Ahmed Usman

AI/ML Engineer

+923350707006 | ahmadusman050@gmail.com |
github.com/ahmedosm0 | linkedin.com/in/ahmedusman050

Objective

AI/ML Engineer specializing in production LLM applications, RAG systems, agentic workflows, voice automation, and applied computer vision. Experienced in building full-stack AI products with FastAPI, Next.js, Supabase/PostgreSQL, LangGraph, vector databases, and cloud deployment. Strong record of shipping practical AI systems that combine model orchestration, retrieval, automation, monitoring, and user-facing product design.

Skills

Programming Languages: Python, JavaScript, TypeScript, C++
Web Frameworks: FastAPI, Django, Flask, Node.js, Express.js
Databases: PostgreSQL, Supabase, MongoDB, MySQL, SQLite, Redis, Pinecone, ChromaDB
AI/ML: TensorFlow, Scikit-learn, OpenCV, DeepFace
LLM & Agent Systems: LangChain, LangGraph, CrewAI, LangSmith, RAG, vector search, semantic caching, tool calling
LLM Providers: OpenAI, Anthropic, Mistral, Groq, Qwen, DashScope, OpenRouter, Ollama, Hugging Face
Speech & Voice: Retell AI, ElevenLabs, Whisper, Coqui TTS, Librosa, SciPy
Automation & Integrations: n8n, Shopify GraphQL, Twilio, Stripe, Buffer, Gmail API, Google Sheets, Klaviyo, GA4, PostHog
MLOps & DevOps: Docker, Jenkins, Git, GitHub, CI/CD, Railway, Model observability
Soft Skills: Communication, Team Collaboration, Problem-Solving, Fast Learning
Languages: English (C1), German (A2)

Experience

Associate AI Engineer at Tech Emulsion - (full time) (Jul, 2025 - Present)

- Design and deploy LLM-driven conversational agents, RAG pipelines, and automation workflows by integrating APIs, structured backends, and third-party tools to deliver scalable AI solutions.
- Develop and maintain backend services and integrations using Python, Django, FastAPI, n8n, and AWS, collaborating with cross-functional teams to deliver end-to-end AI products.

Jr AI/ML Engineer at DevK System - (full time) (Aug, 2024 - Jun, 2025)

- Designed and developed LLM applications including RAG systems, autonomous agents, and agentic workflow prototypes covering the full lifecycle from research to deployment.
- Designed and deployed ML models with attention to latency, scalability, and maintainability; applied MLOps practices using Docker, CI/CD pipelines, and model monitoring.

Back-End Developer at Brandora - (part time) (Sep, 2022 - Jul, 2024)

- Developed REST APIs with Node.js, Express.js, and MongoDB; integrated backend services with React frontends and implemented authentication, query optimisation, and reusable API modules.

Projects

Full-Stack AI Platforms

The Meatery – E-Commerce Intelligence & AI Voice-Agent Platform | *Next.js 14, TypeScript, Retell AI, n8n, Shopify, OpenAI, Anthropic, Twilio*

- Built a full-stack revenue-operations and analytics platform for a US premium-meat e-commerce brand, unifying margin/COGS monitoring, inventory velocity tracking, reorder automation, marketing attribution (Klaviyo, GA4, Search Console, PostHog), competitor price scraping, and GrowthBook A/B pricing experiments within a single operator dashboard.

- Engineered Retell AI inbound and outbound calling agents for abandoned-cart recovery, win-back campaigns, prospecting, and post-delivery support backed by an Express.js callback server that performs live Shopify cart lookup, generates dynamic discount codes, delivers SMS via Twilio, enforces DNC lists, applies per-lead cooldowns, and prevents duplicate calls.
- Designed nightly n8n pipelines that analyse call transcripts with Claude to produce objection rebuttals, competitive battle cards, and prompt-improvement suggestions, then automatically sync the updates into agent-specific Retell knowledge bases; added a streaming LLM concierge shopping agent and AI-driven product-content generation.

AVL Copilot – AI Technical-Support Copilot | *FastAPI, LangGraph, Pinecone, Redis, Supabase, Stripe, OpenAI Responses API*

- Built and deployed a multimodal RAG and agentic support assistant that helps field technicians troubleshoot Audio, Video, and Lighting (AVL) equipment, retrieve manufacturer manuals, and answer complex engineering questions in real time via FastAPI with server-sent event (SSE) streaming.
- Orchestrated a strictly sequential LangGraph pipeline, semantic cache check, conversation summary, intent detection, Pinecone RAG retrieval, query routing, Serper/ScraperAPI web-search fallback with Jina Reader content extraction, streamed generation, enforced with circuit breakers, domain allowlisting, and manufacturer-support reranking.
- Added dual-tier semantic cache (Redis), image-based diagnostics, dynamic token budgeting, tiered model routing, Supabase auth with conversation checkpointing, Stripe billing with per-user quotas, a PDF/URL manual ingestion pipeline, metrics endpoints, and a live SSE log viewer for real-time observability.

Good Food Project – AI Social Content Engine | *FastAPI, Supabase, LangGraph, Next.js 15, React 19, TypeScript, Cloudflare R2*

- Built an AI-powered content generation platform for a UK organic food brand, using a LangGraph multi-node pipeline to create brand-aligned social media posts and route them through structured human review and approval workflows before scheduled publishing.
- Implemented a FastAPI and Supabase backend with async background jobs, multi-provider LLM routing with rate-limit handling and token-cost tracking, RAG over a brand knowledge base (PG Vector), computer-vision image matching, automated quote-image composition, Buffer API and APScheduler for scheduled publishing, Cloudflare R2 for asset storage, and Sentry for observability.
- Delivered a canvas-based image editor built with react-konva, Zustand for state management, TanStack Query for data fetching, Supabase SSR authentication, and Cypress end-to-end test coverage on the frontend..

AI Health Receptionist Voice Agent | *ElevenLabs, Dentally, Pabau, Payment APIs*

- Developed a healthcare voice agent using ElevenLabs TTS/STT that autonomously handles appointment booking, cancellation, and rescheduling integrated with Dentally for real-time calendar management and payment collection within the same call flow, eliminating routine front-desk calls entirely.

Mobile Applications

Lost-N-Fond – AI Campus Lost and Found Mobile App | *React Native, Expo, FastAPI, Supabase, WebSockets, Mistral*

- Built a cross-platform mobile app (React Native + Expo) that digitises the campus lost-and-found process students report items, admins review AI-assisted ownership claims, and claimants are notified in real time via WebSockets.

Education

University of Engineering and Technology, Peshawar
B.S. Electrical Computing & Communication Engineering

Certifications

- Machine Learning Specialization
- Deep Learning Specialization
- Machine Learning in Production
- TensorFlow for AI, ML and DL
- Complete Generative AI Course With LangChain and Hugging Face