HUSNAIN YOUNAS

Senior AI/ML Engineer

+92 302 6801164 | husnainrajpoot425@gmail.com | https://www.linkedin.com/in/husnainyounas

GIFT University (BSSE) | Virtual University (MSCS)

PROFESSIONAL SUMMARY

Senior AI/ML Engineer with 5+ years of specialized experience in architecting and deploying enterprise-grade AI systems across healthcare, automotive, smart city planning, and e-commerce sectors. Expert in Multi-Agent Systems, RAG implementations, Agentic System Design and Computer Vision. Proven track record of building scalable AI microservices, achieving 92%+ accuracy in document processing, and reducing manual workflows by 50-80%. Successfully delivered 8+ production applications and 7+ MVP/POC solutions, serving 10,000+ concurrent users with sub-200ms response times. Specialized in deep learning (NLP, Computer Vision), recommender systems, and end-to-end ML pipeline development from research to containerized deployment.

CORE TECHNICAL EXPERTISE

AI/Machine Learning & Deep Learning

- Large Language Models: GPT-4o, Claude 3.5 Sonnet, Mistral, Gemini AI, PaLM with multi-LLM orchestration and streaming capabilities
- Deep Learning - NLP: LoRA, QLoRA, Peft, RAG (Retrieval-Augmented Generation), Agentic RAG, Multi-Agent Systems, Embeddings, Text-to-Speech (TTS), Speech-to-Text (STT), Conversational AI • Deep Learning - Computer Vision: YOLO, LVM (Large Vision Models), Object Detection, Object Tracking, Multi-level Segmentation, Keypoint Detection, Feature Matching, Image Generation, OCR, Image Generation and Image Upscalers
- Machine Learning: Supervised Learning, Unsupervised Learning, Recommender Systems (Reinforcement Learning-based), ML Forecasting, Custom CNN Models
- Vector Databases & Semantic Search: Pinecone, Qdrant, FAISS with sub-200ms query response times

AI Frameworks & Development Tools

- AI Orchestration: LangChain, LangGraph, CrewAI, vllm, Autogen, OpenAI AgentKit, n8n workflow automation
- AI APIs & Integration: OpenAI API, Azure OpenAI, Anthropic Claude API, Tavily, Perplexity, Tool Calling, Streaming AI, WebSocket integration
- Voice AI: Elevenlabs TTS, LiveKit, real-time speech recognition and voice-enabled conversational interfaces
- Development & Monitoring: LangSmith for debugging and observability, A/B testing capabilities

Backend Development & Cloud Infrastructure

- Backend: Python, FastAPI microservices, WebSockets, Socket.IO, real-time streaming APIs • Databases: PostgreSQL, MongoDB, SQL optimization and query design
- Cloud & DevOps: AWS (EC2, S3, Lambda, SES), Microsoft Azure(Azure ML, Cognitive Services etc), Docker containerization, CI/CD pipelines, nginx
- Authentication & Security: JWT authentication, role-based access control, secure data management

Data Processing & Automation

- Web Scraping & Automation: Selenium, Playwright, Zenrows for large-scale data extraction from dynamic websites.
- Document Processing

PROFESSIONAL EXPERIENCE

Senior AI Engineer

Leading AI innovation initiatives and cross-functional engineering teams to deliver enterprise-scale intelligent automation solutions across multiple industries

- Architected and deployed 8+ enterprise-grade AI applications plus 7+ MVP/POC solutions across healthcare, automotive, smart city planning, and education sectors using GPT-4o, Claude, Mistral, Gemini, and YOLOv8
- Built advanced multi-agent systems, RAG implementations, and recommender systems with LangChain, CrewAI, reinforcement learning, processing 1M+ documents with sub-200ms response times and achieving 92% accuracy in document processing and AI categorization
- Implemented real-time AI streaming infrastructure using WebSockets, FastAPI microservices, and computer vision solutions (YOLOv8, custom CNN, keypoint detection), serving 10,000+ concurrent users and reducing manual inspection time by 75%
- Developed intelligent automation workflows including job scraping, ML-powered inventory forecasting, 3D mapping for smart city planning, and ERP integrations, reducing manual work by 50-80% through conversational AI platforms
- Optimized ML pipelines and containerized deployments using Docker and AWS/Azure cloud architecture, implementing automated CI/CD workflows with comprehensive monitoring, A/B testing capabilities, and 3D modeling solutions

KEY AI PROJECTS

Dronodat – 3D Smart City Planning & Surveillance System
Dronodat : May-2025 — Continue
Organization : Dronodat / Remote

Technologies: Computer Vision, YOLO, Keypoint Detection, Feature Matching, 3D Mapping, Object Detection & Tracking, EDA, Data Pre-Processing, Data Preparation and Validation Pipeline

Advanced 3D modeling platform for smart city planning and facility visualization empowering data-driven insights, streamlined collaboration, and efficient decision-making through computer vision-based surveillance systems.

