# Muhammad Ibrahim Shaikh

Islamabad, Pakistan — +92 333 0336690 — m.ibrahimshaikh357@gmail.com
github.com/mibrahimm1 — linkedin.com/in/ibrahimshaikhh

## SUMMARY

Software Engineering student at NUST with production experience in designing agentic AI systems and automation infrastructure. Proven ability to build end-to-end pipelines in NLP, computer vision, multimodal AI, and full-stack integration. Skilled in Python, PyTorch, LangChain, LLM, RAG pipelines, Agentic workflows, n8n, and production LLM operations. Winner of the AI hackathon at NUST. Seeking roles in AI and Product Engineering to work on multimodal agentic architectures, automated software tools, intelligent systems, and applied AI.

## EDUCATION

**National University of Sciences and Technology (NUST)** — **7th Semester, Ongoing**
Bachelor of Software Engineering — CGPA: 3.59 / 4.0 — Islamabad, Pakistan

## EXPERIENCE

**AI & Product Engineer** — Horquva LLC, United States (Remote) — **Oct 2026 - Present**
- Architected core dependency graph and simulation engine in TypeScript/PostgreSQL to detect enterprise single points of failure; build topological algorithms modeling blast radii when employees depart or third-party APIs fail.
- Shipped 7 features end-to-end, including an interactive dependency map, a "What-If' departure simulator, and a mobile-friendly 60-second magic-link review portal for managers.
- Designed multi-tenant data layer with PostgreSQL Row-Level Security across 58 tables, build a 23-endpoint gateway, and integrate a grounded Gemini copilot; maintain 100% test pass rate across 53 suites.

**AI Automation Engineer & Product Growth** — BizBuddy, Karachi — **Jun 2026 - Aug 2026**
- Architected an enterprise multi-tenant conversational agent powering WhatsApp ordering across 25 restaurant clients, scaling throughput to 157,000+ messages, 28,800+ catalog items, and 16,300+ delivery areas while cutting onboarding from 40 hours to 8 hours.
- Engineered a multimodal ingestion pipeline on WhatsApp Cloud API integrating OpenAI Whisper (speech-to-text with Roman Urdu detection) and Gemini 2.5 Flash Vision for real-time customer audio note and receipt/image processing.
- Developed a tool-calling engine with 9 custom tools interfacing PostgreSQL for real-time catalog browsing, order modification, dynamic promo validation, automated complaint resolution.
- Engineered a two-phase transactional commit pipeline with deterministic Pydantic schema validation, a tiered fuzzy delivery-area resolver, and automated SQL memory correction on failed orders.
- Resolved multi-turn looping and framework execution errors by migrating from Gemini 2.5 Flash to Qwen 3.6 Flash, designing output sanitizers to suppress thinking-bleed while reducing input tokens by 80% and operational costs by 89%.
- Built zero-migration webhook and API integrations from scratch with enterprise POS platforms Blink POS and Technosys POS, while establishing an internal 65-test automated QA evaluation suite achieving a ∼98% order accuracy baseline.

**AI Engineer Intern** — Entropik Labs, Islamabad (Remote) — **Jul 2025 - Sep 2025**
- Architected and automated an end-to-end BiCodec TTS data pipeline; covering tokenization, detokenization, and audio synthesis to produce a large-scale synthetic voice dataset; integrated UTMOS and WER metrics, enabling scalable, synthetic dataset curation.
- Built a multimodal data-curation pipeline using Docling for OCR ground-truth extraction and Llama 4 for structured annotation labeling; fine-tuned three lightweight Vision-Language Models with HuggingFace, establishing a clear correlation between dataset quality and downstream model accuracy.
- Integrated 5+ custom tools (real-time events, live weather, trending news, location geocoding, POI search) into a production LangChain LLM Agent, expanding its ability to resolve location-specific queries.

## PROJECTS

**Autonomous Gmail Monitoring Agent** – *Python, LangGraph, Gmail API, Google Sheets API, Pydantic, Tenacity* — [Github]
- Architected an event-driven agentic workflow using LangGraph (StateGraph) to autonomously monitor inboxes, analyze conversational context, and record tracking entries in Google Sheets.
- Implemented a 6-node deterministic graph featuring structured LLM intent classification, thread analysis for unanswered queries, and deduplication verification via Google Sheets API v4.
- Integrated Tenacity exponential backoff retries and Pydantic validation schemas to guarantee fault-tolerant API execution and eliminate malformed data writes.

