RMTA CHINA 中国现场统计研究会 随机矩阵理论与应用分会

清華大學 统计与数据科学系
Department of Statistics and Data Science, Tsinghua University

大模型数理统计基础研讨会

会议手册

主办单位： 中国现场统计研究会随机矩阵理论与应用分会、清华大学统计与数据科学系、香港中文大学（深圳）人工智能学院

承办单位： 香港中文大学（深圳）　　**协办单位：** 黄大年茶思屋科技网站、华为拉格朗日数学与计算中心

2025 7/23-25　吉林·长白山

RMTA CHINA 中国现场统计研究会 随机矩阵理论与应用分会

清華大學 统计与数据科学系
Department of Statistics and Data Science, Tsinghua University

会议指南

会议时间及地点

- 报道时间：2025 年 7 月 23 日（星期三）14:00-20:00 及 7 月 24 日（星期四）会议期间
- 报道地点：长白山柏悦酒店，长白山国际度假村
- 报道地址：吉林省白山市抚松县长白山国际度假区南区冠冕街 236 号
- 会议时间：2025 年 7 月 24 日 -7 月 25 日

就餐安排

日 期	事 项
7 月 23 日	自助晚餐（18:00-20:00）
7 月 24 日	自助午餐（12:00-14:00）
	晚宴（18:00-20:30）
7 月 25 日	自助午餐（12:00-14:00）
	自助晚餐（18:00-20:00）

会议联系人

廖振宇 13164151928

会议日程

2025 年 7 月 23 日（星期三）长白山柏悦酒店				
	14:00-21:00	研讨会报道	酒店一楼大厅	
	18:00-20:00	自助晚餐		
2025 年 7 月 24 日（星期四）长白山柏悦酒店 宴会厅				
开幕式	09:00-09:05	研讨会介绍		
	时间	演讲人	单位	报告题目
特邀主旨报告	09:05-10:05	范剑青	Princeton University	Measuring Misinformation in Financial Markets
	10:05-10:35	茶歇		
特邀主旨报告	10:35-11:35	陈松蹊	清华大学	Versatile Differentially Private Statistical Learning for general loss functions
	11:35-11:55	鸿蒙十大数学难题发布（发布人：高杰星、黄磊）		
	11:55-14:00	午餐		
主题学术报告	14:00-14:35	李建	清华大学	Understanding the Behaviors of Large Language Models: A perspective based on Komolgorov Complexity and Shannon's Information Theory
	14:35-15:10	苏炜杰	University of Pennsylvania	Aligning Large Language Models Through the Lens of Social Choice Theory: Impossibility and Possibility
	15:10-15:45	雷琦	NYU Courant	Discrepancies are Virtue: Weak-to-Strong Generalization through Lens of Intrinsic Dimension
	15:45-16:10	茶歇		
主题学术报告	16:10-16:45	张新雨	中国科学院	Towards Optimal Neural Networks: the Role of Sample Splitting in Hyperparameter Selection
	16:45-17:20	钟威	厦门大学	A Gaussian Algorithmic Stability Framework for Post-Selection Inference
	17:20-17:55	林伟	北京大学	Demystifying Neural Networks: Overparametrization, Generalization, and Feature Learning
	17:55-18:30	魏玉婷	University of Pennsylvania	To Intrinsic Dimension and Beyond: Efficient Sampling in Diffusion Models
	18:30-20:30	晚宴		

2025 年 7 月 25 日（星期五）长白山柏悦酒店 宴会厅				
	时间	演讲人	单位	报告题目
主题学术报告	09:00-09:35	郑术蓉	东北师范大学	Difference between Large Statistical Model and Medium Statistical Model
	09:35-10:10	李根	CUHK	Faster Convergence and Acceleration for Diffusion-Based Generative Models
	10:10-10:30	茶歇		
主题学术报告	10:30-11:05	袁洋	清华大学	Structural AI: From Pretraining Abstractions to Formal Reasoning Structures
	11:05-11:40	金驰	Princeton University	Goedel-Prover: A Frontier Model for Open-Source Automated Theorem Proving
	11:40-12:15	邹荻凡	The University of Hong Kong	On the Mechanism Interpretability of LLM for Fine-tuning and Reasoning
	12:15-14:00	午餐		
主题学术报告	14:00-14:35	孙强	University of Toronto	Making AI trustworthy
	14:35-15:10	严晓东	西安交通大学	Statistical Detection for Silent Data Corruption during Large-scale Model Training
	15:10-17:30	自由讨论		
	17:30-19:30	晚餐		

