Last updated in April 2025

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

- Analyzed limitations of the LLM-as-a-Judge paradigm, where the performance of an LLM is evaluated by using another LLM to review and score its outputs.
- Developed generalized linear models on pretrained LLM embeddings to produce calibrated scores with uncertainty estimates and worst-case performance guarantees.

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

- Evaluated 13 LLM judge models on scoring model outputs from a multiple-choice QA benchmark, identifying Scott's π as a more reliable metric for evaluating judge models.
- Identified failure cases such as prompt sensitivity and revealed misalignment with human judgments in top-performing LLM judges, with competitive performance from smaller models and simple lexical metrics.

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

- Predicted cognitive load and stress among surgical teams during cardiac surgery using time-series models of heart rate variability (Transformer, LSTM) with MCMC-based imputation for missing sensor data.
- Applied Explainable AI techniques (SHAP, feature ablation, permutation importance) to identify key heart rate variability features driving model predictions, and validated findings against clinical expert knowledge.

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.
- Built causal models for estimating the agent's competence (probability of mission success) with confidence intervals, given the environmental conditions for an upcoming route.

Technical Skills

Languages: Python, R, C++

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

Tools: Git, Linux, Figma

Teaching Experience

University of Massachusetts Amherst: *Data Structures*, *Decarbonization and Data Science*, *Probabilistic Graphical Models*, *Artificial Intelligence*, *Probability Theory — Guest Lecture on Markov Chain Monte Carlo*

Indian Institute of Technology Madras: *Introduction to Machine Learning*

Service and Outreach

- Mentor, UMass Data Science Industry Independent Study: Adobe, Meta *2024–2025*
- Mentor, UMass Ph.D. 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*
- Volunteer Pen-Pal, Letters to a Pre-Scientist *2024–2025*

Relevant Graduate Coursework

Bayesian Statistics, Causal Inference, Probabilistic Graphical Models, Reinforcement Learning, Machine 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