Afzal Khan

My_Website | afzaljawadkhan@gmail.com | +92 337 9755627 | Afzal Khan

SUMMERY

AI Engineer | Machine Learning Engineer | Data Engineer | MLOps Engineer
Results-driven professional with 3+ years' experience designing agentic ecosystems, optimizing deep learning models, and deploying scalable AI pipelines. Skilled in LLM-based automation, data warehouse design, and edge inference using modern MLOps stacks (Docker, FastAPI, Kubernetes).

EXPERIENCE

ClaimbAI — AI Engineer | Remote, California, USA (Jul 2025 – Present)

- Architected and shipped an agent-first ecosystem (Google ADK, n8n, Firebase Functions, Cloud Run) powering web + iOS/Android frontends; delivered end-to-end deployment for **6 secure connectors** and role-based access controls.
- Designed and implemented Retrieval-Augmented Generation (RAG) and local SLM workflows; integrated PostgreSQL vector store + LlamaIndex patterns to enable sub-second retrieval for knowledge access.
- Built production REST APIs (FastAPI) and CI/CD pipelines; automated blue/green deploys and monitoring, reducing rollback time and improving deployment reliability by **40%**.
- Implemented automated questionnaire → Drive → JSON → recommendations pipelines that feed UX components; improved throughput and observability and reduced manual processing steps by **70%**

COMCEPT — Software Engineer | Islamabad, Pakistan (Jan 2025 – Oct 2025)

- Engineered real-time computer vision and multi-object tracking stacks (YOLOv9–v11, GOTURN, CSRT, SAMURAI) and deployed optimized inference on NVIDIA Jetson Orin (aarch64).
- Optimized models via TensorRT, pruning, and quantization — improved inference throughput by **~30%** and reduced false positives by **25%**.
- Developed C++/Qt visualization and monitoring dashboards (40+ screens) to surface telemetry and reduce debugging time by **90%**
- Led system integration across firmware, Radxa/Armbian compute nodes, and networked devices; reduced system boot time and increased uptime by **95%.**
- Applied edge-deployment best practices (ONNX, aarch64 tuning) and performance profiling to meet latency and power constraints.

AIGENIUS — Chief Operating Officer (COO) | Islamabad, Pakistan (Jul 2024 – Dec 2024)

- Directed operations and recruitment while contributing technically; scaled engineering headcount and established hiring pipelines to support product roadmap.
- Led a major data transformation: converted a **200M-row** dataset across **~200** tables into a normalized 3NF data warehouse and developed star-schema semantic layers for analytics and forecasting.
- Built and deployed a Node.js/React agent chatbot and automated ETL using AWS Glue + SQL, improving dashboard refresh times and reducing manual ETL interventions by **4 hours**.
- Delivered proofs-of-concept including player tracking (OpenCV) and production scraping pipelines; improved client reporting cadence and decision-making through automated dashboards.
- Managed client relationships and aligned cross-functional teams (sales, engineering, analytics) to deliver measurable KPIs and on-time deliverables.

Convergent Business Technologies — Consultant, Analytics & Insights | Islamabad, Pakistan (Aug 2023 – Jul 2024)

- Implemented market-basket analysis (Apriori), dimensional modeling, and Tableau dashboards; automated refresh pipelines to reduce manual reporting workload.
- Performed ETL to transform raw sources into 3NF and built a dimensional model semantic layer for analytics consumption.
- Automated Pepsi KSA & UAE analytics project — reduced refresh cycle from 1 week → 3 days (57% reduction) and cut required team size from 6 → 2 through query tuning and scripting.
- Built ML pipelines for transcription and audio analytics (OCR + timestamped ASR + post-processing) for aviation prototyping, enabling time-aligned analytics and faster model iteration.

Upwork — Freelance Data Engineer and Developer | Remote (Aug 2022 – Aug 2023)

- Delivered end-to-end analytics and predictive models (churn, forecasting, inventory optimization) for telco, healthcare, and finance clients; owned data pipelines, feature engineering, model training, and deployment.
- Built automated ETL workflows using SQL and AWS Glue; decreased data latency and improved report freshness by 25%.
- Created Power BI / Tableau dashboards and scheduled reports for executive stakeholders; trained client teams on analytics best practices to ensure adoption.
- Conducted statistical analyses using pandas, NumPy, and scikit-learn to derive actionable insights that informed client strategy.

