

MUHAMMAD SHAHZAD

FULL-STACK AI ENGINEER | ML/DL SPECIALIST

+92-320-6236425 | shahzad.abdulmajeed4894@gmail.com | Lahore, Pakistan

SUMMARY

Full-Stack AI Engineer and ML/DL Specialist with over 5 years of experience bridging the gap between sophisticated machine learning architectures and production-grade web applications. Proven track record as a technical founder, engineering multi-agent cognitive systems (CrewAI/LangChain) and launching commercial AI SaaS platforms that scale to thousands of active users. Expert at transforming raw predictive models into high-throughput backends (FastAPI, Django, Node.js) that solve real-world user bottlenecks and reduce operation cycles.

PROFESSIONAL EXPERIENCE

OpenWriterAI | Lahore, Pakistan

OpenWriterAI

Lead Solutions Architect & Founder

Dec 2025 - Present

- Architected and launched OpenWriterAI, a production-grade, subscription-based AI SaaS platform featuring a decoupled Nest.js backend architecture and a highly reactive Next.js frontend interface.
- Built a secure multi-tenant ecosystem featuring LangChain, OpenAI APIs, and a Supabase Vector Database (RAG) to power context-aware features like instant resume generators and text humanization pipelines.
- Developed a cross-platform companion browser extension, interfacing with core application APIs via secure WebSockets to deliver native, low-latency, on-page writing assistance to active subscribers.
- Integrated robust multi-tenant authentication and Lemon Squeezy webhook gateways, completely automating subscription lifecycle events, recurring billing pipelines, and webhooks processing.

U.S. AI Venture (Pegasus Project) | Remote (U.S. Contract)

Lead AI Architect & Full-Stack Developer

July 2025 - Dec 2025

- Architected and deployed an enterprise-grade, local-first multimodal AI chatbot platform utilizing Ollama and FastAPI to deliver data-private alternatives to cloud-hosted commercial solutions.
- Engineered complex, self-correcting agent execution loops using LangChain and LangGraph, integrating autonomous browser automation to programmatically research and verify live web data.
- Developed a high-throughput backend using asynchronous WebSockets to process low-latency, real-time voice streaming and live video chunk analysis, achieving a streaming response threshold under 3 seconds.
- Designed a node-based administrative workflow interface in Next.js and Tailwind CSS, enabling real-time multi-model routing, seamless model hot-swapping, and automated failover strategies.
- Optimized local hardware resource utilization by implementing model quantization and tensor parallelism strategies, reducing VRAM footprints by 40% while maintaining high token-generation throughput.

Private Equity Professional | Remote Part-Time (U.S. Contract)

PeProfessional

Full-Stack AI Engineer

Jun 2025 - Nov 2025

- Architected and deployed an enterprise-grade RAG pipeline that ingested, structured, and indexed a proprietary database of over 100,000+ private newsletter articles for contextual LLM search.
- Built an end-to-end automated data ingestion pipeline utilizing n8n, handling HTML extraction, text chunking, and vector database syncing to dynamically update the chatbot as new articles published.
- Developed a custom WordPress Single Sign-On (SSO) integration, securing user authentication and providing a seamless interface transition for active portal subscribers.
- Engineered the entire application stack using Python (FastAPI/Django) and React/Next.js, optimizing API retrieval loops to ensure low-latency, production-ready response streaming.
- Slashed GPT API production costs and token consumption by 40% through the implementation of prompt caching, semantic token pruning, and precise vector meta-filtering to minimize redundant LLM payloads.

IkamGroup | Remote (U.K. Contract)

Full-Stack AI & Automation Engineer

Jan 2025 - May 2025

- Architected and developed FacelessLab, an end-to-end autonomous video generation and publishing platform that automates the entire YouTube content creation lifecycle from a single reference link.
- Developed data collection pipelines to scrape YouTube metadata and extract video transcripts, feeding the structured data into fine-tuned GPT models to generate tailored scripts based on user criteria.
- Integrated advanced image generation pipelines utilizing Imagen 4 and Nano Banana models to programmatically generate, refine, and execute head-swapping or canvas-drawing for high-conversion video thumbnails.
- Engineered asynchronous workflows to stream scripts to the Revid.ai API for video rendering, automatically publishing the finalized media directly to target YouTube via the YouTube Data API.
- Implemented robust Redis/Celery background job queues to manage concurrent video rendering tasks, preventing API timeouts and accelerating full-cycle media compilation speeds by 35%.

