

EDUCATION

University of Salford	Manchester, UK
MSc Data Science (Distinction)	Jan 2025 – Mar 2026
Modules: Advanced Databases, Big Data Tools and Techniques, Machine Learning and Data Mining, Applied statistics and Data Visualization	
NUCES FAST University	Pakistan
BSc Computer Science (GPA: 3.16/4)	Aug 2020 – Jun 2024
Relevant Computer Courses: Artificial Intelligence, Statistical Modeling, Deep Learning, MLOps, Generative AI, Cloud Computing, Distributed Data Engineering	

SKILLS SUMMARY

- **Agentic Orchestration:** n8n (Queue Mode, Custom Node Development, Webhooks, Error Handling), LangGraph, Model Context Protocol (MCP), Claude Code, Multi-Agent Systems (ReAct, Plan-and-Execute patterns).
- **Generative AI & LLMs:** Agentic RAG (Self-Correction, Adaptive Retrieval), Prompt Compression (LLMLingua-2), LLM-as-a-Judge (Evaluation & Grading), Fine-Tuning (LoRA/QLoRA), Function Calling/Tool Use, Structured Outputs (JSON Schema/Zod), Azure OpenAI, Local LLMs (Ollama).
- **Evaluation & Ops:** Ragas, Arize Phoenix (Tracing & Observability), Prompt Optimization, Inference Latency Reduction (TTFT Optimization), LLM Cost Optimization (Prompt Caching, TCO Modelling).
- **Data & Vector Engineering:** Qdrant, Pinecone, PostgreSQL (pgvector), Redis (Semantic Caching), Airtable (as flexible memory).
- **Core Development:** Python, TypeScript, Next.js, FastAPI, Flask, .NET (C#), Docker, AWS Lambda.
- **Other Skills:** Excellent Communication, Collaboration, Problem Solving

WORK EXPERIENCE

AI Engineer (Innovation Team) POWWR	Mar 2026 – Present
○ AI Contract Management Platform: Built an end-to-end document-AI pipeline (Next.js, TypeScript, Azure OpenAI GPT-4o, Docling) that converts supplier energy price files (PDF/Excel/email) into validated contracts, applying layout-aware parsing, strict JSON-schema extraction (Zod), fuzzy supplier-name normalization (Jaro-Winkler matching), and idempotent unit conversion into a standardized open charge schema. ○ Trust & Verification Layer: Engineered a deterministic verification layer that grounds every extracted value to its source text, applies plausibility bounds, and measures self-consistency so only low-confidence fields reach human review, achieving ~99.9% schema compliance with zero extra LLM calls for grounding. ○ Automation Platform: Co-developed an in-house pipeline-orchestration platform (Angular, .NET, Dagster) replacing a commercial RPA product; its Open/Closed adapter framework registers a new automation target with one adapter class and one DI line, cutting new-integration delivery from weeks to days. ○ Production Pipelines: Built reconciliation pipelines on the platform that combine, key, and CRM-match 12,000+ records per run at ~96% auto-match with automated alerts for unmatched and duplicate data, fully replacing a manual, error-prone process. ○ AI Enablement: Delivered foundational training for the company-wide Claude rollout as one of the lead trainers, covering agentic workflows, model selection, token economics, and MCP connectors. ○ Local Model Research (ongoing): Benchmarking open-weight local models (Ollama) against GPT-4o for document extraction and designing a confidence-gated model cascade projected to cut extraction costs by 70-95%; prompt caching and token compaction already reduced input tokens by 15-30% per document. Skills: TypeScript, Next.js, Azure OpenAI, Ollama (Local LLMs), Model Context Protocol (MCP), Claude Code, Docling, .NET (C#), Angular, Dagster, Prompt Caching, LLM Evaluation, Jira, Confluence	
Generative AI Engineer Sparkix Technologies	Sep 2024 – Dec 2024
○ Developed advanced Generative AI applications using Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) techniques. ○ Created robust and scalable Flask and FastAPI endpoints to enable efficient interaction with AI models. ○ Worked extensively with OpenAI APIs for chatbot development, vision models for quality reporting, structured JSON outputs using LangChain, and creating various chains.	

