Muhammad Haseeb

AI Engineer | Islamabad, Pakistan
+92 314 9946492 | mhaseeb1604@gmail.com | linkedin.com/in/haseebai | haseebai.tech

SUMMARY

AI Engineer specializing in production Agentic RAG systems and LLM orchestration. Shipped AI assistants that drove $600K in pre-seed funding and cut enterprise research time by 95% on 300GB+ corpora. Built FastAPI backends scaling to 1K+ Concurrent Requests, multi-tenant SaaS architectures with RBAC, and Dockerized CI/CD pipelines on AWS. Strong fundamentals in async Python, PostgreSQL, vector search, and end-to-end AI pipelines.

SKILLS

AI / ML: Agentic RAG, LLM Orchestration, LangChain, LlamaIndex, PydanticAI, LiveKit Agents, OpenAI, Anthropic Claude, embeddings, vector search, hybrid retrieval, prompt engineering, deep learning (PyTorch / Tensorflow), NLP
Backend: Python, FastAPI, Flask, async I/O, REST APIs, SQLAlchemy, pytest
Data: PostgreSQL, MySQL, SQLite, Milvus, Qdrant, ElasticSearch
Infra & DevOps: Docker, CI/CD, AWS (EC2, DynamoDB), GCP, Azure, Git, Linux
Observability & Tools: Langfuse, OpenTelemetry, Arize Phoenix, Jira

EXPERIENCE

AI Engineer | *Stixor Technologies* Apr 2024 – Jun 2026

*Across the projects below, also owned the backend and infrastructure layer end-to-end: FastAPI services scaling to **1K+ Concurrent Request** with async processing and caching, SQLite→PostgreSQL migration with query optimization and connection pooling, and Dockerized CI/CD pipelines for automated, reliable deployments.*

- **Mari Energies – Intelligent Search & Chatbot System**
  Cut engineer research time by 95% across 300+ GB of unstructured technical data. Designed and deployed Agentic RAG assistants (chatbot + search) processing PDFs, Word, Excel, PowerPoint, and SAP exports. Built custom parsers (RAGFlow Deepdoc + in-house) and optimized RAG pipelines for **1K–6K page** technical documents, enabling engineers to retrieve answers instantly with source verification and document downloads. Drove system adoption across multiple engineering teams and improved compliance workflows.
- **Malakah – AI Legal Assistant**
  Drove $600K in pre-seed funding with a 98% hit rate on Saudi legal corpora. Built an Agentic RAG assistant combining law retrieval and query-based legal support. Partnered with KSA legal experts to refine contextual recommendations, ensuring jurisdiction-specific, reliable outputs. Became a core product feature, validating the platform's vision of AI-driven legal services.
- **Multi-Tenant AI SaaS Platform (MVP)**
  Architected and prototyped an Agentic RAG system with multi-tenancy and **role-based access control enforced at the vector-search level**, designed to enable businesses to deploy AI-powered document intelligence without custom development. Collaborated with backend engineers to design PostgreSQL schema supporting tenant isolation, permissions management, and document processing. Implemented hybrid retrieval system with PydanticAI framework, integrating the AI layer with backend services for secure, performant querying.
- **Bilingual Voice Agent on LiveKit – English / Urdu**
  End-to-end real-time voice agent with RAG-augmented context injection. Built on LiveKit integrating Kokoro TTS, remote Whisper STT, and Silero VAD, with Langfuse and OpenTelemetry tracing across the full agent loop. Diagnosed and removed CPU-bound TTS buffering to materially reduce end-to-end response latency; refactored custom observability onto LiveKit's native *metrics_collected* event, removing manual instrumentation across the core agent loop. Deployed as a reception kiosk at an industry exhibition.

FastAPI Backend Engineer | *TriTech Solutions (Contract)* Dec 2023 – Feb 2024

- Developed a FastAPI RESTful API for an e-commerce platform, integrating AWS DynamoDB with ORM to enable scalable catalog management and user session handling as the platform grew.
- Deployed services on AWS EC2 with production-grade configuration and implemented Pytest-driven unit testing that reduced bugs, improved reliability, and ensured faster, safer feature releases for end users.

Remote AI Developer | *Freelance* Mar 2023 – Apr 2024

- Developed a ResNet-based real-time model for classifying Pakistani currency notes, achieving 99% accuracy in image-based recognition tasks.
- Architected a hybrid deep learning model for precipitation nowcasting using the CIKM radar dataset, forecasting local weather patterns with reduced error (RMSE = 11 dBZ) and providing more accurate local weather predictions.
- Prototyped a multilingual voice-command eCommerce system using GCP Speech-to-Text and Translation APIs, allowing customers to interact with online stores in multiple languages through natural voice queries, improving accessibility and user engagement.

EDUCATION

BS Software Engineering | *City University of Science & IT, Peshawar* Sept 2019 – Sept 2023

- **Final Year Project:** Urdu Text Sentiment Analysis – Created sentiment analysis models using deep learning and pretrained models, experimented with various embedding and NLP techniques for Urdu text. Constructed a hybrid architecture that achieved a 96% F1-Score.

CERTIFICATIONS & PUBLICATIONS

- **Certifications:** Buildables Fellowship (2025) | Building Agentic RAG with LlamaIndex, DeepLearning.AI | AI Orchestration with LangChain and LlamaIndex, LinkedIn Learning | Multimodal Prompting with Google's Project Gemini, LinkedIn Learning
- **Publication:** M. M. Danyal, **M. Haseeb**, S. S. Khan, B. Khan, S. Ullah. "Opinion Mining on Movie Reviews Based on Deep Learning Models." *Journal on Artificial Intelligence*, vol. 6, pp. 23–42, 2024.

LEADERSHIP

ML/DL Track Lead | *Bytewise Limited* Jun 2024 – Oct 2024

- Led the ML/DL mentorship track for 30+ learners across CNNs, RNNs, transformers, and deployment workflows. Designed learning paths, reviewed code, and provided feedback that moved mentees to shipping their own models.

Outreach Lead | *Google Developer Group Peshawar* Dec 2022 – Present

- Drive community partnerships and event programming for one of Pakistan's most active GDG chapters. Engaged 1.3K+ participants in DevFest and coordinated 20+ volunteers on event execution.

Campus Lead | *Google Developer Student Clubs (GDSC)* Jul 2022 – Aug 2023

- Selected as 1 of 76 GDSC Leads nationwide. Organized 35+ technical events engaging 9,000+ attendees, including a female-focused ideathon, plus 35+ mentorship sessions covering technical and career skills.