(607) 391-1145
kk983@cornell.edu
github.com/kimzemian

Kimia Kazemian

Gates Hall, Cornell University
Ithaca, NY, 14850
kimzemian.github.io

Objective and Summary

Machine learning researcher bridging rigorous theoretical frameworks with large-scale model optimization. Research integrates the mechanistic interpretability of transformers, large-scale post-training for LLMs, and building scalable predictive models for multivariate time-series data and control-affine systems.

Education

PhD in Computer Science **Cornell University** **Aug 2022 - now**

- Advised by Sarah Dean and John Thickstun, Kilian Weinberger's Lab member.
- Relevant Courses: Deep Probabilistic and Generative Models, Machine Learning, Reinforcement Learning, Analysis of Algorithms, Advanced Programming Languages (*OCaml*), Applied High Performance and Parallel Computing(C++), Differential Manifolds.
- Other coursework: Equity investment research, Entrepreneurship for Scientists and Engineers.

PhD in Mathematics **Cornell University** **Aug 2019 - May 2022 (Transferred to CS)**

- Advised by Ravi Ramakrishna and Birgit Speh, Research in Automorphic Forms.
- Relevant Courses: Lie Groups and Lie Algebras, Algebraic Topology, Commutative Algebra, Algebra I & II, Complex Analysis.

B.Sc. in Mathematics **Sharif University of Technology** **Aug 2015 - May 2019**

- Relevant Courses: Game Theory, Algebraic Graph Theory, Probability, Numerical Analysis (*Matlab*), Real Analysis.

Work Experience

Machine Learning Intern **Microsoft - Turing team** **June 2024 - September 2024**

Post-Training (Expert Iteration) for LLM Reasoning:

- Engineered an automated Expert Iteration pipeline for LLM reasoning, utilizing vLLM to overgenerate tens of thousands of Chain-of-Thought (CoT) trajectories and implementing masked prompt supervised fine-tuning (SFT) via Hugging Face (PyTorch, DeepSpeed).
- Designed a robust rejection sampling and verification framework utilizing regex-based extraction and mathematical equivalence checking, successfully improving model format adherence from 34% to 98% through systematic prompt tuning.
- Ensured dataset quality by clustering sentence transformer embeddings (via PCA and t-SNE) to measure CoT diversity, and increased reasoning accuracy on the Hendrycks/MATH dataset from 36% to 44% (Llama-3.1-8B) and 40% to 43% (Phi-3-small) by training exclusively on self-verified, in-distribution CoT trajectories.

Siegel PiTech Impact Fellow **Oxfam** **May 2023 - August 2023**

Unsupervised Topic Modeling & NLP at Scale:

- Engineered a high-throughput streaming pipeline to decompress, filter, and structure 2TB of historical Reddit archives (spanning 10 years) into Hugging Face datasets.
- Designed an unsupervised topic modeling architecture utilizing SentenceTransformers (`all-MiniLM-L6-v2`) for contextual text embeddings, UMAP for dimensionality reduction, and HDBSCAN for hierarchical density-based clustering (*Pytorch*, *Hugging Face*).
- Extracted and quantified highly granular thematic clusters regarding Amazon and Walmart warehouse labor conditions (e.g., unionization, automation, wages) to directly inform executive-level worker advocacy initiatives.

Research Experience

Cornell University **ML Core; Thickstun Lab; Dean Lab** **August 2022 - now**

Spatial Induction Heads: In-Context Learning of Multidimensional Cellular Automata:

- Formalized "spatial induction heads", two-layer transformer circuits that infer unseen multi-dimensional cellular automata rules in context by reconstructing spatial neighborhoods and retrieving matching transitions.
- Proved that the positional dimension required for spatial routing is independent of grid size and trajectory length; constructed attention mechanisms that approximate Bayesian posterior prediction with explicit approximation guarantees.
- Validated generalization empirically on unseen 1D and 2D rules, achieving $\geq 99.99\%$ deterministic rollout cell accuracy and < 0.005 nats KL from Bayesian optimal predictions; supported the proposed circuit through attention ablations and linear probes.