**Multimodal Medical Report Analyzer** – *YOLOv8, FastAPI, LangChain, Google Gemini, Next.js, Docker* — [Github] — [Live]
- Fine-tuned YOLOv8 on 14,000+ Chest X-Rays to detect 14 thoracic abnormalities; achieved mAP@50 of 0.99 for healthy scans, 0.88 for Cardiomegaly, and 0.87 for Aortic Enlargement across 25 epochs on a T4 GPU.
- Architected a multimodal FastAPI backend integrating the fine-tuned YOLO vision engine, clinical text extraction, and a LangChain orchestration chain that cross-references detections with parsed report findings to generate patient-friendly summaries.
- Deployed a full-stack application featuring a real-time split-screen dashboard with an AI Grounding toggle for interactive bounding box overlays on annotated X-Rays.

**Sonyc - Production Multimodal RAG System** – *Python, FastAPI, LangChain, FAISS, ChromaDB, Hugging Face* — [Github]
- Designed and deployed a complete RAG system ingesting PDFs, GitHub repositories, YouTube transcripts, and live web content into a unified vector knowledge base with semantic chunking, BGE/Gemini embeddings, and dual FAISS/ChromaDB indexing
- Built LangChain pipelines with Runnables, Prompt Templates, Output Parsers, and multi-step Chains for context-aware LLM response generation
- Deployed FastAPI backend on AWS EC2; embedding caching and async processing reduced average response latency by 40%

**Edge IoT Voice Assistant** – *Raspberry Pi, PyAudio, WebRTC VAD, FastAPI, Groq Whisper, Llama 3* — [Github]

- Developed a real-time hardware voice assistant interfacing a Raspberry Pi edge client with a distributed FastAPI inference backend over local HTTP streams.
- Implemented client-side voice activity detection using PyAudio and WebRTC VAD, streaming digitized audio only during active speech to eliminate idle bandwidth.
- Chained Whisper for sub-second STT, Llama-3.3 for conversational reasoning, and an asynchronous TTS pipeline streaming speech responses directly back to Raspberry Pi hardware speakers.

**AI-Powered Video Calling Application** – *WebRTC, Socket.IO, Google Gemini, Google Cloud Translation API, Vercel* — [Github]
- Built a real-time peer-to-peer video calling platform supporting low-latency streaming and WebRTC connection handshakes.
- Integrated Gemini 2.5 Flash for real-time speech-to-text transcription with automatic language detection with live translation

## ACHIEVEMENTS

- **Winner, Code Connect 2025 AI Hackathon** – 1st among 20+ teams across Pakistan, delivering a production-ready AI solution under competition pressure
- **NTF Scholarship Recipient** – Awarded 100% Fully Funded NUST Need-cum-Merit Scholarship (Tuition covered).
- **GPA Merit Scholarship Recipient** – Recognized as one of the top GPA scorers, awarded a merit-based scholarship.

## TECHNICAL SKILLS

- **Languages:** Python, Java, C++, TypeScript, JavaScript, HTML, CSS, SQL
- **AI/ML:** PyTorch, TensorFlow/Keras, Scikit-learn, HuggingFace Transformers, Fine-tuning, Computer Vision
- **LLM and Agentic AI:** LangChain, LangGraph, LlamaIndex, Pydantic, OpenAI API, Qwen, Hugging Face, Whisper, Transformers, RAG, Agentic Tool Calling, Model Context Protocol (MCP), Prompt Engineering, Semantic Chunking, Embeddings, Vector Databases
- **Automation:** n8n, API Integration, WhatsApp Business API, Webhook Orchestration
- **Data Visualization:** Pandas, NumPy, Matplotlib, Seaborn, Altair, Streamlit
- **Web / Backend:** React.js, Next.js, FastAPI, Bootstrap, REST APIs, WebRTC, Socket.IO
- **Tools:** AWS (EC2, S3), PostgreSQL, SQLite, Git/GitHub, CI/CD (GitHub Actions), Jupyter, Docling, Whisper, Wav2Vec2

## VOLUNTEERING & LEADERSHIP

**Secretary Media & Executive Panelist** — Software Society, NUST **Jul 2024 - Jun 2025**
- Directed a creative committee managing brand identity, digital outreach strategy, and multimedia event coverage across department-wide software engineering initiatives.
- Redesigned the official visual identity for the Software Society; led typesetting and layout design for the Computer & Software Engineering Department's annual magazine, *Resurgence*.
- Served on the Executive Director's Panel for DevCon '25, spearheading branding and digital media campaigns for the national hackathon.

**Eco-Intern** — WWF-Pakistan **Jul 2022**
- Completed WWF-Pakistan's Eco-Internship Program, contributing to environmental sustainability campaigns, ecological conservation workshops, and community outreach.