

Mudassar Bin Qaiser

Senior AI & ML Engineer | Agentic AI Systems & Cloud/On-prem Architecture

Lahore, Pakistan · +92 310 4779862 · mudasserqaiser14@gmail.com · [linkedin.com/in/mudassarbinqaiser](https://www.linkedin.com/in/mudassarbinqaiser) · github.com/mudassarbinqaiser

PROFESSIONAL SUMMARY

Senior AI/ML Engineer with 3+ years designing and shipping enterprise-grade, cloud-native agentic AI platforms end to end, from problem framing and architecture through production reliability. Owns the whole system rather than the model layer — agent runtime, transport, data stores, deployment, and the operational cost of each choice. Selects datastores and transports on access pattern, data lifetime and scaling ceiling rather than defaulting to one engine, and reverses an architecture when the evidence justifies it: consolidated a thirteen-agent topology into a single agent and removed roughly 93,000 lines after running both in parallel behind a feature flag. Deep hands-on experience with AWS-centric AI systems (Bedrock, Nova, Bedrock Data Automation, Lambda, ECS, S3, IAM), model selection, prompt orchestration, RAG versus agentic retrieval strategy, and cost-aware design. Backend platform depth in Python and TypeScript, with on-premise delivery under data-sovereignty constraints.

CORE SKILLS

Agentic AI & LLMs: Agentic AI system design, multi-agent orchestration (LangChain, LangGraph, DeepAgents), ReAct, RAG and Agentic RAG, prompt orchestration, tool and function calling, MCP, structured output enforcement (Pydantic), model evaluation, guardrails, hallucination and PII controls, human-in-the-loop, LLM observability (Langfuse); GPT-4 / GPT-4o, Claude, Gemini, DeepSeek, LLaMA, Amazon Nova; Hugging Face, Semantic Kernel, AutoGen, CrewAI, LlamaIndex

Data, Storage & Scaling: Schema design and data modelling, multi-tenant isolation (PostgreSQL row-level security), polyglot persistence and store selection by access pattern, caching, TTL and eviction strategy, horizontal-scaling constraints and remediation, memory and latency profiling; PostgreSQL, MySQL, MongoDB, Redis / Redis Stack, Memgraph, Kafka; vector databases including Pinecone, Chroma, Weaviate, FAISS, pgvector

API & Transport Design: REST API design, event-driven and pub/sub architecture (MQTT, EMQX), delivery-guarantee and QoS trade-offs, token streaming protocols, WebRTC media planes, backpressure and reconnection handling, contract and versioning design

Solution Architecture & Delivery: Requirements analysis and solution scoping, effort estimation, architecture and design documentation, phased delivery planning (MVP to scale to autonomy), architecture reviews, stakeholder and executive communication, acceptance testing and sign-off, production incident triage, AI engineering standards, mentoring and technical onboarding

Voice & Conversational AI: LiveKit, STT and TTS pipelines, voice activity detection, turn detection and endpointing, real-time streaming, latency and concurrency tuning, multilingual agents (English, Urdu)

Backend & Full Stack: Python (FastAPI, Django, DRF), Node.js, NestJS, Golang, TypeScript, React, Next.js, Flutter, REST API design

Cloud & Deployment: AWS (Amazon Bedrock, Bedrock Data Automation, Nova, Textract, Lambda, ECS, EC2, S3, IAM, VPC), Docker and Docker Compose, on-premise and air-gapped delivery, private registries, CI/CD; working exposure to Azure, GCP, Alibaba Cloud

Automation & Integration: n8n, Make.com, RPA scripting, REST APIs, legacy and third-party system integration, data pipeline automation

Domain Experience: Telecom network operations (FTTH, PON/OLT, optical telemetry), IBM Watson AIOps / CP4AIOps, enterprise customer support, recruitment marketplaces, real estate, content moderation and compliance

PROFESSIONAL EXPERIENCE

Senior AI/ML Engineer — AirCod Technologies

Oct 2025 – Present · Lahore, Pakistan · Full-time

Leading the design and production delivery of enterprise agentic AI systems and telecom AIOps integrations, owning two flagship initiatives.

