Sankaran Vaidyanathan

✉ sankaranv@cs.umass.edu
sankaranv.com
in linkedin.com/in/sankaranv8/

Education

Sep '21–Dec '26 *(expected)* **Ph.D., Computer Science**, University of Massachusetts Amherst

Sep '19–May '24 **M.S., Computer Science**, University of Massachusetts Amherst *GPA: 3.97/4.0*

Aug '13–May '17 **B.E., Electrical and Electronics Engineering**, Anna University *GPA: 8.45/10*

Research Experience

Jan '20–present **Research Assistant**, Knowledge Discovery Lab, University of Massachusetts Amherst
Advisor: David Jensen

Jul '17–Jun '19 **Project Associate**, RISE-IIL Lab, Indian Institute of Technology Madras
Advisor: Balaraman Ravindran

Publications

(* denotes equal contribution)

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

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

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

Neural Networks vol. 173 (2024) **Data-driven Learning of Chaotic Dynamical Systems using Discrete-Temporal Sobolev Networks**
Connor Kennedy, Trace Crowdis, Haoran Hu, **Sankaran Vaidyanathan**, Hong-Kun Zhang

Applied Network Science, 5, 52 (2020) **Hypergraph Clustering by Iteratively Reweighted Modularity Maximization**
Tarun Kumar, **Sankaran Vaidyanathan**, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, Balaraman Ravindran

Complex Networks (2019) **A New Measure of Modularity in Hypergraphs: Theoretical Insights and Implications for Effective Clustering**
Tarun Kumar*, **Sankaran Vaidyanathan***, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, Balaraman Ravindran

Technical Skills

- **Programming Languages**: Python, C++, R
- **Frameworks**: PyTorch, PyMC, HuggingFace Transformers, Box2D, TransformerLens
- **Tools**: Figma, Arduino, Kubernetes, Git, Linux, Adobe After Effects, Inkscape, Ableton

Selected Projects

Mar '24–present **Circuit Generalization and Component Reuse in LLM Mechanistic Interpretability**
Supervised by David Jensen

- Identifying circuits – small sets of components within large neural networks that are responsible for solving specific tasks – in the GPT-2 small and Gemma2-2B large language models (LLMs).
- Evaluated circuit behavior on prompt variants designed to challenge underlying assumptions in the task.
- Demonstrated that circuits in GPT-2 small generalize to prompt variants by reusing the majority of components and mechanisms from the original circuit.
- Explaining the underlying causes of circuit generalization using Sparse Autoencoder (SAE) features.

Jan '23–present **Automated Discovery of Actual Causes**
Supervised by David Jensen and Scott Niekum

- Extending the theory of actual causality, a framework based on legal reasoning and human judgment that formalizes explanations, blame, and harm in AI systems, to continuous and high-dimensional domains.
- Designed a tractable inference approach for identifying actual causes by detecting context-specific independencies, which indicate if particular observed events did not affect the agent and can be ignored.
- Detected and explained the most relevant interactions between an agent and objects in a set of physical reasoning and reinforcement learning (RL) domains, by learning to identify actual causes.

Feb '24–Dec '24 **Evaluating Alignment and Vulnerabilities in LLMs-as-Judges**
Supervised by Dieuwke Hupkes (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 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.

Mar '24–present **Identifying Causes of Patent Claim Rejection using LLMs**

- Using LLM-based methods to evaluate patent claims for potential causes of rejection and map them to relevant parts of the Manual of Patent Examining Procedure, with a focus on *indefiniteness* issues.
- Evaluated alignment between human annotators, a patent lawyer, and fine-tuned BERT in identifying lack of antecedent basis, a common patent claim error.

May '23–May '24 **Analysis and Prediction of Cognitive Load Among Teams During Cardiac Surgery**
In collaboration with the 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 techniques including SHAP, feature ablation, and permutation importance, to identify key features that the models prioritized when predicting cognitive load.

May '20–Aug '22 **Competence-Aware Machine Learning**
Joint work with David Jensen (UMass Amherst), Joydeep Biswas (UT Austin), and Charles River Analytics

- 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.

Teaching Experience

Sep-Dec '24 **COMPSCI 590X: Decarbonization and Data Science**, University of Massachusetts Amherst

Feb-May '23 **COMPSCI 688: Probabilistic Graphical Models**, University of Massachusetts Amherst

Sep-Dec '22 **COMPSCI 383: Artificial Intelligence**, University of Massachusetts Amherst

Dec '21 **MATH 605: Probability Theory**, University of Massachusetts Amherst

- Gave a guest lecture on sampling methods, Markov Chain Monte Carlo, and Hamiltonian Monte Carlo.

Jan–May '19 **Introduction to Machine Learning**, Indian Institute of Technology Madras

Coursework

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