CORE COMPETENCIES

- **AI / LLMs & Agents:** LangChain, LlamaIndex, RAG, Generative AI, local SLMs (Ollama/Self-hosted), Google Agent Development Kit, n8n, zapiar, make, gohighlevel
- **APIs & Integrations:** OpenAI, Anthropic, Gemini, REST/FastAPI, Zapier/Make, webhook-based integrations, OAuth/connectors for email/calendar/tasks
- **ML / DL:** PyTorch, TensorFlow, Hugging Face, Transformer models, Whisper, Librosa, YOLO (v9–v11), Predictive Analytics
- **Model Optimization & Edge:** TensorRT, ONNX, quantization, pruning, NVIDIA Jetson Orin (aarch64), Edge AI, ARM64 tuning
- **MLOps & Cloud:** Docker, Kubernetes, Cloud Run, Firebase Functions, CI/CD, MLflow, AWS Glue, SageMaker (familiar)
- **Data & Analytics:** ETL, Data Pipeline Automation, 3NF & dimensional modeling, PostgreSQL (incl. vector DB usage), SQL, Tableau, Power BI

PROJECTS

- **ADVANCED SPORTS PERFORMANCE ANALYTICS PLATFORM:** Architected and deployed predictive models for player positioning and performance metrics; delivered interactive dashboards used by coaching staff to improve tactical decisions and reduce match analysis time by 45%.
- **FACE RECOGNITION SOFTWARE:** Developed a high-performance facial recognition pipeline (MATLAB + transfer learning on AlexNet) achieving 92% classification accuracy on a custom dataset of 10,000+ images; optimized batch inference to improve throughput by 35%.
- **AI AGENT ECOSYSTEM FOR BUSINESS AUTOMATION:** Architected a multi-agent automation system across Sales, HR, Marketing, and Engineering; central controller orchestrated 10–15 specialized agents per department, automating repetitive tasks and reducing manual workload by ~60%.
- **ADAPTIVE LANGUAGE LEARNING WEB APP:** Firebase + Gemini API; LLM-curated lessons, conversational practice, and progress tracking.
- **LOCAL RESEARCH AI AGENT:** Local-only agent using n8n + Ollama + SERP API; used PostgreSQL vector DB for knowledge retrieval and persistent memory.
- **PERSONAL ASSISTANT AI AGENT**: Developed a personal assistant agent using OpenAI's API and n8n to manage calendars, Gmail, Telegram, WhatsApp, and other personal tasks.
- **CLAUDE DESKTOP MCP CLIENT:** Created a local MCP client for Claude Desktop, enabling the AI to add, remove, and read files from the local machine.

EDUCATION

MIT | MicroMasters, Statistics and Data Science (In Progress) Virtual

Courses: Probability - The Science of Uncertainty and Data, Statistics and Data Science

NUST | Mechatronics Islamabad, Pakistan

Bachelors, Mechatronics Engineering

CERTIFICATIONS

- McKinsey Forward Program | McKinsey
- Tableau Data Analysis | Philip Burton
- Google Data Analytics | Coursera
- Advanced SQL for data analysis | Udacity
- UI/UX Design | Calarts
- CS50 | Harvard

SKILLS & TECHNOLOGIES

- **AI / LLMs & Agents:** LangChain, LlamaIndex, RAG, Generative AI, Ollama, n8n, Google ADK
- **Machine Learning:** PyTorch, TensorFlow, Hugging Face, Transformer Models, YOLO, Whisper
- **MLOps & Cloud:** Docker, Kubernetes, FastAPI, MLflow, Firebase, Cloud Run, AWS Glue, SageMaker
- **Data Engineering**: SQL, PostgreSQL, Vector Databases, ETL, Data Pipeline Automation, 3NF & Dimensional Modeling
- **Analytics & Visualization:** Tableau, Power BI, Predictive Analytics
- **Programming:** Python, C++, Node.js, React, Qt
- **Additional Expertise:** Prompt Engineering, Model Deployment, API Integration, Edge AI, CI/CD