

Stephen Robinson

AI Systems Architect & Data Architect

Global Head of Risk IT Reporting — Nomura

s.robinson@samuraicomputing.com

020 3780 8679

steve.motlogs.com

Shefford · Greater London, UK

Thirty years architecting data at banking scale — now building the production AI that runs on it. I bring 30 years of banking-grade data architecture to production AI: I designed and shipped the MOTLogs estate — five live production systems spanning a voice booking agent, a self-grading agentic RAG knowledge base, a predictive MOT-log analytics engine, a 136M-record risk-scoring engine, and a code-intelligence platform exposed to AI agents over MCP.

Global Head of Risk IT Reporting, Nomura

Microsoft SQL Server MVP

BSc Computer Science, Loughborough

30+ years in enterprise data

30+

Years architecting data

49TB+

Risk platform tuned at Nomura

5,000+

Regulated reports / day

MVP

Microsoft SQL Server MVP

Track record — Nomura, 2009 – present

Global Head of Risk IT Reporting · Risk SQL Team Lead. Before MOTLogs, three decades making data dependable at the scale and regulatory rigour a global bank demands. This is the foundation the AI work is built on — not a pivot away from it.

- ▶ Architectural guidance and leadership across multiple global teams

▶ Introduced Data Governance & Master Data Management into Risk Reporting

▶ Tuned and maintained 49TB+ risk SQL platforms powering VaR
- ▶ 5,000+ reports a day delivered across Market & Credit Risk

▶ Designed a new global reporting platform on event-driven microservices

▶ Doubled the team within 8 months

The estate — five production systems, architected end to end

Agent — Voice AI

p95 · latency held as an SLA gate

Customers ring a normal phone number and talk to an AI receptionist that books appointments against a live calendar, answers FAQs from the garage's knowledge base, takes messages, and transfers to a human when it matters. Every conversation flow is data-driven, so a new garage is configured from an admin UI — no code change.

Wiki — Agentic RAG + governance

6 · surfaces, one RAG core

A tenant-isolated knowledge base that answers customer and staff questions with retrieval-augmented generation and grounded citations. When confidence is low it escalates to a human over Twilio Video, and it continuously learns from ratings, detects drift, and flags weak answers for a governance reviewer.

MOTLogs — Predictive log analytics

200M+ · vectors searched per query

MOTLogs ingests vehicle MOT test logs from the DVSA history API and user uploads, turns each test's failures and advisories into a semantic narrative, and predicts the likely outcome of a vehicle's next test by comparing it against cohorts of look-alike vehicles. It powers a garage-facing web app and an analytics API for compliance and test-quality review.

Scout — Risk-scoring engine

136M · vehicle records scored

Scout ingests the entire UK DVSA MOT dataset — ~49GB, 136 million records — and gives every vehicle a deterministic 0–100 risk score from an eight-stage pipeline that weighs defect patterns, mileage, ownership behaviour and model-specific failure rates. A registration lookup then returns an instant, mechanic-grade risk read for auction buyers.

Second Brain — Code-intelligence MCP

36 · agent tools in one MCP server

OpenBrain parses every repository in the estate into a hybrid semantic + structural index — a code knowledge graph — and serves it to AI agents over MCP. An agent can ask where a function lives, what calls it across repos, why it exists, and what a change would break — and get grounded, cited answers instead of guesses.

Capabilities mapped to 2026 demand

Agentic RAG, done for real

A Wiki knowledge base with hybrid dense + full-text retrieval, a cross-encoder reranker (bge-reranker-base, int8 ONNX), GraphRAG, and grounded citations carrying source URLs into every answer.

LLMOps & evaluation

Drift detection, self-evaluation runs, RLHF reweighting, active-learning queues and golden-set ranking validation — wired as hosted services with a governance portal a human reviewer signs off in.

Governance, scale & audit

The instinct from 30 years of regulated banking data: multi-tenant row-level security, DSAR/privacy, a guardrail hook that blocks agents answering before they consult the source of truth, and audited overrides.

MCP & tool-using agents

A production MCP server (mot-ledger) exposing 25+ tools — semantic search, dependency graph walks, 'why does this exist' provenance, invariant guards, refactor planning — that turn a whole codebase into a queryable substrate for AI agents.

Voice & real-time AI

A Twilio ↔ ElevenLabs ConvAI orchestrator with a custom-LLM contract, spike-driven engineering, and explicit, measured SLA budgets (p95 latency gates) — built on telephony I'd already shipped at PhoneBounce.

Ship & operate it

Reproducible Hetzner infrastructure (idempotent bootstrap, Tailscale, disaster-recovery runbook), Stripe billing, nginx/Kestrel, and commit-driven re-indexing — the unglamorous parts that make AI a business.

How I work

Scale and audit are not afterthoughts

Thirty years delivering regulated reporting at a global bank means I design for governance, lineage and recovery from day one — the questions a board will ask are answered before they're asked.

Agents that build software

I orchestrate AI agents to do real engineering — a workflow-driven QA loop ran parallel testers, adversarially verified findings, auto-fixed and redeployed, driving confirmed bugs 13 → 2 to convergence.

Evidence over vibes

Embedder choices A/B-tested, latencies measured live, SLA budgets re-based against reality with a written justification. Decisions leave a trail you can audit.

Reproducible or it didn't happen

Lose the server, run one idempotent bootstrap, you're back. Infrastructure as a runbook, not as tribal knowledge in one person's head.

Stephen Robinson · Samurai Computing

s.robinson@samuraicomputing.com · linkedin.com/in/samuraicomputing/