Shehryar Ashraf

Lahore, Pakistan • +92 312 4937976 • shehryarashraf99@gmail.com
github.com/shehryarashraf99 • linkedin.com/in/shehryar-ashraf-728b30100

SUMMARY

Senior AI Developer and Machine Learning Engineer with 6 years of experience taking AI systems from mathematical modelling to production across generative AI, agentic systems, natural language processing (NLP), computer vision and optimization. Currently building retrieval-augmented generation (RAG) and multi-agent products for UK legal, fintech and mortgage clients. Previously led a real-time voice AI platform (950 ms p95 latency), fine-tuned an 8B large language model (LLM) for text-to-SQL (35% to 88% accuracy), and was sole architect of a solar-site optimization engine for a US client. Owns pipelines end to end, from modelling and implementation to AWS deployment and monitoring, and mentors the engineers around them.

PROFESSIONAL EXPERIENCE

CONSULTANCY OUTFIT (CO-UK)

Senior AI Developer
June 2026 - Present • Lahore, Pakistan

- **Loophole Larry (UK legal AI)**: Building the RAG pipeline for an AI case solver over UK law, where a user submits a case and the system generates a legal strategy. Own the data pipeline, knowledge-base indexing and retrieval optimization for long-document queries: a single "query" is a case file running to hundreds of pages rather than a short prompt, which rules out standard RAG retrieval patterns. Work in progress as more data is ingested.
- **Data privacy**: Implemented PII anonymization and redaction of case data with Microsoft Presidio to meet UK GDPR requirements before documents enter the index.
- **CossmoneyX and CurrencyGram (fintech)**: Designed and pitched to the CEO a 13-agent compliance system that assesses readiness for FCA registration (crypto-asset for CossmoneyX, e-money for CurrencyGram): scrapes the regulator's published requirements, indexes them alongside the company's own policies in a vector database, and runs 11 parallel analyst agents (AML, sanctions, Travel Rule, financial promotions, etc.) plus a coordinator and a reviewer agent to produce a decision-ready readiness report with strengths, gaps and suggested improvements.
- **Property Peak (UK mortgage platform)**: Scoped and prioritised nine AI agents (case tracking, summarisation, handoff checks, ticket triage, onboarding, commission updates) across six user roles, each with role-based data access and GDPR constraints; building them as LangGraph agents behind FastAPI REST APIs with platform data exposed via the Model Context Protocol (MCP).

DEVSINC

Associate Team Lead - AI
October 2025 - April 2026 • Lahore, Pakistan

- **Voice AI**: Led the design and development of a real-time conversational video/voice AI agent with Speechmatics and Deepgram speech-to-text, RAG for grounded answers and LangGraph agents for multi-step dialogue.
- **Latency**: Migrated the stack from a Pipecat telephony prototype to the ElevenLabs API as the product moved to video, bringing end-to-end response latency to a consistent 950 ms p95; replaced a custom LangGraph web-search tool with OpenAI's built-in web search, which was simpler to integrate and faster.
- **Deployment**: Deployed the platform on AWS and set up production monitoring.
- **Mentoring**: Trained and mentored junior and intern data scientists on LLM application development, code review and deployment practices.

A1 STRATEGEZ

Data Scientist
August 2024 - October 2025 • Lahore, Pakistan

- **Text-to-SQL**: Fine-tuned Llama 3.1 8B for text-to-SQL so business users could query inventory data in plain English, raising accuracy on the client's dataset from 35% to 88% with a small, cheaply hosted model.
- **ChippyGPT**: Co-developed and deployed the backend for a ChatGPT-style assistant customised for a client's electronic-component data: proprietary prompts, the client's own database, and fine-grained control over model selection and structured output format. Optimized the end-to-end pipeline, enforced uniform output standards and monitored and refined it after launch.
- **Schema unification**: Built a text-processing pipeline that unified heterogeneous inventory schemas into one standard, improving output quality and cutting LLM token consumption by routing simple cases through classical NLP instead of the model.

ZIGRON

Math / Algorithms Expert

April 2020 - June 2024 • Lahore, Pakistan

Solar Power Generation and Optimization - Terrasmart (US)

- Sole developer and architect of an end-to-end engine that models a solar site and optimizes its energy production; reported directly to the Director.
- Built a 3D mathematical model of site layout, sun position over discrete time steps and inter-row shading between panels, then optimized panel tilt angles to maximize total daily energy output.
- Solved the constrained optimization with MATLAB Global Optimization Toolbox solvers (pattern search, particle swarm, genetic algorithm) and tuned their parameters to speed up convergence.
- Reduced runtime by vectorizing the pipeline, parallelizing it with the Parallel Computing Toolbox and porting hot paths to C++.

AI Dentistry - Automated Analysis of Dental X-rays

- Sole ML developer (initially); demonstrated the viability of AI for tooth numbering, dental diagnosis and bone-loss estimation from radiographs.
- Processed and annotated a dataset of several hundred dental images; achieved 90% accuracy on tooth numbering.
- Ran R&D on model architectures including DETR (DEtection TRansformer) and on preprocessing such as CLAHE and wavelet transforms to improve detection on low-contrast X-rays.
- Deployed the solution on a local server and presented the model's capabilities and limitations to academics at University College Dublin.

TECHNICAL SKILLS

- **Generative AI and LLMs**: Large language models (LLMs), retrieval-augmented generation (RAG), prompt engineering, LLM fine-tuning (Llama), OpenAI API, AWS Bedrock, Google Vertex AI, Azure AI, vector databases, embeddings
- **AI Agents**: LangGraph, LangChain, CrewAI, Model Context Protocol (MCP), multi-agent systems, tool use, agentic workflows
- **Voice and Speech AI**: ElevenLabs, Pipecat, Deepgram, Speechmatics, real-time conversational agents
- **Machine Learning and Computer Vision**: PyTorch, TensorFlow, scikit-learn, OpenCV, Transformers (DETR), deep learning, natural language processing (NLP), NumPy, pandas, Seaborn
- **Optimization**: Constrained and global optimization (pattern search, particle swarm, genetic algorithms), convex optimization, vectorization, parallel computing
- **Cloud, MLOps and Backend**: AWS (SageMaker, Lambda, Glue, Bedrock), GCP (Vertex AI), Docker, CI/CD (GitLab), Git, FastAPI, REST APIs, Streamlit, Apache Spark, MongoDB, Microsoft Presidio (PII redaction), GDPR compliance
- **Programming Languages**: Python, MATLAB, C++, SQL

EDUCATION

LAHORE UNIVERSITY OF MANAGEMENT SCIENCES (LUMS) Lahore, Pakistan

- MS Electrical Engineering, 2019 (Machine Learning, Deep Learning, Convex Optimisation, Information Theory)
- BS Electrical Engineering, 2015