Guy Freeman, PhD

AI Evaluation Scientist — LLM & agent evaluation · experimental design · Bayesian statistics

Hong Kong (permanent resident) · mail@gfrm.in · gfrm.in ·
linkedin.com/in/guyfreemanstat · github.com/gfrmin

Open to permanent, contract and consulting roles · open to relocation or remote

Summary

Statistician (Warwick PhD, Oxford BA, 1,750+ citations) who designs AI workflows and measures whether they actually work, including when the result is bad news for my own system. At Booking.com I designed the evaluation for an agentic LLM recommender and used it to redesign the system: task completion up ~2.7×, latency down an order of magnitude. At HSBC I design and measure LLM workflows for compliance and fraud. My independent research is on evaluating and overseeing autonomous AI agents: leak-proof scoring, power analysis, checks proven against adversarial candidates, and pre-registered criteria that AI builders cannot edit.

Core Skills

Evaluation design: offline evaluation frameworks, LLM-as-judge, variance decomposition, ablations, paired and A/B comparisons, power analysis, pre-registration, error analysis, proper scoring rules, ship / no-ship decisions

Statistics: Bayesian inference, causal inference, decision theory, value of information, sequential and anytime-valid testing · **Tools:** Python, SQL, R, Julia; Claude, GPT, Gemini; Arize, Langfuse

Experience

Senior Data Scientist (Contract) — HSBC · Hong Kong · Aug 2026 – Present

- Designed and measured a RAG workflow, with guardrails, that answers third-party vendor compliance questionnaires.
- Designing and measuring improved address remediation to catch fraud: turning raw customer address data into signals of financial crime.

Senior ML Scientist (Contract) — Booking.com · Tel Aviv · Mar 2026 – Jun 2026

- Designed a multi-arm offline evaluation framework (LLM-as-judge scoring, variance decomposition, automated prompt optimisation) to compare configurations of an agentic LLM recommender and choose models.
- Used it to redesign the system from a multi-step agentic loop into a leaner deterministic workflow: task completion up ~2.7×, latency down an order of magnitude, from research prototype to production.
- Separated a genuine model regression from an evaluation artefact through ablations; defined metrics and judge design for a new RAG itinerary-extraction feature; built judge-based metrics, trace-level observability (Arize) and same-day A/B harnesses.

Lead AI Engineer — Slice | Global Equity · Tel Aviv · Jul 2025 – Jan 2026 · Regulated fintech

- Designed the document-AI pipeline (OCR plus Claude and Gemini extraction) that replaced manual onboarding: 500+ legal and financial documents in 20 minutes at 95%+ accuracy.

Machine Learning Engineer — Velotix · Ramat Gan · Oct 2024 – May 2025

- Designed anomaly detection that surfaced unauthorised database permissions across 1M+ tables and 50k+ users.

Senior Data Scientist (Team Lead) — Willis Towers Watson · Tel Aviv (Remote) · May 2022 – Oct 2024

- Led the data science behind an LLM job-evaluation product deployed to enterprise clients; early GPT-3 adopter; led a distributed team.

Earlier · Hong Kong and Tel Aviv · 2011 – 2022

- Data Scientist, Entrée Capital (deal-sourcing model: 3 pre-seed startups found before announcement, all funded); HSBC; Prudential Asia; EY; founder, Data Guru; Postdoctoral Fellow, University of Hong Kong (causal inference).

Research

Evaluating self-improving AI agents (assize) · 2026

- Sealed evaluation protocol scoring agents on decisions (utility on never-issued test seeds) and on predictions committed before each observation (log-loss), with compute priced in. Bayesian bootstrap power analysis: a 0.5-achievement improvement in Crafter is detectable at 33 episodes per arm.
- Every check validated against adversarial candidates (clairvoyant, sandbox-escaping, crash-hiding, misreporting compute): 81 of 81 seeded faults caught. AI builder, reviewer and auditor sessions with every forecast scored by log-loss.

Pre-registered evaluation of AI-built systems · 2026 · public essays at gfrm.in

- Acceptance tests signed before any code exists, so an AI builder cannot fit the checks; a red team that found 15 ways past my own verification gate, each turned into a seeded failure the gate must catch; a pre-committed adoption gate that refused to adopt my own tool.

Credence: decision-theoretic LLM agents · 2024 – 2026

- When a one-line greedy ablation beat my full Bayesian agent under fair conditions, I retired my own published headline and reported what survived: the value of exploration depends on attribution quality (+27 with clean categories, −12 with inferred ones).
- Governing a coding agent's tool calls by expected utility: wasted repeat calls blocked at precision and recall 1.0 on held-out sessions; an injected exfiltration flagged at 0.94 precision, interrupting 1.2% of safe sessions. Model routing: −96% cost, −52% latency vs always-Sonnet.

Education & Publications

PhD, Statistics and **MSc (Distinction), Statistics** — University of Warwick · **BA, Mathematical Sciences** — University of Oxford

1,750+ citations, including *BMC Medicine* (697) and *Int. J. Epidemiology* (190). Essays include "The Ablation That Beat the Agent" and "The Bitter Lesson Has No Utility Function" (Hacker News front page). English (native), Hebrew (advanced).