

Foad Namjoo

foad.namjoo@utah.edu | Website | Google Scholar | GitHub | Hugging Face | LinkedIn | Salt Lake City, UT

Summary

CS Ph.D. student, University of Utah. LLM interpretability researcher (benchmark evaluation, datasets and activation steering) grounded in high-dimensional geometry and statistics; also builds wearable-health sensing systems.

Skills

Languages: Python, C++, JavaScript, Dart

ML & Data: PyTorch, Hugging Face, scikit-learn, TransformerLens, Data Curation, LLM evaluation & benchmarking

Systems: Flutter, BLE, Docker, Node.js, MongoDB, REST APIs, CI/CD, pybind11, CMake

Research Experience & Publications

Graduate Research Assistant | University of Utah

Jan 2023 – Present

- **LLM Benchmark Auditing — Surface-Feature Leakage (TruthfulQA-476) | 2026**

Showed that binary-choice **TruthfulQA** is separable by answer *style* alone: a **question-blind** 6-feature classifier reaches **68.9% accuracy (AUC 0.715)**, above 16 of 18 models on the llm-stats leaderboard (Sept 2026), and several of 13 other benchmarks also leak, worst on HaluEval QA (AUC 0.973). Designed **Audit-Prune** and released the **TruthfulQA-476** dataset, cutting leakage to near chance (**AUC 0.528**) while preserving model rankings (Spearman $\rho = 0.915$).

Paper: *Judging by the Cover: Cleaning LLM Truthfulness Benchmarks to Avoid Surface-Level Feature Leakage* — arXiv 2026, under review at ICLR 2027, [\[paper\]](#) [\[code\]](#) [\[Hugging Face dataset\]](#).

- **Multi-Attribute LLM Steering and Robustness under Dataset Corruption | 2025–Present**

With **Martian AI**, co-authored a study of steering under noisy and adversarial dataset corruption (Llama-3.2, Mistral-7B, OLMo-2), mitigating most of it with robust high-dimensional mean estimation; built contrast datasets and publicly released three [\[Hugging Face\]](#). Co-developed **orthogonal, parameterized multi-attribute steering**, so correlated traits (e.g. power and coordination, cosine 0.87) steer jointly without over- or under-steering.

Papers: *Orthogonal and Parametrized Multi-Attribute Steering of LLMs* — under review at ICLR 2027, [\[code\]](#); *Understanding and Mitigating Dataset Corruption in LLM Steering* — arXiv 2026, under review at ACL ARR, [\[paper\]](#).

- **Efficient Multidimensional Two-Sample Testing (dKS) | 2025**

Multi-dimensional Kolmogorov–Smirnov distance: a unit-invariant metric with a near-linear $O(n \log n)$ approximation; open-source C++/pybind11 library $\sim 76,000\times$ faster than exact computation at 1M points (2D) [\[code\]](#) [\[website\]](#).

Paper: *Efficient and Stable Multi-Dimensional Kolmogorov–Smirnov Distance* — submitted to Foundations of Data Science; authors alphabetical, [\[paper\]](#).

- **MotionPI — Wearable Health Sensing Platform (NCI award R37CA276365) | 2023–Present**

Sole developer of the app, BLE protocols, backend, and offline-first sync layer, within a study team spanning Ohio State’s wearable-hardware group, Utah’s health-study group, and the Huntsman Cancer Institute. Streamed wristband ENMO, accelerometry, and survey signals with schema-validated sync (raw high-rate PPG/IMU archived on-device); sustained **$\sim 7.7\text{M}$ records/day with zero malformed writes**. Built the study team’s participant-monitoring report for weekly compliance follow-ups; typical per-participant wristband records rose from ~ 400 to **800** (study-team reported). Built an ENMO threshold-calibration app for activity-trigger tuning and data-quality review [\[code\]](#); identified an upstream Flutter SDK bug, triaged P2 by Google’s Flutter maintainers [\[issue\]](#).

Paper: *Designing a Secure and Resilient Distributed Smartphone Participant Data Collection System* — SmartSP 2025 (first author), Springer LNICST, [\[paper\]](#).

- **Anomaly Detection — Region-Aggregated Spatial Scan Statistics | 2023–2026**

Replaced SaTScan-style centroids with 20–50 sampled points per region: higher detection power, and all 3,108 contiguous U.S. counties scanned in 0.33s ($\sim 3,000\times$ faster than FlexScan, scan step only) [\[code\]](#).

Paper: *ACM SIGSPATIAL 2026* (first author; accepted as a full paper, 58/248), [\[paper\]](#).

Graduate Research Assistant | University of Tehran

2019 – 2022

- **Computer Vision — High-Dimensional Spectral–Spatial Change Detection | 2022**

Dual-stream 3D/2D CNN with SE attention; accuracy $> 96\%$, $\kappa > 0.9$, and fewer false positives vs. baselines.

Paper: *Hyperspectral Change Detection (Double-Stream CNN and Attention Module)*, 2024, [\[paper\]](#).

Education

Ph.D., Computer Science | University of Utah | Jan 2023 – Expected May 2028 | TA, [Data Mining](#) (Fall 2026)

Advisor: Prof. Jeff Phillips | GPA: 3.84/4 | Artifact Evaluation Committee, [ALENEX 2027](#) (SIAM algorithms symposium)

M.Sc., Algorithms and Computation (Computer Science) | University of Tehran | 2019 – 2022

Thesis: Graph-theoretic modeling of bushfire propagation | Entry exam rank: 29/20,000 (top 0.15%)