Rise Marketing Engineer | Remote (U.S. Contract)

Interactive Real Estate

Full-Stack Engineer

May 2024 - Jan 2025

- Engineered and deployed a production-ready interactive real estate management platform from foundational architecture to live deployment using Next.js and NestJS.
- Built an interactive plot mapping and inventory dashboard, enabling administrators to manage property layouts and broadcast real-time availability states across embedded client views.
- Implemented a secure multi-tenant infrastructure using Supabase, developing role-based access control (RBAC), user authentication, and scalable TypeScript REST APIs.
- Managed the production infrastructure, deployment, and performance monitoring pipelines on Render, optimizing client-side rendering speeds for external third-party site embeds.
- Boosted map embed loading speeds by 45% through PostGIS spatial index optimizations and aggressive Redis caching of active plot geometry payloads, drastically cutting server overhead.

Full-Stack Engineer & Data Pipeline Architect

Jan 2024 - May 2024

- Engineered and deployed Attronaut, an enterprise-grade multi-tenant e-commerce analytics platform consolidating real-time storefront revenue and advertising performance data into a unified dashboard.
- Configured secure OAuth 2.0 authentication flows and complex webhook gateways to ingest live marketing, spend, and conversion metrics directly from Shopify and Google Ads API environments.
- Built high-throughput Django backends paired with asynchronous background workers to execute scheduled data-gathering loops, eliminating manual cross-platform reporting overhead.
- Developed a responsive, fluid data-visualization dashboard in Next.js and Tailwind CSS, achieving sub-second interaction latencies and implementing domain-verified email gateways for strict data security.

Programmers Venture | Lahore, Pakistan

Software Engineer & Full-Stack Developer

March 2021 - Jan 2024

- Deployed scalable Python (Django) and JavaScript enterprise backends, integrating custom machine learning models that boosted application data-forecasting accuracy by 25%.
- Architected and optimized real-time computer vision (CNN) data pipelines, streamlining live object detection features to process video streams with zero frame degradation.
- Built context-aware NLP modules utilizing advanced text-parsing algorithms, significantly improving automated real-time customer sentiment classification.
- Engineered secure Django REST middleware, eliminating common API vulnerabilities and safeguarding multi-tenant data access points against unauthorized execution.
- Collaborated with cross-functional product teams in an Agile environment to migrate a monolithic system to an optimized, microservices-driven architecture.

TECHNICAL SKILLS

- AI Frameworks & Orchestration: CrewAI, LangChain, LangGraph, AG-UI, Browser-use, RAG Systems, Multi-Agent Cognitive Workflows, MLFlow, ONNX Runtime, CUDA, vLLM,
- Deep Learning & Computer Vision: TensorFlow, Keras, Scikit-learn, CNNs, Custom OCR, Real-time Object Detection, LLM Architechure, Transformers, NLP, diffusers , TensorRT
- Generative AI & Multimodal: Openai APIs, Stable Diffusion, Imagen 4, Nano Banana, Ollama, Local LLM Fine-Tuning, PEFT, LoRA, QLoRA, OstrisToolKit, Timm, distillation, quantization
- Full-Stack Frameworks: Python (FastAPI, Django), JavaScript/TypeScript (Next.js, React, Node.js, Nest.js), microservices
- Databases & Cloud Infrastructure: PostgreSQL, Supabase, Vector Databases, Docker, Render, Git, CI/CD, GCP, Azure, Amazon, Vercel, Netlify, Hostinger, Cloudflare, Oracle, Huggingface Space, etc.
- Automation & Integrations: n8n, ComfyUI, Make.com, GoHighLevel

EDUCATION

Virtual University of Pakistan
Bachelor of Science in Software Engineering (BSSE) | 2016 – 2020

ADDITIONAL INFORMATION

- **Languages:** English, Urdu, Learning German
- **Portfolio:** shahzad.openwriterai.com
- **LinkedIn:** www.linkedin.com/in/shahzad-abdulmajeed-618887220
- **GitHub:** https://github.com/MShahzadAbdulmajeed
- **HuggingFace:** https://huggingface.co/its-shezi