- Deployed APIs on cloud servers using Azure and AWS.
- Conducted research on innovative applications like music and lyrics generation and integrated SSML features into solutions.
- Collaborated with cross-functional teams to define project requirements and contribute to the successful delivery of AI-powered solutions.

Skills: Langchain, LangGraph, LangServe, Prompt Engineering, OpenAI, Vector Databases, AWS, MySQL, Jupyter Notebook, Excellent Communication, Collaboration, Problem Solving, Flask

Generative AI Engineer | Horizon Tech Services

May 2024 – Nov 2024

- Working on a Speech Enhancement project, improving audio quality using Generative AI models **GAN** (PyTorch-based) and **Transformer Based Models** (PyTorch-based).
- Prepared and augmented data to enhance model training.
- Selected, trained, and fine-tuned sep-former models and GAN models for optimal performance.

Skills: LoRA (Low-Rank Adaptation), Transfer Learning, Hyperparameter Optimization, Post-Training Quantization, Mixed Precision Training, VS code, TensorFlow, PyTorch, SpeechBrain, Pandas, Numpy, Flask, Scikit-Learn, Matplotlib, Jupyter Notebook, Visual Studio Code, Problem Solving

Deep Learning Intern | GrayHat & FYP (Final Year Project)

Sep 2023 – May 2024

- Leveraged Large Language Models and developed a backend system using Flask to automatically dub videos into other languages with an accuracy of 85%.
- Conducted fine-tuning of the **Whisper** model by OpenAI on Urdu datasets, resulting in a 4% reduction in Word Error Rate (WER).
- Collaborated closely with team members to ensure timely achievement of project objectives.

Skills: Fine-tuning, TensorFlow, PyTorch, Generative AI models (GAN, sep-former, diffusion models, U-net, BERT, BART, XTTS-V2, Bark, Whisper, RVC)

PROJECTS

Omni-Channel AI Executive Assistant (n8n + Meta APIs)

- **System Architecture:** Architected a centralized n8n orchestration layer that unifies communication across WhatsApp Business, Instagram Graph API, and Facebook Messenger, acting as a 24/7 personal executive assistant.
- **Agentic Workflow:** Designed a router agent that uses LLM classification to autonomously categorize incoming messages (e.g., "Urgent," "Casual," "Business Opportunity"). Routine inquiries trigger automated, context-aware replies, while high-priority items are routed to a Human-in-the-Loop verification stage via a private Telegram bot.
- **Activity Reporting:** Implemented a "Observer Agent" that logs daily interactions and task completions into Notion/PostgreSQL. The system runs a nightly cron job to synthesize these logs into a concise "Daily Activity Briefing" using LLM, allowing for effortless tracking of personal productivity and social obligations.
- **Latency Optimization:** Optimized the webhook handling pipeline to process messages in parallel (n8n Queue Mode), ensuring sub-second response times even during high-volume message spikes.

Autonomous Recruitment Agent with "LLM-as-a-Judge" Evaluation

- **Evaluation Pipeline:** Built a production-grade resume screening agent that uses the LLM-as-a-Judge pattern to grade candidates. Instead of generic summaries, the "Judge Agent" evaluates resumes against a strict rubric (e.g., "Has 3+ years in Python," "Experience with Vector DBs") and outputs a confidence score for each criterion.
- **Reliability Engineering:** Validated the agent's performance by creating a "Golden Dataset" of 100 human-reviewed resumes. Used Ragas and Arize Phoenix to measure the alignment between the AI Judge and human recruiters, achieving an 85% correlation score and reducing false positives by 40%.
- **Tools & Protocols:** Standardized the data ingestion using Model Context Protocol (MCP) servers to securely fetch resumes from Google Drive and update candidate status in Airtable without brittle API integrations.

