(650) 440-8143
kk983@cornell.edu
github.com/kimzemian

Kimia Kazemian

Gates Hall, Cornell University
Ithaca, NY, 14850
kimia.boo

Objective and Summary

- PhD candidate in AI, specializing in optimization, foundation models, and novel architectures for time series analysis and reasoning.

Education

PhD in Computer Science **Cornell University** **Aug 2022 - May 2026 (Expected)**

- Research in AI, optimization, Reinforcement Learning, NLP.
- Relevant Courses: 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)**

- Research in Number Theory and Representation Theory.
- 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*).

Work Experience

Data Scientist Intern **Microsoft - Turing team** **June 2024 - September 2024**

Expert Iteration(RL) to improve reasoning in LLMs:
- Built a pipeline to iteratively overgenerate diverse chain of thought(CoT) data and supervise-fine-tune(SFT) an LLM on rejection sampled CoT, to self-improve, and converge to golden policy (*Python*, *Huggingface*, *DeepSpeed*).
- Introduced a new metric for measuring model improvement, and demonstrated that perplexity is not a reliable loss function for evaluation and training in this domain (*Wandb*).
- Improved reasoning abilities of multiple Llama 3.1 (8B and 70B) and Phi3 models by 2-4%.

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

Analyzing sentiment of social media at scale:
- Developed a robust data processing pipeline to analyze reddit discussions - handles 2T of zipped reddit data smoothly.
- Extracted key sentiments and insights using topic modeling, and showcased the effectiveness of encoder models in understanding large-scale social media conversations (*Python*, *Huggingface*).
- Delivered a comprehensive analysis that informed strategic decisions on worker advocacy initiatives.

Research Experience

Cornell University **ML Core & Dean Lab** **August 2022 - now**

Cross-Platform Time Series Forecasting Framework:
- Formalized Lead-Lag Forecasting (LLF) as a novel time series prediction paradigm for cross-channel forecasting with substantial temporal gaps between early signals and long-term outcomes. Addressing the critical gap in existing forecasting benchmarks that focus on short-horizon, single-channel predictions.
- Created large-scale multi-domain benchmark datasets including arXiv (2.4M papers) and GitHub (3M repositories), orders of magnitude larger than existing time series benchmarks and enable systematic study of real-world impact prediction.
- Established comprehensive evaluation framework and baselines through empirical analysis demonstrating lead-lag dynamics and benchmarking multiple machine learning approaches.

Fast representation learning in affine state-action space:
- Proposed two novel classes of nonlinear feature representations which capture action affine structure, and formalized representation guarantees via introduction of a novel affine dense kernel.
- 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*).
- My model reduced both time and space complexity from cubic and quadratic to linear without sacrificing 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. Worked on infinite dimensional representations of real reductive Lie groups, and automorphic representations of GL(n).

Sharif University of Technology **Akbari Group** **2018**

Signed graphs characterization:
- Identified necessary and sufficient criteria for signed graphs to have full rank.

Publications and Papers

- **K. Kazemian**, Z. Liu, et al. Lead-Lag Forecasting in social platforms. *(in submission) NeurIPS 2025*
- **K. Kazemian**, Y.Sattar, S.Dean. Random features approximation for control-affine systems. *L4DC 2024*
- A. Kovacs, J.Chee, **K. Kazemian**, S. Dean. Datasets for Navigating Sensitive Topics in Preference Data and Recommendations. *NeurIPS 2024 Workshop on Safe Generaative AI*
- **K. Kazemian**, Sarah Dean. Random Features Approximation for Fast Data-Driven Control. *Neurips Workshop on Gaussian Processes, Dec 2022*
- S. Akbari, A. Ghafari, **K. Kazemian**, M. Nahvi. Some Criteria for a Signed Graph to Have Full Rank. *Discrete Mathematics, Volume 343, Issue 8. August 2022*
- **K. Kazemian**. Kleene Algebra with Domain and Kleene Modules. *Expository paper, Dec 2019*

Misc. Projects

- *Navigating Sensitivity*: Constructed a database of movie ratings by incorporating content warnings. Assessed the correlation of ratings and content sensitivity.(*Python*, *Numpy*, *Scikit-Learn*, *SciPy*, *Pandas*). 2024
- *Calibration in Fairness*: Theoretically proved possibility of achieving fairness measures like calibration and predictive parity by appropriately thresholding continuous calibrated scores. 2023
- *Namegen*: Implemented recurrent neural network (RNN) model to generate unique names for babies/startups (*Python*, *Pytorch*). 2022
- *Kleene Modules*: Detailed the algebraic exposition of Kleene algebras with domains and Kleene modules for action-state modeling in program verification. 2020
- *CarOBs*: Implemented a game where cars orbit obstacles to avoid them (*Java*). 2015

Leadership

Teaching Assistant **Cornell University: Math and CS** **Aug 2019 - Aug 2022**

- Held recitations, labs, office hours, and grading for courses: Introduction to Combinatorics, Differential Equations and Dynamical Systems, Matrix Groups, Multivariable Calculus **(class of 500+ students)**, Analysis of Algorithms, Introduction to Python **(200+ students)**.

Editor-in-Chief **Sharif Mathematical Journal** **Aug 2017 - Aug 2019**

Revitalized discontinued Sharif Mathematical Journal, curated **a team of 30 people** and Orchestrated every aspect of production of second series, issues 9 and 10.

Awards and Honors

- 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)
- GRE General: Verbal 159/170 (83rd percentile), Quant 170/170 (96th percentile)

Skills

- **Programming Languages**: Python, C++, OCaml, MATLAB, Java
- **Frameworks**: PyTorch, Huggingface, Scikit-Learn
- **Tools**: DeepSpeed, Wandb, NumPy, CVXPY, SciPy, Pandas, Seaborn