Huzefa Nalkheda Wala

India / United Arab Emirates (Hybrid-Remote)

☎ +91 9685595352 ✉ huzaifanalkhedaemp@gmail.com

LinkedIn Portfolio GitHub

Summary

AI Product Engineer blending deep LLM & RAG expertise with full-stack product delivery. Engineered production-grade AI pipelines for document intelligence, multimodal vision, email intelligence, and agentic automation. Creator of **MedGenius LLaMA-3.2B** and curator of a **40K+ Medical Intelligence Dataset**. Proven track record of end-to-end ownership from dataset → model → API → deployment. **Impact: ≈1,000 GitHub contributions (2025)**, **40+ merged PRs**, **45+ production features shipped**.

Technical Expertise

- **LLMs & Retrieval:** LLaMA, LoRA/QLoRA, RAG, BGE embeddings, SentenceTransformers, Faiss, Qdrant
- **Multimodal & CV:** OCR pipelines, floor-plan analysis, image preprocessing, OpenCV
- **Backend & API:** FastAPI, Django, Node.js, TypeScript, Redis, PostgreSQL, SSE/WebSockets, JWT/OAuth2
- **Production & Infra:** Docker, CI/CD, Railway, S3, monitoring & health checks, memory optimization
- **Integration & Automation:** n8n nodes, WhatsApp automation, webhook systems, OpenRouter
- **Research & Communication:** Technical publications, open-source models/datasets, developer outreach

Experience

AI Product Engineer Nov 2024 – Present

CleverFlow (UAE-based automation platform) *Hybrid (India / UAE)*

Building AI features that transform a no-code workflow platform into an intelligent automation stack.

- **Engineering Velocity:** ≈1,000 GitHub contributions (2025); 40+ merged PRs; 45+ production features shipped across multiple repositories.
- **Document Intelligence API:** Engineered a FastAPI + CV + LLM pipeline for parsing UAE government documents (IDs, title deeds, licenses) with production-grade accuracy and monitoring.
- **Email Intelligence System:** Architected production RAG pipeline—MSG parsing, HTML sanitization, SentenceTransformer embeddings, Qdrant multi-vector indexing, hybrid semantic search (weighted vector fusion), score aggregation, deduplication, context-aware LLM generation, conversation threading, PostgreSQL chat management, Redis caching, JWT authentication, memory optimization.
- **AI-Powered CRM Systems:** Architected WhatsApp lead-status agent, AI follow-up agent, transcription & summarization pipeline, SSE workflows, and CRM automation logic (frontend + backend).
- **Marketing Website & CMS:** Developed full Blog/CMS stack with SEO architecture, AI content tooling, and production optimizations for improved search visibility.
- **RAG Chatbot:** Architected PDF → embeddings → vector search → Redis caching pipeline with streaming responses for enterprise knowledge search.
- **Multi-Model OCR System:** Designed vision model routing, image compression, structured metadata parsing, and document extraction for various document types.
- **Real-Estate CV Tool:** Floor plan comparison system for architectural change detection and compliance reporting in property workflows.
- **Platform Integrations:** Engineered official n8n node, Dockerized microservices, SSE/WebSocket event streams, S3 audio pipelines, and deployment infrastructure.

Creator & Lead Engineer Jan 2024 – Oct 2024
MedGenius Project *Remote*

Architected an open healthcare AI stack and dataset:

- **MedGenius LLaMA-3.2B:** LoRA/QLoRA fine-tuned model for diagnostics, education, and telemedicine.
- **Medical Intelligence Dataset (40,443 records):** Open-sourced dataset (diseases, symptoms, treatments, dialogues) adopted by the community.
- **PHMS Medical Device (Registered Design):** Co-inventor on a registered low-cost wearable design for vital monitoring (Design No. 375474-001, Govt. of India, 2022).

Software Developer Mar 2024 – Jun 2024
RailWorld India Pvt. Ltd. *Indore, India*

Developed internal automation tools and system dashboards using modern full-stack practices; led team project demonstrating collaborative development capabilities.

Growth Strategist Intern Jan 2024 – Feb 2024
Growth Theory *Mumbai, India*

Drove digital strategy initiatives, growth experiments, and campaign execution.

Education

Indian Institute of Technology, Ropar Aug 2024 – Nov 2025 (Expected)
Major Program in Artificial Intelligence
Key coursework: LLMs, RAG systems, TinyML, Reinforcement Learning, Computer Vision, NLP, Prompt Engineering.

Parul University 2019 – Apr 2023
Bachelor of Technology (B.Tech), Information Technology Grade: 7.20/10
1st Place, Vadodara Startup Fest with PHMS medical device project (among 250+ startups, 25K+ participants).

Key Projects

- **Email Intelligence System:** Architected production RAG pipeline—MSG parsing, HTML sanitization, SentenceTransformer embeddings, Qdrant multi-vector indexing, hybrid semantic search (weighted vector fusion), score aggregation, deduplication, context-aware LLM generation, conversation threading, PostgreSQL chat management, Redis caching, JWT authentication, memory optimization.
- **Document Intelligence API:** Production API for parsing UAE government documents with high accuracy and real-time monitoring.
- **AI-Powered CRM Stack:** Lead engineering for AI agents in CRM workflows: WhatsApp interpreter, follow-up agent, transcription + summarization, real-time eventing.
- **Marketing Website & CMS:** Full CMS and SEO stack with AI content generation tools (JSON-LD, sitemap, search visibility optimizations).
- **Enterprise RAG Chatbot:** PDF ingestion → embedding → vector retrieval → streaming chat with caching layer.
- **Multi-Model OCR System:** Vision model routing and structured extraction for title deeds, invoices, and identity documents.
- **Floor Plan Inspector:** CV-based system to compare architectural plans and detect modifications for compliance workflows.

Research & Open Source

- **Medical Intelligence Dataset** (40,443 records) — Kaggle & Hugging Face
- **MedGenius LLaMA-3.2B** — Open fine-tuned model (LoRA/QLoRA)
- **Publications:** Byte Latent Transformer (BLT) analysis, DeepSeek v2.5 evaluation, Marco O1 reasoning research, XAI & LLM engineering guides
- **Community:** 45+ GitHub repositories, 5 models on Hugging Face, 3K+ LinkedIn followers

Skills

- **AI/ML:** PyTorch, Transformers, LLaMA family, LoRA/QLoRA, BGE, Faiss, SentenceTransformers, Qdrant, OpenCV
- **Backend / API:** Python, FastAPI, Django, Node.js, TypeScript, PostgreSQL, Redis, JWT/OAuth2
- **Infra:** Docker, Railway, S3, CI/CD, WebSockets/SSE
- **Tools / Platforms:** GitHub, Hugging Face, Kaggle, n8n, OpenAI, OpenRouter, Google AI