Manqing Liu

Boston, MA, 02116
manqingliu@g.harvard.edu
manqingliu.github.io

Summary

My current research develops methods to detect reasoning pathologies in large language models, including post-hoc rationalization, encoded reasoning, and internalized reasoning. I construct model organisms exhibiting each pathology and build end-to-end evaluation pipelines for training-time monitoring and scalable oversight. I've also worked on causal machine learning research in collaboration with Dr. Andrew Beam and Dr. James Robins. I am seeking Research Scientist or Research Engineer roles in model evaluation, post-training, alignment, and safety.

Education

2021–Present **Ph.D. in Causal Machine Learning**, *Harvard University*

2022–Present **Secondary field in Computer Science and Engineering**, *Harvard University*

2021–Present **M.Sc in Biostatistics**, *Harvard University*

2017-2020 **Post-Baccalaureate Studies in Maths/Statistics**, *University of Pennsylvania*

2014–2016 **MHS in Epidemiology**, *Johns Hopkins University*

Relevant Coursework

Causality	Advanced Epidemiologic Methods, Models for Causal Inference
Maths and Statistics	(MIT) Matrix Methods in Data Analysis & Signal Processing, (MIT) Introduction to Functional Analysis, Probability, Statistical Inference, Advanced Regression and Statistical Learning, Bayesian Inference
Computer Science	Systems Development for Computational Science, High Performance Computing for Science and Engineering, Stochastic Methods for Data Analysis, Inference and Optimization
Machine Learning	AI safety, (MIT) Machine Learning, (MIT) Quantitative Methods for NLP, Deep Learning for Biomedical Data, Geometric Methods for Machine Learning, Algorithms for Data Science

Research Experience

July 2025 – Sept 2025 **Diagnosing Pathological Chain-of-Thought in Reasoning Models**
Completed MARS fellowship developing novel metrics to detect CoT pathologies in LLMs. Developed and implemented comprehensive evaluation metrics to identify and monitor pathologies in Chain-of-Thought (CoT) reasoning across large language models, including post-hoc, internalized, and encoded reasoning patterns. Collaborated on fine-tuning open-weight LLM models to elicit internalized and encoded reasoning capabilities in model organisms.

Oct 2024 – Jan 2025 **Doubly Robust MCTS for LLM reasoning**
Integrated doubly robust estimator into Monte Carlo Tree Search (MCTS), enabling large language models to perform complex, multi-step reasoning and planning with higher accuracy and sample efficiency in real-world scenarios.

June 2023 – Dec 2024 **DAG aware Transformer**
Engineered a noval DAG-aware transformer model to precisely estimate causal effects, addressing foundational challenges in unifying causal effect estimation under various scenarios.

Publications

2025 **Diagnosing Pathological Chain-of-Thought in Reasoning Models**, *Liu M.*, Williams-King D., Caspary I. et al. Under review at ICLR, 2026. [PDF]

2025 **Doubly Robust Monte Carlo Tree Search**, *Liu M.*, Beam A. Under reivew at ICLR, 2026. [PDF]

2024 **DAG-Aware Transformer for Causal Effect Estimation**, *Liu M.*, Bellamy D., Beam A. Causal Representation Learning workshop at NeurIPS 2024. Available at: arXiv:2410.10044

2022 **Development of Machine Learning Algorithms Incorporating Electronic Health Record Data, Patient-Reported Outcomes, or Both to Predict Mortality for Outpatients with Cancer**, Parikh R.B., Hasler J.S., Zhang Y., *Liu M.*, Chivers C., et al., *JCO Clinical Cancer Informatics*, 6.

2021 **Trajectories of Mortality Risk Among Patients with Cancer and Associated End-of-Life Utilization**, Parikh R.B., *Liu M.*, Li E., Li R., Chen J., *npj Digital Medicine*, 4(1):104.

2020 **Validation of a Machine Learning Algorithm to Predict 180-Day Mortality for Outpatients with Cancer**, Manz C.R., Chen J., *Liu M.*, Chivers C., Regli S.H., et al., *JAMA Oncology*, 6(11):1723-1730.

Professional Experience

Summer 2025 **Research Fellow**, *Cambridge AI Safety Hub*, Cambridge, UK, MARS Fellowship Program
Completed competitive 3-month research fellowship focused on AI safety and alignment. Conducted independent research on detecting pathological reasoning behaviors in large language models under mentorship from University of Cambridge with collaborators from UCL and Mila. Developed novel evaluation methodologies for Chain-of-Thought pathologies with implications for AI system reliability and trustworthiness.

Summer 2024 **Technical AI safety Fellowship**, *AI safety student team*, Cambridge, MA, USA
Attended a 8-week reading group on AI safety, covering topics like neural network interpretability, learning from human feedback, goal misgeneralization in reinforcement learning agents, and eliciting latent knowledge.

Skills

Programming Languages: Python, C++, R, SAS, STATA

Libraries and Frameworks	PyTorch, Tensorflow, Pandas, NumPy
Others	Causal Inference, Machine Learning, Deep Learning