Privacy-Preserving PII Redaction Pipeline (Small Language Models)

- **Edge AI Implementation:** Engineered a local-first NLP pipeline to detect and redact Personally Identifiable Information (PII) from sensitive documents.
- **Cost & Privacy:** Replaced expensive cloud APIs with a fine-tuned Microsoft Phi-3 (SLM) running locally. This reduced inference costs to near-zero and ensured 100% data sovereignty, making the solution compliant with strict GDPR/privacy standards.

Agentic Travel Intelligence Dashboard

- **Autonomous Workflow:** Engineered a self-directed agentic workflow that leverages **Google Maps API** and **function calling** to autonomously query, filter, and rank hotel data based on complex user constraints (price, rating, proximity).
- **Data Transformation:** Built a robust ETL pipeline that transforms raw API JSON responses into structured business intelligence, feeding a real-time **PowerBI** dashboard for actionable insights.

Speech Enhancement using Transformer Architectures

- **Model Optimization:** Fine-tuned a Transformer-based architecture for audio restoration, achieving industry-leading metrics of **14 dB SI-SNR** and **2.25 PESQ**.
- **Real-Time Serving:** Deployed the enhanced model via **Streamlit Cloud**, optimizing the inference pipeline to handle real-time audio streams with minimal latency.

Localized Text-to-Speech Engine (Fine-Tuned XTTS-V2)

- **Domain Adaptation:** Fine-tuned the **XTTS-V2** generative model on a custom dataset of Asian English accents, utilizing **Transfer Learning** to significantly improve prosody and naturalness for underrepresented dialects.
- **Data Pipeline:** Developed an automated audio preprocessing pipeline to clean, normalize, and segment training data, ensuring high-fidelity model convergence.

End-to-End FOG Prediction Pipeline (MLOps)

- **Automated CI/CD:** Implemented a full-scale MLOps pipeline using **GitHub Actions** and **Docker**, automating the data collection, preprocessing, model training, and deployment lifecycle.
- **Lifecycle Management:** Leveraged **MLflow** for experiment tracking and model registry, ensuring full reproducibility and version control (**DVC**) across all training iterations.

CERTIFICATES

Advanced Agentic AI & Orchestration

- **Claude Code: A Highly Agentic Coding Assistant** – *DeepLearning.AI (2026)*
- **MCP: Build Rich-Context AI Apps with Anthropic** – *DeepLearning.AI (2026)*
- **AI Agentic Design Patterns with AutoGen** – *DeepLearning.AI (Jan 2026)*
- **Agentic AI** – *DeepLearning.AI (Dec 2025)*
- **Building Agentic RAG with LlamaIndex** – *DeepLearning.AI (Dec 2025)*
- **Building AI Voice Agents for Production** – *DeepLearning.AI (Jan 2025)*
- **Building Live Voice Agents with Google's ADK** – *DeepLearning.AI (Dec 2024)*

Core Generative AI & Engineering

- **Pretraining LLM** – *DeepLearning.AI (Jan 2026)*
- **Prompt Compression and Query Optimization** – *DeepLearning.AI (Jan 2026)*
- **Vector Databases: From Embeddings to Applications** – *DeepLearning.AI (Aug 2024)*
- **LangChain for LLM Application Development** – *DeepLearning.AI (Jul 2024)*
- **Preprocessing Unstructured Data for LLM Applications** – *DeepLearning.AI (Jun 2024)*
- **Generative AI with Large Language Models** – *Coursera (Nov 2023)*

Cloud & MLOps

- **AWS Academy Graduate: Microservices & CI/CD Pipeline Builder** – *AWS (Apr 2024)*
- **Open Source Models with Hugging Face** – *DeepLearning.AI (May 2024)*