AAIRA — Enterprise Agentic Customer Support Platform *(end-to-end owner)*

- Own the architecture and production delivery of a conversational AI platform serving a telecom operator's customer base across chat and voice, deployed on-premise via Docker to satisfy data-sovereignty constraints that ruled out SaaS agent platforms (LangGraph, FastAPI, Python 3.13, EMQX, Redis Stack, Memgraph).
- Led the consolidation from a thirteen-agent orchestrator topology to a single agent with fourteen runtime-loaded Markdown workflows**, after running both architectures in parallel behind a feature flag: removed roughly 93,000 lines, eliminated the cross-agent state loss occurring at handoff boundaries, and reduced prompt maintenance from thirteen files to one.
- Designed routing as an authorisation boundary: the agent loads a domain workflow through a gated filesystem backend that runs an entitlement check on read, so intent routing and access control resolve at the same point across a 47-tool surface.
- Selected a streaming XML response envelope over JSON mode and structured-output APIs because partial JSON is unparseable — enabling token-by-token text and field-by-field widget rendering in the browser rather than buffer-then-render; accepted the loss of schema validation as the explicit trade-off and specified schema-generated prompt and parser as the remediation.
- Rebuilt the personalisation layer** from an LLM-rewritten profile document to a deterministic projector mapping executed tool calls into a Memgraph knowledge graph, so facts enter memory because a tool ran rather than because a model remembered; LLM scope narrowed to behavioural signals only.

- Designed the four-store state model — Redis for short-term messages and session state, LangGraph checkpoints for conversation history, Memgraph for the user graph, FAISS for the diagnostics corpus — choosing per store on access pattern and data lifetime rather than standardising on one engine.
- Built the real-time voice layer as a separate LiveKit worker holding no LLM: speech-to-text publishes to the same MQTT topic as the web client, so chat and voice share one session, one checkpoint and one tool set, and users switch modality mid-conversation with no state transfer.
- **Cut voice time-to-first-word from over 10 seconds to sub-second** by replacing full-response buffering with a bounded async queue streamed directly into TTS; tuned VAD thresholds, STT endpointing and a two-phase turn detector to stop the agent interrupting itself.
- Kept the runtime portable across OpenAI, Anthropic and DeepSeek behind a single provider switch, including normalisation of divergent streaming payload shapes and system-message handling that had silently produced empty responses in production.
- Implemented the access-control layer required by the client's compliance review: entitlement gating at workflow-load time, prompt-injection defence, and authentication hooks in the middleware pipeline, with denial paths written so the agent never discloses internal systems to the customer.
- Delivered English and Urdu on the same runtime, and a single agent codebase serving both chat and voice through a unified message contract.
- **Produced the scaling analysis and remediation path** off single-instance deployment: session indices moved from in-process state into Redis, leader-elected MQTT subscription ownership to allow more than one replica, and TTL and eviction policy on conversation checkpoints and unbounded profile documents.
- Operate the system in production: diagnosed and resolved TTS provider timeouts, silent stream-discard failures, checkpoint corruption from orphaned tool calls, and voice-activity-detection CPU starvation, using Langfuse tracing and latency instrumentation.

ZALA Studio — Multi-Tenant Agent Platform *(companion product)*

- Designed the enterprise multi-tenant platform for building, deploying and managing agents on the same runtime: a 114-table PostgreSQL schema with row-level-security tenant isolation, .NET modular monolith backend, and Flutter cross-platform client.
- Delivered full Google Agent-to-Agent (A2A) Protocol v0.3 compliance with automatic Agent Card generation, real-time compliance scoring, and a nine-step agent-creation wizard, so agents export cleanly for cross-vendor interoperability and marketplace publishing.

AIOps — IBM Watson AIOps Integration for Telecom Network Operations

- Serve as AirCod's embedded technical resource inside a Latin American telecom operator's engineering team, working alongside client engineers and a third-party vendor team on IBM Watson AIOps / CP4AIOps.
- Ran discovery and scoping on Bad WiFi, PON/OLT failure and failure-prediction use cases, converting operational pain points into staged ML pipelines covering label design, feature engineering over FTTH KPIs and optical-power telemetry, and technician dispatch.
- Integrated the AIOps platform with the client's existing Servassure and ITOS systems, and built Python and RPA automation to close gaps the platform did not cover out of the box.
- Debugged pipeline failures inside the client's environment across multiple production cases and drove each to resolution jointly with the vendor engineering team, coordinating across three organisations without direct authority over any of them.
- Produced the artefacts the engagement is judged on: architecture reference documents, acceptance-test evidence, and executive-facing case reports for client leadership.

