open Source

1 PROJECT

OptiML — LLM Compression & Fine-Tuning Framework

An open-source framework that turns fine-tuning into a compression pipeline, so models come out smaller and ready for production. One config-driven API covers multiple techniques — **pruning (magnitude, Wanda, SparseGPT), quantization, LoRA, and MLA** — on a dynamic graph execution model that adapts to the target hardware. Scales across multi-GPU setups (NVIDIA and Intel) via PyTorch DDP, with SageMaker JumpStart pipelines and lm-eval benchmarking built in. Also built the supporting GPU infrastructure: a control plane that rents cost-capped Vast.ai GPUs into a K3s cluster (Ray, FastAPI job API, Tailscale networking) and serves the compressed models for **self-hosted, optimised inference** via Text Generation Inference. Led a team of 4 — architecture, CI/CD, and releases.

STACK Python · PyTorch · Hugging Face · AWS SageMaker JumpStart · Ray · Kubernetes (K3s) · FastAPI · Text Generation Inference · Docker · GitHub Actions

11 · HR Operations / Agent Tooling (MCP)

1 PROJECT

AI Onboarding Ops — Employee & Client Onboarding as an MCP Server LIVE

Onboarding operations delivered two ways from one codebase: a web console for people, and a **Model Context Protocol (MCP) server** that Claude Desktop, Claude in Slack, or any agent can call with a personal token. One instruction starts an employee onboarding — person record, task list from a template, document requests with upload links, account provisioning — and tasks that need paperwork stay closed until the document is actually accepted; client onboarding (contract, workspace, billing, security review, kickoff, go-live) runs on the same state machine. Policy questions are answered only from the company's own **versioned handbook**, with a citation for every claim and a refusal when the documents are silent. The design centre is **25 narrow tools behind a single guard** that runs role check → idempotent replay → handler → audit write, on success and on failure, so the console, the in-app chat agent, and the MCP endpoint cannot change data without a record. Five roles with row-level scoping, restricted personal data behind a separate audited call that is never exposed to chat or MCP, and every model call priced at write time from a versioned price table with daily/weekly/monthly cost views. 24 tools, 3 resources, and 3 prompts over a hand-rolled JSON-RPC 2.0 MCP endpoint; the seed data is produced by driving the real tools, so the audit log is genuine.

STACK Python · FastAPI · SQLAlchemy 2.0 (async) · Alembic · PostgreSQL · MCP (JSON-RPC 2.0) · BM25 retrieval · OpenRouter (Claude Sonnet 4.5) · React 18 · TypeScript · Vite · TanStack Router · Tailwind · Docker · Railway

LINKS onboarding.kognio.ai ↗ · Case study on GitHub ↗