Cross-Platform Time Series Forecasting Benchmark Dataset:

- Established Lead-Lag Forecasting (LLF) as a novel time-series paradigm by formalizing cross-channel prediction, where early interaction signals (lead) forecast delayed outcome metrics (lag) across unseen entities.

- Curated and released two large-scale benchmark datasets comprising 2.3M arXiv papers (accesses → citations) and 3M GitHub repositories (pushes/stars → forks), providing the first public datasets that jointly test cross-channel and cross-series generalization at scale.
- Trained and benchmarked deep learning models (custom Transformers, Time-MoE) alongside classical methods (Logistic Regression, KNN, MLP) to validate predictive signals and establish baseline performance across classification and regression tasks.

Fast representation learning in affine state-action space:
- Proposed the Affine Dot Product (ADP) and Affine Dense (AD) nonlinear feature representations to capture control-affine structures, formalizing theoretical guarantees for these novel kernels.
- Illustrated the utility via a theoretical case study of data-driven optimization-based control, and then empirical simulation stabilizing a double pendulum with unknown dynamics online from data (*Python*, *Numpy*, *CVXPY*, *Seaborn*).
- Demonstrated empirically that these random feature (RF) approximations reduced time and space complexity from cubic and quadratic to linear without sacrificing predictive accuracy.

Cornell University **Ramakrishna and Speh Groups** **2019 - 2022**

Research in Langlands program:
- *Galois theory*: Computationally verified Buhler's thesis on modularity of Icosehedral Galois representations.
- *Lie theory*: Research on Cohomological Parabolic Induction. Constructed infinite dimensional representations of real reductive Lie groups, and automorphic representations of GL(n).

Selected Publications

Kimia Kazemian*, Menghan Xu*, John Thickstun, Sarah Dean. Spatial Induction Heads: In-Context Learning of Multidimensional Cellular Automata. *In Submission*

Kimia Kazemian*, Zhenzhen Liu*, Yangfanyu Yang, Katie Luo, Shuhan Gu, Audrey Du, Xinyu Yang, Jack Jansons, Kilian Weinberger, John Thickstun, Yian Yin, Sarah Dean. Benchmark Dataset for Lead-Lag Forecasting in Social Platforms. *KDD 2026 on Benchmark and Datasets*

Kimia Kazemian, Yahya Sattar, Sarah Dean. Random Features Approximation for Control-Affine Systems. *L4DC 2024*

Kimia Kazemian, Sarah Dean. Random Features Approximation for Fast Data-Driven Control. *NeurIPS 2022 on Gaussian Processes*

Saeed Akbari, Afsaneh Ghafari, **Kimia Kazemian**, Mina Nahvi. Some Criteria for a Signed Graph to Have Full Rank. *2022 Discrete Mathematics Journal*

Skills, Awards & Leadership

Programming & Frameworks: Python, C++, OCaml, MATLAB, Java, PyTorch, Huggingface, Scikit-Learn

Tools: DeepSpeed, Wandb, NumPy, CVXPY, SciPy, Pandas, Seaborn

Awards and Test Scores:
- Recipient of the Cornell Graduate School Recruiting Fellowship as a mathematics PhD student.
- Finalist, Iranian National Olympiads in Mathematics (2011–2013) and Informatics (2013).
- GRE Subject Mathematics: 860/900 (88th percentile)

Leadership: Editor-in-Chief, Sharif Mathematical Journal (2017–2019). Curated a team of 30 people to revitalize the discontinued journal and orchestrated the production of issues 9 and 10.

Mentorship & Teaching:
- Mentored and designed research projects for 5 master's and undergraduate students in CS.
- Teaching Assistant for 6 Math and CS courses; Principles of Large-Scale Machine Learning Systems, Introduction to Combinatorics, Differential Equations and Dynamical Systems, Matrix Groups, Multivariable Calculus (class of 500+ students), Analysis of Algorithms, Introduction to Python (200+ students).