Core Expertise & Contributions:

- Computer Vision Engineering: Implemented complete surveillance system pipeline with custom-trained object detectors and keypoint detection algorithms for smart city applications
- 3D Mapping & Reconstruction: Developed 3D mapping solution using keypoint detection and feature matching techniques for accurate facility visualization
- Object Detection & Tracking: Built multi-object detection and tracking system using YOLO for real-time city surveillance and monitoring
- Data Pipeline Engineering: Designed comprehensive EDA and data pre-processing pipelines for training custom CNN models on urban infrastructure data

CHEX.AI – Intelligent Vehicle Inspection Platform
Chex.ai : Mar-2023 — Continue
Organization : ChexAI LLC

Technologies: Florence2, YOLO, Detectron2, , Multi-level Segmentation, OCR , GPT-4o Vision, Object Detection & Tracking, Feature Matching, Data Augmentation

AI-powered automated vehicle inspection system revolutionizing damage detection and assessment processes using advanced computer vision and OCR technologies.

Core Expertise & Contributions:

- Multi-level Segmentation: Engineered sophisticated multi-object segmentation pipeline for precise identification and localization of vehicle damage across different severity levels
- Automated Data Pipeline: Built end-to-end automated data preparation pipeline including data

augmentation, model training, and validation workflows, reducing manual preprocessing time by 75% • OCR Integration: Implemented advanced Optical Character Recognition systems for license plate & vin number extraction and vehicle identification with 95%+ accuracy.
• Old Damage details and damage location, feature matching for old and new damages. • Historical Damage Analysis: Developed intelligent pipeline for retrieving and analyzing old damage details, enabling comprehensive vehicle history tracking
• Detailed Damage Report.

Site Assist – Automated Custom RAG Agent Platform
Site Assist Org : Jan-10-2026 — continue
Organization : Techling Pvt ltd

Technologies: RAG, Agentic RAG, LangChain, Qdrant , Web Scraping, Multi-Agent Systems, Lead Generation, User chat sentiment.

Intelligent automated site agent platform that creates custom RAG systems instantly from website URLs, YouTube videos, and project documents, extending to comprehensive customer service and sales lead generation.

Core Expertise & Contributions:

- Agentic RAG Architecture: Designed and implemented automated RAG system that crawls, scrapes, and processes website content and multimedia sources to create custom knowledge bases instantly • Multi-Source Integration: Built intelligent data ingestion pipeline supporting website URLs, YouTube video transcripts, and document uploads for comprehensive knowledge extraction
- Conversational AI Agent: Developed advanced chatbot with RAG capabilities serving as both customer service and sales agent, understanding user intent and generating qualified leads
- Lead Intelligence System: Engineered automated reporting system that analyzes user conversations to identify intent, interest levels, and contact preferences for sales teams
- Future Integration Ready: Architected extensible platform with planned integrations for email, calendar, WhatsApp, and CRM tools for seamless business workflow automation

Teach Track AI – Intelligent Learning Management System
Freelance : Feb-21 — Aug-21

Technologies: FastAPI, MongoDB, Pusher, Pinecone, OpenAI GPT-4, Claude 3, PaddleOCR, LangChain, Elevenlabs, Semantic Search

Advanced AI-powered Learning Management System with multimodal intelligence, real-time collaboration, and adaptive learning featuring secure authentication, voice/text chatbot, OCR-based content extraction, automated question generation, AI-driven grading, and intelligent presentation generator.

Key Features:

- Multimodal AI Integration: Conversational AI with voice and text capabilities, OCR-based content extraction, automated question generation, and AI-driven grading systems
- Intelligent Presentation Generator: Automated educational presentation creation with semantic search integration and intelligent image sourcing for academic institutions
- Personalized Learning Paths: AI-driven adaptive learning with role management, real-time collaboration, and customized study paths

Milele – AI-Powered Automotive Export Platform
May-2021 — jan-2024
Organization : Milele UAE

Technologies: Recommender Systems, Reinforcement Learning, Multi-Agent Systems, Sales AI Agent, Marketing Content AI Agent, Customer Service AI

World leader in tax-free motor vehicle export to Africa, Asia, and Europe with specialized AI systems for

vehicle procurement, export management, and customer service technology.

