Kian Maleki — AI Safety Research | Reasoning & Evaluation | Machine Learning | Physics PhD

+1 (737) 600 9803 | kian@drkianmaleki.com | drkianmaleki.com | linkedin.com/in/kian-maleki-phd | github.com/drkianmaleki
University of Iowa (2025) PhD in Physics (GPA: 4.03) | Creighton University (2017) MS in Physics (GPA: 4.00)

Research Profile

- Empirical AI-safety researcher focused on reasoning-model behavior, grounded in 10+ years of mathematical modeling, statistical inference, and scientific computing as a physics PhD.
- Completed a pre-registered study of chain-of-thought faithfulness (Aug 2026), using prefix interventions and resampling on reasoning traces; all hypotheses adjudicated against frozen thresholds and the full analysis is reproducible from the public repository.
- Author of two machine-learning methodology manuscripts, one submitted to *Machine Learning* and one under review at *Pattern Recognition*, alongside peer-reviewed physics publications in *Physical Review B* and *Nano Letters*.

AI Safety Research

Decided Mid-Thought: Locating Answer Commitment in Reasoning Models 2026

Methods/Tools: Python, OpenRouter/DeepInfra APIs, Qwen3.6-27B (thinking mode), Prefix Interventions, Resampling Experiments, Pre-registered Study Design

link: https://github.com/drkianmaleki/decided-mid-thought

- Designed and pre-registered a study testing a core assumption behind chain-of-thought monitoring and oversight: whether visible reasoning traces reveal when a reasoning model commits to its answer, probed via prefix interventions and resampling on thinking traces.
- Constructed underspecified variants of famous problems (e.g., a bat-and-ball problem with a premise deleted) where the model recalls the famous version in its chain of thought on essentially every sample, yet its final answer is governed almost entirely by how the answer options are phrased.
- Found no single-sentence commitment point (largest one-sentence causal effect $+0.28$ against a pre-registered 0.4 threshold); the only threshold-clearing change was a pre-specified 15-sentence region ($+0.48$), while the default answer was $\sim$80% determined by the prompt alone and the minority answer was constructed in-trace ($0.12 \rightarrow 1.00$), with 84.6% of the trace generated after the answer had saturated (500 resampled continuations).
- Froze the pre-registration before the resampling runs and committed all scripts, item files, and raw run data for full reproducibility, with pinned model/provider settings and documented sampling-variance and backend-drift limitations.
- Exercised the single permitted classifier revision after a seeded, blinded hand-check (19/20 agreement) exposed extractor defects; every one of the 35 re-scored records was validated by eye, and the headline provenance claim reproduces from a clean checkout with one committed script.

Machine Learning Methodology Research

Ambiguity Range Framework 2026

Methods/Tools: Python, scikit-learn, XGBoost, pytest, Statistical Testing, Model Evaluation

link: https://github.com/drkianmaleki/ambiguity-framework

- Developed a model-agnostic diagnostic framework for detecting decision-boundary ambiguity that aggregate metrics such as AUC-ROC can hide, directly supporting trustworthy classifier evaluation.
- Introduced ambiguity-focused metrics that quantify prediction concentration near decision boundaries and threshold-sensitive degradation in precision, recall, and F1.
- Built reproducible experiments with calibration analysis, bootstrap confidence intervals, paired statistical tests, and automated test coverage.

Sequence Acceleration Benchmark for Gradient Boosting 2026

Methods/Tools: Python, XGBoost, scikit-learn, NumPy, Statistical Testing, Reproducible Benchmarking

link: https://github.com/drkianmaleki/sequence-acceleration-benchmark

- Developed a benchmark for finite-horizon learning-curve prediction in gradient boosting, evaluating 51 sequence-acceleration methods from 13 families across 18 synthetic convergence regimes.
- Found that classical sequence accelerators largely fail under realistic noisy, finite-horizon conditions while simple curve fits extrapolate more reliably; developed failure-detection and conservative extrapolation rules for operational use.

Research Experience

Creighton University | Resident Assistant Professor of Physics Aug 2025 – Jul 2026

- Taught undergraduate physics courses emphasizing quantitative reasoning, modeling, optics, electronics, and applied problem-solving.
- Developed interdisciplinary lab curriculum integrating AI tools and chatbots into scientific workflows while emphasizing responsible and ethical use.

University of Iowa | Graduate Research Assistant, Physics Aug 2018 – Aug 2025

- Developed high-dimensional numerical models for complex physical systems, combining statistical inference, optimization, and large-scale simulation.
- Optimized computational workflows, reducing simulation runtime from days to under one hour through algorithmic improvements and efficient implementation.
- Analyzed noisy, high-dimensional experimental and simulation data to identify structure, extract signal, and validate theoretical predictions.
- Contributed to peer-reviewed publications in condensed matter physics, quantum materials, and solid-state integration of quantum devices.

Creighton University | Graduate Research Assistant, Physics Aug 2015 – Aug 2018

- Built simulation-based models for physical systems using Python and Fortran, including N-body simulations and molecular dynamics.
- Designed parallelized computation pipelines for large-scale simulations, improving scalability with HPC resources.
- Developed interactive scientific software tools for simulation and analysis, enabling user-driven exploration of model behavior.

Technical Expertise

- **Research & Evaluation:** Pre-registration, experimental design, statistical inference, hypothesis testing, bootstrap confidence intervals, paired statistical tests, uncertainty analysis, Monte Carlo methods.
- **LLM Research:** Python, PyTorch, Transformers, OpenRouter/OpenAI-compatible APIs, reasoning-trace analysis, model evaluation, calibration.
- **Scientific Computing:** NumPy, Pandas, Polars, Numba, Fortran, Linux/Bash, HPC, MPI, numerical optimization.
- **Reproducibility:** Git, GitHub Actions, pytest, Docker, reproducible experiment pipelines.

AI Safety and Machine Learning Training

BlueDot Impact | Technical AI Safety Project Aug 2026 – present

- Developing *Decided Mid-Thought* as the program research project.

TripleTen | AI & Machine Learning Program (36 weeks) Nov 2025 – Apr 2026

- Completed 15+ end-to-end projects spanning supervised and unsupervised learning, deep learning, NLP, computer vision, and model evaluation.

AI-Powered Document Intelligence Extern | Pfizer-Sponsored Externship Apr 2026 – Jul 2026

- Built document intelligence and retrieval-augmented question-answering pipelines for healthcare documents using native PDF text extraction, LLMs, and RAG architectures.

Selected Publications

- Maleki, K. (2026). "Finite-Horizon Learning-Curve Prediction for Gradient Boosting: Why Sequence Acceleration Fails and Simple Curve Fits Suffice." *Machine Learning*, submitted.
- Maleki, K. (2026). "Quantifying Decision-Boundary Uncertainty in Binary Classifiers: The Ambiguity Range Framework." *Pattern Recognition*, under review. Preprint: https://ssrn.com/abstract=7118401.
- Maleki, K. et al. "Crystal fields, exchange and dipolar interactions, and noncollinear magnons of erbium oxide." *Phys. Rev. B* 112, 064434 (2025).
- Kavand, M., Maleki, K., et al. "A General and Modular Approach to Solid-State Integration of Zero-Dimensional Quantum Systems." *Nano Lett.* 25, 13787–13794 (2025). Related to a patent-pending quantum device.