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 environments.

TECHNICAL SKILLS

Languages Python (primary), Java

AI & ML PyTorch, OpenAI, Gemini, LangChain, LangGraph, RAG, LLM, Evaluation, function-calling, prompt engineering, agentic systems, scikit-learn, unstructured.io

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, NumPy, pandas, polars, matplotlib

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. Day-to-day owner of implementation across retrieval and evaluation systems; presented architectural decisions through ADR reviews with senior architects.

- **Productionised retrieval over 120 policy documents** by replacing fixed-window chunking with a structure-aware pipeline: pre-processed PDFs to Markdown via Unstructured.io, then split on header hierarchy using LangChain to preserve logical document structure; embedded and indexed chunks into a Postgres store for the assistant's RAG layer.
- **Shipped the LLM model upgrade for the 140,000-user assistant** ahead of Google's deprecation of the prior Gemini version, clearing the Model Risk Management governance gate. Built the end-to-end ETL pipeline backing the upgrade case: batch evaluation, calibration against expert labels, and alignment between automated metrics and human judgement, all reproducible and auditable.
- **Engineered a queue-based batching layer around the Gemini API** to operate within a 20 RPM quota at evaluation scale, enabling large batch runs without throttling or partial-failure recovery issues; prerequisite for the upgrade evaluation to complete on schedule.
- **Identified 12% of routes as unstable on the production agentic function-calling router** and remediated them via targeted prompt tuning. Designed controlled experiments isolating prompt and tool-description changes against fixed conversational datasets, quantifying routing consistency against a baseline to surface regressions before release.
- **Developed diagnostic metrics to triage model failures at scale**, combining variance across repeated runs with bias against human gold labels into a composite loss; used percentile-based ranking to construct prioritised failure sets, focusing scarce human-review effort on items most likely to be both wrong and unstable.
- **Shipped a score-conditioned self-critique technique into production** to improve model–human alignment without relabelling: ranked failure cases by composite loss, then prompted the model to reason about why each scored low and what would raise it to improve assistant performance in production.

Provenir AI · *London, UK*

Software Engineer · Sep 2022 – Oct 2024 (25 months)

ML engineering for a major credit bureau's KYC and fraud-screening platform, used by banks and lenders to verify customers at onboarding.

- **Architected and developed a large-scale ML pipeline for financial fraud detection** through graph analysis and classical ML model inference (XGBoost, LightGBM). 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 HonoursGraduated 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.