Core Expertise & Contributions:

- Reinforcement Learning Recommender: Implemented advanced recommendation system using reinforcement learning for personalized vehicle suggestions and procurement optimization • Sales & Customer Service Agent: Developed intelligent AI agent combining sales expertise with technical support capabilities for enhanced customer experience
- Marketing Content AI Agent: Created automated social media script writing and promotional content generation agent for multi-platform marketing campaigns

DeftGPT – Multi-LLM Chat Platform
Organization : Remote / Nektek

Technologies: FastAPI, Multi-LLM, RAG, Agentic RAG, ChromaDB, Tavily, Perplexity, Docker, Microservices, GenAI, Text-to-Image

Production-grade microservice for multi-LLM chat platform with advanced streaming capabilities, RAG integration, document chat, image generation, private chats, and containerized deployment.

Key Features:

- Multi-LLM Architecture: Integrated multiple LLM providers with intelligent routing and streaming capabilities for optimal performance
- RAG Pipeline Implementation: Built comprehensive RAG system for document chat functionality with ChromaDB vector storage and semantic search
- Production Microservices: Developed FastAPI-based microservices with Docker containerization for scalable deployment and maintenance

Additional Enterprise Projects

PatientEd AI Avatar System: AI-powered medical consultation platform with 3D streaming avatars (HeyGen), real-time speech recognition, multilingual ophthalmology education, and secure patient data management using Next.js, TypeScript, FastAPI, GPT-4o, PostgreSQL, and WebSockets.

Job Scraping & AI Categorization: Intelligent job scraper with GPT-4o-powered classification achieving 92% accuracy across 10,000+ listings and 35% improvement in recruitment matching using Selenium, Playwright, PostgreSQL, and Zenrows.

CAPA Automotive Dealership System: Automotive dealership management software with integrated ERP automation (Odoo, QuickBooks) for inventory and accounting processes, reducing manual work by 50% using FastAPI and PostgreSQL.

Fouani ML Forecasting: ML-powered inventory forecasting system with LLM natural-language query interface for multi-warehouse procurement insights and demand prediction.

OmniPlex AI Platform: Multi-LLM AI chat and search platform with plugin tool orchestration, persistent chat history, and serverless API architecture using Next.js, TypeScript, Gemini AI, OpenAI, Claude, LangChain, and LangGraph.

Talk2Taste: Webhook-based real-time ordering system with bidirectional data streaming, live cart synchronization, and AI voice integration using Node.js, Socket.IO, WebSockets, and Elevenlabs. **AI Mental Health Assistant:** Conversational AI therapeutic assistant with structured assessment workflows, vector-based knowledge retrieval using FastAPI, OpenAI GPT-4, LangChain, FAISS, and RAG. **Multi-Agent Newsletter Automation:** Intelligent 3-agent workflow automating newsletter research, content generation, and HTML formatting with structured output validation using n8n, Gemini AI, and Gmail Integration.

PROFESSIONAL CERTIFICATIONS

Deeplearning.ai Specializations:

- Machine Learning Specialization by Deeplearning.ai and Stanford University
- Deep Learning Specialization by Deeplearning.ai
- MLOps: Machine Learning Operations Specialization by Duke University

Advanced AI & LLM Courses:

- Embedding Models: Architecture to Implementation (Deeplearning.ai)
- MCP: Build Rich-Context AI Apps with Anthropic (Deeplearning.ai)
- Large Language Models with Semantic Search (Deeplearning.ai)
- Building AI Voice Agents for Production (Deeplearning.ai)
- Prompt Engineering for Developers (Deeplearning.ai)
- LangChain Developer Essentials (Deeplearning.ai)
- AI Agents in LangGraph (Deeplearning.ai)
- Reasoning with o1 (Deeplearning.ai)
- AI for Everyone (Coursera)

KEY ACHIEVEMENTS

- Architected 8+ enterprise-grade AI applications and 7+ MVP/POC solutions using GPT-4o, Claude, Mistral, Gemini, and YOLOv8 across healthcare, automotive, smart city planning, and education sectors
- Achieved 92% accuracy in AI-powered classification systems, processing 1M+ documents with sub-200ms response times and serving 10,000+ concurrent users
- Engineered multi-agent automation workflows, computer vision solutions (3D mapping, object detection, segmentation), and recommender systems, reducing manual processes by 50-80% and inspection time by 75%
- Led implementation of reinforcement learning-based recommender systems and intelligent document processing with ERP integrations for enterprise workflow automation

EDUCATION

Master of Science in Computer Science (MSCS) Virtual University of Pakistan | In Progress

Bachelor of Science in Software Engineering (BSSE)GIFT University | Graduated 2021