SAMEEN ISLAM

Ealing, London | 07455 083989 | sameen11@hotmail.com | sameen.info | linkedin.com/in/sameen-islam

Profile

Applied AI engineer specialised in building and evaluating agentic pipelines, RAG systems, and LLM-powered workflows at scale within highly regulated enterprise environment.

Technical Skills

Languages	Python (primary), Java
AI & ML	PyTorch, OpenAI, Gemini, LangChain, LangGraph, RAG, LLM evaluation, function-calling, prompt engineering, agentic systems, signal processing, time-series ML
Cloud & Infra	FastAPI, MLflow, Docker, Linux, Bash, uv, pip, Microsoft Azure (AI Foundry, Static Web Apps, Functions, Cosmos DB, Batch, Blob Storage, DSVM), Amazon AWS (EC2, Lambda, S3, DynamoDB, Batch)
Data & ML Tooling	SQL, PostgreSQL, pgvector, FAISS, DuckDB, scikit-learn, NumPy, pandas, polars, matplotlib, seaborn, Metaflow

Professional Experience

Citigroup — London, UK
Generative AI Engineer — AVP — Nov 2024 – Mar 2026 (16 months)

Applied engineering team for Citi Assist — a generative AI assistant deployed to 140,000+ employees across 8 countries, built on Google Vertex AI with Gemini and Claude models.

- Built an end-to-end pipeline (ETL architecture) for LLM evaluation supporting a Gemini model upgrade governance gate for the 140,000-user production system, performing model calibration from human feedback and ensuring support for full reproducibility and auditability.
- Designed and executed controlled evaluation experiments on a production agentic function-calling router, defining independent variables (system prompt, function descriptions), control variables (conversational dataset, trial count), and dependent variables (matched, mismatched, and uniquely chosen routes) to systematically identify unstable routes and quantify routing consistency against a controlled baseline.
- Developed custom per-item diagnostic metrics (model variance across 5 runs, absolute bias against human gold labels, and a composite loss score combining both) to surface items that were simultaneously wrong and unstable; followed by percentile-based rank-and-select failure set construction to prioritise human review.
- Developed a score-conditioned self-critique prompting technique: after ranking failure cases by composite loss, prompted the model to reason about why each item scored low and what was required to score higher, targeting improved alignment with human annotators without modifying model weights or evaluation data.
- Collaborated with subject-matter experts across business lines to construct human-labelled datasets for regression evaluations and human correlation benchmarking, communicating methodology and results to cross-functional stakeholders.

Provenir AI — London, UK
Software Engineer — Sep 2022 – Oct 2024 (25 months)

- Architected and developed large-scale ML pipeline for financial fraud detection through graph analysis. Deployed on AWS compute cluster — owned full R&D, system design, and implementation — achieved up to 4× reduction in end-to-end runtimes (from days to hours) through profiling, parallelisation, and containerisation, with material reduction in aggregate compute utilisation.
- Acted as technical owner for key ML-based fraud detection systems, with the work recognised in a successful £400,000 HMRC R&D tax relief claim — validating the technical novelty of the engineering.

Earlier Experience

Disperse.io	Junior Computer Vision Engineer	London, UK	Jun 2022 – Jul 2022
Accenture	Software Developer, Analyst	London, UK	Nov 2018 – Aug 2019
Airbus Defence & Space	Spacecraft Software Developer	Stevenage, UK	Sep 2016 – Sep 2017

Education

University of Southampton — Southampton, UK
Master of Science (MSc) Artificial Intelligence — Graduated Sept 2021 (12 months)

Thesis — Computational Biomedical Signal Processing: Real-Time Seizure State Classification from Multi-Channel EEG — `github.com/codexsameen/chbmit-seizure-prediction`
numpy • scikit-learn • yasa • matplotlib

Built a real-time biomedical ML system to classify brain states (interictal vs preictal) from raw multi-channel EEG data. Open-sourced and adopted within the biomedical ML research community.

- *Feature extraction:* Engineered two independent pipelines — (1) a time-domain ARMA pipeline using sliding-window least-squares estimation (Moore-Penrose pseudoinverse) to extract AR parameters per channel, and (2) a spectral pipeline using Welch FFT to compute relative bandpower across six neural frequency bands (δ, θ, α, β, γ, Γ).
- *Classification:* Trained class-balanced SVM (Support Vector Machine) classifiers (linear and RBF kernels) on extracted features for binary state prediction.
- *Signal smoothing:* Implemented a Kalman filter from scratch — full predict/update cycle with tunable process noise (Q) and measurement noise (R) hyperparameters — to regularise the real-time prediction signal.
- *Delivery:* Model training and streaming inference with CLI.

Queen Mary University of London (QMUL) — London, UK
Bachelor of Science (BSc) Computer Science — First Class with Honours — Graduated July 2018 (48 months)

Open-Source Projects

Job Role Evaluator — `job-eval.tools.sameen.dev`
Azure Functions (Python) • Cosmos DB (NoSQL) • Azure OpenAI • GitHub OAuth • BeautifulSoup • httpx

- Built and deployed a full-stack LLM-powered job evaluation tool using a serverless API backend; taking the product from concept to live deployment. Core pipeline: web scraper fetches and parses job descriptions from arbitrary URLs, then an LLM scores each JD against a user-specific rubric via structured JSON output with prompt-engineered YAML templates.
- Implemented Dirichlet-prior empirical Bayesian weight adaptation: using asymmetric positive/negative evidence accumulation, enabling the scoring rubric to learn user preferences over time.
- Designed an async evaluation architecture using background threads as a cost-effective workaround instead of durable serverless functions for long-running LLM calls.