Sankaran Vaidyanathan

Amherst, MA sankaranv@cs.umass.edu sankaranv.com sankaranv8 sankaranv

Goal

Developing principled tools grounded in causal reasoning for explaining and evaluating complex AI systems, including large language models and reinforcement learning agents.

Education

University of Massachusetts Amherst *Sept 2021–Dec 2026*
Ph.D. in Computer Science

University of Massachusetts Amherst *Sept 2019–May 2024*
M.S. in Computer Science

Anna University *Aug 2013–May 2017*
B.E. in Electrical and Electronics Engineering

Experience

Research Assistant *Amherst, MA*
Knowledge Discovery Lab, University of Massachusetts Amherst *May 2020–present*

Project Associate *Chennai, India*
Robert Bosch Center for Data Science and AI *July 2017–June 2019*
Indian Institute of Technology Madras

Publications

[1] **Adaptive Circuit Behavior and Generalization in Mechanistic Interpretability**
Jatin Nainani, **Sankaran Vaidyanathan***, AJ Yeung, Kartik Gupta, David Jensen*
arXiv:2411.16105 (under review), 2024

[2] **Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges**
Aman Singh Thakur, Kartik Choudhary*, Venkat Srinik Ramayapally*, **Sankaran Vaidyanathan**, Dieuwke Hupkes*
arXiv:2406.12624 (under review), 2024

[3] **Automated Discovery of Functional Actual Causes in Complex Environments**
Caleb Chuck, **Sankaran Vaidyanathan***, Stephen Giguere, Amy Zhang, David Jensen, Scott Niekum*
arXiv:2404.10883 (under review), 2024

[4] **Data-driven Learning of Chaotic Dynamical Systems using Discrete-Temporal Sobolev Networks**
*Connor Kennedy, Trace Crowdis, Haoran Hu, **Sankaran Vaidyanathan**, Hong-Kun Zhang*
Neural Networks, Volume 173, May 2024, 106152

[5] **Hypergraph Clustering by Iteratively Reweighted Modularity Maximization**
*Tarun Kumar, **Sankaran Vaidyanathan**, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, Balaraman Ravindran*
Complex Networks and Their Applications VIII, 2019

[6] **A New Measure of Modularity in Hypergraphs: Theoretical Insights and Implications for Effective Clustering**
Tarun Kumar, **Sankaran Vaidyanathan***, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, Balaraman Ravindran*
Applied Network Science 5 (1), 52

Collaborative Research Projects

Sequential Circuit Discovery in LLMs *Feb 2025–May 2025*
AI Safety Camp

- Investigating multi-token mechanisms in LLMs that explain how specific components and features in the model influence not just the next token generated, but also tokens further ahead in the sentence generated.

- Extending causal mediation analysis with time-varying treatments and mediators to identify how specific model components and features affect token sequences over multiple prediction steps.

Quantitative LLM Judges *Feb 2025–May 2025*
Adobe Research

- Analyzing the limitations of the LLM-as-a-Judge paradigm, a popular approach for evaluating the performance of an LLM by having another LLM assess its outputs and generate a score.
- Designing judge models that are explicitly trained to produce calibrated, quantitative scores with uncertainty estimates and worst-case performance guarantees, while still leveraging pretrained LLMs.

Evaluating Alignment and Vulnerabilities in LLMs-as-Judges *Feb 2024–Dec 2024*
Meta

- Evaluated the performance of 9 exam-taker models solving a multiple-choice question answering test using 13 different LLM judge models, to study the performance and behavior of the judges.
- Discovered significant performance gaps between the highest-performing judge models and human evaluators, while demonstrating competitive accuracy from smaller models and simple lexical metrics.
- Identified Scott's π as a more reliable metric for evaluating judges, and revealed further issues with judge models such as leniency, sensitivity to prompt quality, and struggles with underspecified answers.

Analysis and Prediction of Cognitive Load During Cardiac Surgery *May 2023–May 2024*
National Institute of Health and Harvard Medical School

- Modeled and visualized various measures of heart rate variability, to predict cognitive load and stress among members of a surgical team while performing cardiac surgery.
- Developed Transformer and LSTM models for time-series prediction of heart-rate variability, and an MCMC based imputation scheme to fill in missing data from faulty heart rate monitors.
- Leveraged Explainable AI (XAI) techniques including SHAP, feature ablation, and permutation importance, to identify key features that the models prioritized when predicting cognitive load.

Competence-Aware Machine Learning *May 2020–Aug 2022*
DARPA Competence-Aware Machine Learning Program

- Determined the causes of failure for a pre-trained reinforcement learning agent navigating in the AirSim driving environment, by estimating causal effects of various environmental conditions on mission failure.
- Learned causal models that estimated the agent's competence, or probability of mission success, for a route with pre-specified environmental conditions.
- Developed a system that allowed a human operator to specify environmental conditions for a new episode prior to deployment, and returned an upper and lower bound on the agent's estimated competence.

Technical Skills

Languages: Python, R, C++

Frameworks: PyTorch, Pyro-PPL, Box2D, TransformerLens, SAELens

Tools: Git, Linux, Figma

Teaching Experience

- Data Structures *UMass Amherst, Spring 2025*
- Decarbonization and Data Science *UMass Amherst, Fall 2024*
- Probabilistic Graphical Models *UMass Amherst, Spring 2023*
- Artificial Intelligence *UMass Amherst, Fall 2022*
- Probability Theory (Guest Lecture: Markov Chain Monte Carlo) *UMass Amherst, Fall 2021*
- Introduction to Machine Learning *IIT Madras, Spring 2019*

Service and Outreach

- Data Science Industry Mentor: Adobe, Meta *2024–2025*
- Mentor, UMass PhD Applicant Support Program *2021–2024*
- Co-organizer, UMass Machine Learning and Friends Lunch *2019–2020, 2023–2024*
- M.S. Graduate Representative, UMass College of Information and Computer Sciences *2020*
- Social and New Student Committee, UMass College of Information and Computer Sciences *2023–2025*
- Volunteer Pen-Pal, Letters to a Pre-Scientist *2024–2025*

Relevant Graduate Coursework

Bayesian Statistics, Intro to Causal Inference, Probabilistic Graphical Models, Machine Learning, Reinforcement Learning, Artificial Intelligence, Advanced Natural Language Processing, Neural Networks: A Modern Introduction, Optimization in Computer Science, Math Statistics, Research Methods in Empirical CS, Probability Theory, Distributed and Operating Systems, Quantum Information Systems, Fixing Social Media