

MUDASSAR AWAN

AI/ML Engineer · Automation · Voice Agents · LLM Training · Computer Vision · MLOps · FAST NUCES '25 (AI)

writetomudassarawan@gmail.com · github.com/mudassar531 · linkedin.com/in/mudassar-a1b1c1d1 · Pakistan · Remote

CAREER SUMMARY

AI/ML Engineer with hands-on experience designing, training, and deploying ML models across deep learning, NLP, computer vision, and large language models. Proficient in Python, PyTorch, TensorFlow, and Keras, with production experience deploying models as REST APIs on AWS and Azure. Contributed to Qalb — the world's largest Urdu LLM (1.97B tokens, SOTA on 7 benchmarks for 230M+ speakers). Currently building across Agents Limited (CounterVision), Ember, and NavAI, with a focus on **automation, real-time voice agents, and LLM training**.

WORK EXPERIENCE

Product Owner — Agents Limited (CounterVision)

Jan 2026 – Present

agentslimited.com/countervision · Remote

- Own the product and led CounterVision v2 — real-time person tracking and face recognition — benchmarking 10 model+tracker configurations on an NVIDIA RTX 5060.
- Architected the multi-stage pipeline: YOLO26 detection, ByteTrack/BoT-SORT tracking, SCRFD-10G face detection, ArcFace w600k_r50 512-d embeddings — 96.4% tracking continuity at 18.7 FPS.
- Engineered quality-gated recognition (~60-70% compute saved), multi-frame embedding fusion, clothing-color HSV histograms for body re-ID, and self-correction logic.

Software Engineer — Ember

May 2026 – Present

ember.new · Remote

- Shipped a meeting-recorder connector end to end across an Onyx-based AI knowledge & operator-brain platform: backend connector, admin-panel UI, and the operator-brain auto-trigger pipeline.
- Indexing → semantic search → cited answers; validated end to end and merged to main with full CI and adversarial review.

Voice AI Engineer — NavAI

Apr 2026 – Present

navai.pro · Tashkent, Uzbekistan (Remote)

- Built the end-to-end Uzbek voice-agent pipeline: custom STT/ASR + TTS, GPT-4o, LiveKit real-time transport, and SIP telephony (migrated off Yandex).
- Stood up a full observability platform — Prometheus + Grafana + cAdvisor + Node Exporter on GCP, Langfuse tracing, and Telegram alerting via a webhook relay.
- Built a model playground + A/B testing harness and a Post-Call QA agent (MongoDB); led the NAV-175 root-cause investigation across 750+ calls.

Junior AI Engineer — CareCloud MTBC

Jul 2025 – May 2026

carecloud.com · Rawalpindi, Pakistan

- Built a claim-denial classifier on millions of 837-EDI insurance claims: 99%+ accuracy on accepted, 80%+ on denied.
- Shipped a CPT/ICD mapping RAG pipeline + clinical-document classification, replacing manual workflows for 500+ staff.
- Built an AI dashboard over 2007–2024 clinical records and contributed to "Status AI", a voice agent for doctors.

AI Research Intern — Genesis Lab

Feb 2024 – May 2024

- Researched and implemented LLMs, GANs, and RAG systems for document automation.

ML Intern — IMARAT Group of Companies

Sep 2023 – Dec 2023

imarat.com.pk · Islamabad, Pakistan

- Trained classification/regression models for real-estate valuation and optimized YOLOv8 detection pipelines.

PROJECTS

Hearsay — Open-Source Media → Markdown Library — MIT · Python · github.com/mudassar531/hearsay

- "crawl4ai for video & audio" — one command turns any YouTube video, podcast, or local recording into clean, timestamped, chunked, LLM-ready markdown. Solo-built.
- Captions-first with automatic faster-whisper fallback; batch playlists & podcast feeds; JSON sidecars for RAG; ships an MCP server for Claude and other agents.

Qalb — World's Largest Urdu LLM — arXiv: 2601.08141 · LLaMA 3.1 8B, PyTorch, crawl4ai, HuggingFace, LoRA

- Contributed to the first dedicated Urdu LLM (1.97B tokens) for the 230M+ speaker gap; scraped 1.84B+ tokens with a 67.8%-retention cleaning pipeline.
- 90.34 weighted average across 7 Urdu NLP benchmarks — outperforming prior SOTA by +3.24 points.

VANOS — Voice Agent Neural Operating System — Pipecat AI, Deepgram Nova-3, Groq, Kokoro TTS, Silero VAD

- Real-time voice AI desktop assistant with sub-second latency on consumer hardware; implemented 5 Voice-to-Action (V2A) tools and solved critical mic-compatibility challenges.

Claim Denial ML Pipeline — End-to-End MLOps — Python, Scikit-learn, PyTorch, FastAPI, AWS, Docker

- Complete pipeline for 837-EDI ingestion, classification, and live claim validation with automated retraining.

DashGrab — Smart Retail Automation (FYP) — YOLOv11, Detectron2, OpenPose, MediaPipe, SAM

- Computer-vision system detecting items picked by shoppers and auto-generating bills.

EDUCATION

B.S. Computer Science — Artificial Intelligence · FAST University Islamabad (FAST NUCES) · 2021 – 2025

SKILLS

Languages: Python, SQL, JavaScript

Computer Vision: YOLO26/v11, RT-DETR, InsightFace (SCRFD, ArcFace), ByteTrack, BoT-SORT, OpenCV, SAM, OpenPose, MediaPipe

LLMs & NLP: Continued pre-training, LoRA/PEFT/QLoRA, RAG, LangChain, OpenAI/Groq, Ollama, tokenization, crawl4ai, MCP

Observability: Prometheus, Grafana, cAdvisor, Node Exporter, MongoDB aggregation, Telegram Bot API

ML Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, HuggingFace, Ultralytics, Unsloth

Voice & Speech AI: Pipecat, Deepgram Nova-3, Kokoro TTS, Silero VAD, LiveKit, SIP Telephony, Groq, ElevenLabs, Langfuse

Deployment: FastAPI, Flask, Docker, AWS, Azure, ONNX Runtime (GPU), Streamlit, NVIDIA A100/RTX

Soft Skills: Product Ownership, Technical Writing, Research, Root-Cause Analysis, Benchmarking