

# HAMZA AZIZ

AI Product Engineer • Generative AI & LLM Systems • Computer Vision • Conversational AI

Ghaziabad, Uttar Pradesh, India | +91 8447644188 | hamza.aziz.dev@gmail.com

Nationality: Indian | Visa Status: UAE Employment Visa Sponsorship Required | Notice Period: 30 Days

Open to Relocation: Abu Dhabi & Dubai, UAE

[LinkedIn](#) | [GitHub](#) | [Portfolio](#) | [Medium](#)

## PROFESSIONAL SUMMARY

AI Product Engineer with 3+ years of experience building production Generative AI, Computer Vision, and Conversational AI systems for enterprise and multi-tenant platforms. Delivered agentic LLM applications using LangGraph, Retrieval-Augmented Generation (RAG), and Google Gemini, and built real-time facial recognition and video analytics systems on NVIDIA DGX A100 and Jetson Orin Nano achieving >99.99% accuracy. Experienced across the full production stack — FastAPI services, vector databases (Milvus), TensorRT/DeepStream inference optimization, and CI/CD-driven MLOps — for low-latency, large-scale deployments. Seeking Generative AI, Computer Vision, or Applied AI Engineering roles in the UAE (Abu Dhabi / Dubai).

## SKILLS

- **Programming Languages:** Python, SQL, Bash
- **AI / Machine Learning:** PyTorch, TensorFlow, OpenCV, YOLO, InsightFace, Scikit-learn, NumPy, Pandas
- **Generative & Conversational AI:** LangChain, LangGraph, Agentic AI (ReAct), Rasa, Google Gemini, LLMs, NLP, Prompt Engineering, RAG, vLLM
- **Inference & Edge AI:** TensorRT, ONNX, DeepStream, NVIDIA Jetson, DGX A100
- **Backend & MLOps:** FastAPI, Next.js, React, TypeScript, Docker, CI/CD, Git, Linux
- **Databases & Messaging:** Milvus (Vector DB), PostgreSQL, MongoDB, Redis, Kafka, MinIO
- **Languages:** English (Professional Working Proficiency), Hindi (Native)

## WORK EXPERIENCE

**AI Product Engineer** | *VoicePlug Inc.* | September 2025 – Present

- Ship production-grade DriveThru voice ordering features for a multi-tenant Quick-Service Restaurant (QSR) platform deployed across 2 live restaurants and 4+ QSR brands
- Migrated core conversational workflows from legacy Rasa NLU to an agentic LangGraph + Google Gemini architecture, improving entity extraction, dialogue control, and tool-use orchestration
- Raised live client success metrics from ~25–30% to 60%+ by resolving intent-recognition gaps, entity ambiguity, menu-understanding failures, and fallback-handling issues
- Hardened POS-integrated deployments by improving exception handling, debug observability, fallback escalation, and restaurant-specific prompt templating, reducing recurring production issues by 6%
- Converted insights from 500+ real customer interactions into product improvements by partnering with product, support, and QA teams to reduce ordering friction and improve deployment readiness

**Assistant Engineer – AI** | *Global InfoVentures Pvt. Ltd.* | August 2023 – August 2025

- Led development of 4 commercial-grade Facial Recognition Attendance Systems on DGX A100, RTX 4090, and Jetson Orin Nano, achieving >99.99% accuracy
- Developed the G6 Voice Assistant, converting natural language voice queries into precise SQL queries via Rasa, Gemini 2.0, and Milvus Vector DB; handled ambiguous queries with advanced intent disambiguation, achieving near-100% intent classification accuracy
- Mitigated domain shift in facial recognition using a CI/CD-style pipeline combining mobile and CCTV datasets — improved cross-domain recognition by 14%
- Integrated PostgreSQL for recognition logs and engineered an API sync module with the legacy SIM ERP Portal, enabling real-time attendance updates and SMS alerts for 2,000+ students

**AI Intern** | *Global InfoVentures Pvt. Ltd.* | February 2023 – July 2023

- Conducted 10+ hands-on AI/ML training sessions and mentored 50+ students across CSE, AIML, and Data Science domains
- Guided 10+ object detection projects using YOLOv8 and CNNs (VGG, ResNet, InceptionNet), resulting in successful full-cycle model deployments

## EDUCATION

---

### **Master of Technology (Executive), Artificial Intelligence and Machine Learning**

Birla Institute of Technology and Science (BITS), Pilani — WILP | 2025 – 2027

### **Bachelor of Technology, Computer Science and Engineering**

Dr. A. P. J. Abdul Kalam Technical University, ABES Institute of Technology | 2019 – 2023

## PROJECTS

---

### **DriveThru Voice Ordering Modernization** | *VoicePlug Inc.*

- Shipped production features for a real-time, multi-tenant DriveThru voice ordering agent powered by FastAPI, Socket.IO, LangGraph ReAct workflows, Google Gemini 2.5 Flash fallback logic, and per-call token-cost accounting
- Refactored entity parsing, size normalization, and combo/multi-size guard logic to eliminate cart crashes and enforce semantically valid orders before persistence
- Hardened entity extraction against ASR disfluencies using fuzzy matching and added latency-masking filler scheduling that suppresses duplicate TTS when the real model response arrives in time
- Extended POS-compatible ordering workflows across 4 integrations (QUBeyond, Clover, Speedline, Deliverect) using Jinja2-templated restaurant prompts, Redis-backed session state, and MongoDB-backed order-event audit trails

### **G6 Voice Assistant — Natural Language to SQL Converter** | *Global InfoVentures Pvt. Ltd.*

- Integrated SpeechRecognition, Rasa, and Gemini 2.0 to convert user voice input into structured SQL queries, enabling seamless access to ERP data across Admission, Employee, Fee, and Library modules
- Achieved ~100% intent classification accuracy using Rasa by training with 50+ voice query variations per module; devised secondary intent resolution via contextual follow-ups for ambiguous queries
- Formulated 100+ field-level schema mappings per module for Gemini and integrated Milvus Vector DB-based suggestions, reducing LLM hallucinations and boosting SQL query precision by over 30%
- Solved heterogeneity in field values using vector-based field and value disambiguation, improving precision in SQL generation by over 40%

### **Computer Vision Attendance Systems** | *Global InfoVentures Pvt. Ltd.*

- Deployed CCTV, PTZ, and edge-based facial recognition attendance systems using SCRFD, w600k\_r50, TensorRT, DeepStream, Milvus, DGX A100, RTX 4090, and Jetson Orin Nano
- Orchestrated 20+ live CCTV streams and automated face-vector refreshes from elevated-angle surveillance footage, improving real-world recognition accuracy from 85% to 99%
- Enabled low-latency, contactless attendance tracking across classroom, office, and surveillance environments, reducing manual roll calls, proxy attendance, and biometric contact dependency

## CERTIFICATES

---

- Rasa Developer Certification — 2025
- Introduction to Data Science, IBM — 2025
- LangChain — Develop AI Agents with LangChain & LangGraph, Udemy — 2025
- Generative AI with Large Language Models, Coursera — 2024
- Building Video AI Applications at the Edge on Jetson Nano, NVIDIA Deep Learning Institute — 2024
- Getting Started with AI on Jetson Nano, NVIDIA Deep Learning Institute — 2024
- Artificial Intelligence and Machine Learning, Global InfoVentures Pvt. Ltd. — 2023
- Getting Started with Deep Learning, NVIDIA Deep Learning Institute — 2022