Mohammad Kasra Sartaee

Evaluation and Optimisation of LLM and Generative AI Systems

Dubai, United Arab Emirates • +971 50 180 2335 • kasrasartaee@yahoo.com
kasra-sartaee.github.io • github.com/Kasra-Sartaee

Research Profile

I work on how to evaluate and configure LLM and generative AI systems reliably when their behaviour is stochastic and their objectives conflict: experimental design and statistics for randomised systems, multi-objective optimisation and model selection, uncertainty and calibration, and the efficiency and energy cost of AI. My MSc dissertation compared evolutionary, Bayesian and random search for tuning a multi-agent LLM repair pipeline under a matched budget; my BSc work, published in the AAAI Symposium Series, fine-tuned a generative music model and evaluated it against both objective metrics and listeners. I am first author of a survey on uncertainty, disagreement and information fusion in AI systems. I am seeking a funded PhD on principled evaluation of AI systems.

Education

MSc Artificial Intelligence and Data Engineering 2025 – 2026
University College London (UCL), London

- Dissertation submitted September 2026; final classification expected December 2026.

BSc (Hons) Computer Science (Artificial Intelligence), First Class Honours 2021 – 2025
Heriot-Watt University, Dubai

- Deputy Principal's Award for Academic Excellence, 2021–2025 (four consecutive years). 100/100 in Logic and Proof.

Publications

- **M. K. Sartaee** and K. Karim. "EchoScript: Enhancing AI Music Generation for Cinematic Scoring via Script-Aware Fine-Tuning." *Proceedings of the AAAI Symposium Series*, 6(1):323–332, 2025. doi:10.1609/aaaiss.v6i1.36069
- **M. K. Sartaee** and M. Q. Al-Shihani. "Uncertainty, Disagreement, and Information Fusion in Modern AI Systems: A Comprehensive Survey." *International Journal of Data Science and Advanced Analytics*, 7(2):473–484, 2025. doi:10.69511/ijdsaa.v7i2.327

Research Experience

MSc Dissertation, University College London 2025 – 2026
Multi-Objective Search over the Decoding Policy of a Multi-Agent LLM Pipeline for Automated Program Repair
pymoo, BoTorch/GPyTorch, Ollama, Intel RAPL github.com/Kasra-Sartaee/apr-decoding-policy-search

- Formulated the joint decoding-and-retry policy of a four-stage analyse–repair–verify–optimise LLM pipeline (qwen2.5-coder 1.5B) as an 11-dimensional search problem under three objectives: test-pass fraction, cyclomatic complexity, and execution energy of the repaired program.
- Compared NSGA-II, multi-objective Bayesian optimisation (qLogNEHVI over per-objective GPs) and uniform random sampling against a repeated default policy at a matched 750-evaluation budget, five seeds per arm (791 host-hours); searched on Codeflaws (C) and transferred unchanged to RunBugRun (Python).
- Random sampling beat NSGA-II on held-out correctness with complete separation across seeds (Cliff's $\delta = 1$, $p_{\mathrm{Holm}} = 0.024$). Selecting on training correctness (exploratory) raised held-out correctness by 3.4–5.2 points in every arm–benchmark cell, whereas a balanced three-objective rule did worse than random selection from the same front. Correctness rankings transferred across languages ($\rho = 0.63$, 665 policies); energy did not.
- Designed a two-phase energy protocol separating concurrent generation from serial RAPL measurement, cutting run-to-run spread from 4–17% to 2–3%, and identified a measurement-epoch confound (1.68$\times$ energy shift across a reboot with correctness unchanged).
- Analysis at the repetition unit: Mann–Whitney U, Cliff's δ, Holm correction, task-grouped bootstrap, and hypervolume, IGD$^+$ and ε^+ with cardinality control and a power analysis.

BSc Dissertation, Heriot-Watt University 2024 – 2025
Script-Aware Fine-Tuning for Cinematic Music Generation Supervisor: Kayvan Karim
PyTorch, Hugging Face, MusicGen Graded First Class; published at AAAI-SS 2025

- Fine-tuned MusicGen on a curated dataset annotated with genre, mood, instrumentation, tempo and narrative context, conditioning soundtrack generation on film scripts.

- Evaluated against the base model with Fréchet Audio Distance (4.37 vs. 4.55) and structured listening tests, in which participants preferred the fine-tuned model for musical quality and narrative alignment.

Selected Projects

Probabilistic Forecasting of Geomagnetic Indices 2025 – 2026
UCL | PyTorch

- Transformer and LSTM forecasters with probabilistic output heads, calibrated by temperature scaling and compared against a physics-informed hybrid baseline.

Formal Verification of Python Standard-Library Code 2025
UCL | CrossHair, Z3

- Verified CPython standard-library functions by symbolic execution (32 of 33 contracts), and seeded a deliberate counterexample to test where the method fails.

Distributed ML Inference Platform 2025 – 2026
UCL | Kubernetes (K3s), Terraform, Ansible, Docker, MinIO

- Provisioned a 5-node inference cluster (22 cores, 140 GB RAM) as infrastructure-as-code with object storage and a message-queue task layer; root-caused failures under resource pressure and re-architected the storage layer to remove them.

Teaching

Teaching Assistant: Functional Programming; Compiler Design 2024 – 2025
Heriot-Watt University, Dubai

- Tutorials and lab support for 30+ undergraduates (lambda calculus, type systems, parsing); wrote supplementary materials and contributed to problem sets and marking rubrics.

Skills

ML and optimisation: PyTorch, Hugging Face Transformers, scikit-learn, BoTorch/GPyTorch, pymoo; LLM fine-tuning, local LLM serving (Ollama), quantised models, calibration, multi-objective and Bayesian optimisation.

Statistics and evaluation: SciPy, statsmodels, pandas; experimental design for stochastic systems, non-parametric testing, effect sizes, multiple-comparison correction, bootstrap, listening tests.

Systems: Linux, Docker, Kubernetes, Terraform, Ansible, Git; hardware energy measurement (RAPL); long-running reproducible experiments (checkpoint/resume, provenance manifests).

Programming: Python (expert), C, Java, OCaml, SQL, Bash, LaTeX.

Languages: English (IELTS Academic 8.5/9, C2), Persian (native).

References

Available on request.