Technical Leadership

- Lead architecture reviews and translate business objectives into execution-ready designs with explicit in-scope and out-of-scope boundaries and phased delivery (MVP to scale to autonomy).
- Establish AI engineering standards across the team covering model evaluation, prompt orchestration, vector-retrieval optimisation, inference efficiency and production reliability.
- Mentor AI/ML engineers on multi-agent design, LLM orchestration and production-grade AI system design, using architecture reviews and written design documents as the teaching surface rather than ad-hoc correction after the fact.
- Produce the protocol and interface documentation the rest of the stack integrates against — the MQTT message contract, topic map and streaming envelope — so client and voice teams can build against the platform without direct hand-holding.
- Documented the platform's architecture, decision history and failure modes — routing, transport, memory, deployment and the reasoning behind each reversal — as the onboarding reference for engineers joining the system.

AI/ML & Backend Engineer — GoodToGo

Aug 2024 – Oct 2025 · Lahore, Pakistan · Full-time

- Owned end-to-end solution architecture and delivery of AI systems across two production platforms (NewZhenghe, StrangerUs), translating business requirements into scalable, cloud-native designs and shipping MVPs in five to eight week cycles while holding SLA compliance.
- **Replaced Elasticsearch-based matching with two custom ML recommendation systems** (employer to candidate, job to jobseeker), removing a managed search cluster from the critical path and materially reducing infrastructure cost.
- Designed and deployed an agentic AI content-moderation platform (ReAct, LangChain, LangGraph) with Explainable AI, pre-persistence guardrails and field-level explainability, so moderation decisions were auditable by compliance stakeholders rather than opaque.
- Packaged moderation as reusable infrastructure via decorators and interceptors, giving consistent policy enforcement across backend services without per-team rework.

- Architected AWS-native deployments (EC2, S3, Lambda) tuned for cost, latency and scale across LLM inference and data pipelines.
- Led backend architecture for multilingual, multi-region platforms, improving user retention by 30%.

AI/ML Full Stack Engineer — BotBrain

Jul 2023 – Aug 2024 · Wyoming, USA · Remote

- Designed and delivered a customer-deployed agentic AI chatbot platform, focused on multi-agent orchestration, autonomy and production-scale reliability rather than standalone chatbot features.
- Architected multi-agent workflows (ReAct, LangChain, LangGraph) supporting task planning, execution and validation inside complex end-user workflows.
- Built Agentic RAG pipelines with controlled context boundaries, memory strategies and prompt orchestration; **reduced response latency by 25%** through vector-retrieval and embedding optimisation.
- Shipped autonomous document-processing pipelines that **reduced manual support effort by 60%**.
- Standardised deployment architecture and automated training workflows, **cutting new-customer deployment time by 40%**.
- Ran feedback-driven refinement cycles directly with customers, improving customer satisfaction by 30%.

EDUCATION

Bachelor of Science in Computer Science — Government College University, Lahore · 2023

SELECTED PROJECTS

- **RealtorAI — Real Estate Automation Suite.** Multilingual voice agents orchestrating lead capture, cold-calling, CRM integration, call transcription and scheduling, with GDPR-compliant workflows balancing autonomy against regulatory auditability.
- **NewZhenghe — Multilingual Job Marketplace.** Multi-tenant platform supporting five user types across Chinese, Japanese, Korean and Singaporean markets, with scalable data models, localisation strategy and AI-driven employer-candidate matching.
- **StrangerUs — Real-Time Social Application.** Location-aware social platform using QR codes and geo-hashing for low-latency discovery, with a backend optimised for moderation, user safety and real-time scale.
- **BotBrain — AI-Powered Website Chatbot Builder.** Agentic chatbot platform on LangChain and LangGraph enabling autonomous task execution, controlled retrieval and feedback-driven continuous learning.
- **Digital Human Generator (SadTalker).** Digital human generation system integrating conversational agents with avatar-driven visual output, with controlled interaction flows and production-ready deployment pipelines.
- **Real-Time Science Tutor (LLM Avatar).** Multilingual, voice-based AI tutoring system delivering personalised, age-adaptive learning experiences with conversational control flows.
- **AI Course Planner (RAG + LLM).** Agent-driven curriculum planner using RAG and vector retrieval to generate personalised learning paths, schedules and content recommendations.
- **Voice-Activated Visual Shopping Assistant.** Mobile voice assistant performing on-screen object detection, masking and reverse image search to surface visually similar products in real time; final-year university project.

LICENSES & CERTIFICATIONS

Full list of licenses and certifications available on LinkedIn.