报告摘要

Jianqing Fan

Bio: Jianqing Fan is Frederick L. Moore Professor at Princeton University, where he directs labs in financial econometrics and statistics. He earned his PhD from UC Berkeley and previously held academic positions at UNC–Chapel Hill, UCLA, and the Chinese University of Hong Kong. A former president of both the Institute of Mathematical Statistics and the International Chinese Statistical Association, he has served as editor for leading journals, including JASA, Annals of Statistics, Probability Theory and Related Fields, Journal of Business and Economics Statistics and Journal of Econometrics. His research spans statistics, machine learning, financial economics, and computational biology, with over 300 highly cited publications and four books. Recognized with numerous honors—including the COPSS Presidents' Award, Guggenheim Fellowship, Guy Medal, Noether Distinguished Scholar Award, and Le Cam Award and Lecture —he is a fellow of multiple scientific societies and an elected member of Academia Sinica and the Royal Academy of Belgium. His current interests include high-dimensional statistics and AI.

Title: Measuring Misinformation in Financial Markets

Abstract: We propose a framework for measuring firm-level misinformation. By leveraging advanced machine learning and AI technologies, we transform and categorize unstructured text into comparable information, extract "reliability-weighted consensus" from each set of comparable information, and quantify the degree of misinformation based on divergence from the "consensus" . Applying our framework to analyze 254.8 million textual materials, we validate its effectiveness in quantifying misinformation. We find that firms with weaker balance sheets and poorer governance structures exhibit higher misinformation, and misinformation spikes during major corporate events. We also demonstrate that misinformation significantly impacts investors' attention, trading volumes, stock returns, and risks.

(Joint work with Qingfu Liu, Yang Song, Zilu Wang)

Song Xi Chen

Bio: Dr. Song Xi Chen is a University Chair Professor at Statistics and Data Science of Tsinghua University, Peking, China. He received his B.Sc. and M.Sc. in Mathematics from Beijing Normal University in 1983 and 1988, respectively, and his Ph.D. in Statistics from Australian National University in 1993. His primary research interests include inference for high-dimensional data, environmental modeling, empirical likelihood, econometric theory and financial econometrics. He became a member of Chinese Academy of Sciences in 2021 and was elected as Fellow of the American Statistical Association and Fellow of the Institute of Mathematical Statistics (IMS) in 2009. He is also an Elected Member of International Statistical Institute, an Elected Council Member of IMS during 2016 – 2019 and an Elected Board Member of the International Chinese Statistical Association (ICSA) during 2008 – 2013. He is currently serving as Scientific Secretary of Bernoulli Society since 2019 and was selected to give Peter Hall Lecture at the 12th ICSA International Conference.

Title: Versatile Differentially Private Statistical Learning for general loss functions

Abstract: This paper aims to provide a versatile privacy-preserving release mechanism along with a unified approach for subsequent parameter estimation and statistical inference. We propose a privacy mechanism based on Zero-Inflated symmetric multivariate Laplace (ZIL) noise, which requires no prior specification of subsequent analysis tasks, allows for general loss functions under minimal conditions, imposes no limit on the number of analyses, and is adaptable to the increasing data volume in online scenarios. We derive the trade-off function for the proposed ZIL mechanism that characterizes its privacy protection level. Within the M-estimation framework, we propose a novel doubly random (DR) corrected loss for the ZIL mechanism, which provides consistent and asymptotic normal M-estimates for the parameters of the target population under differential privacy constraints. The proposed approach is easy to compute without numerical integration and differentiation for noisy data.

(Joint work with Qilong Lu and Yumou Qiu)

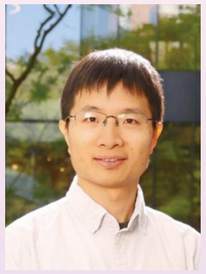
Jian Li

Bio: Jian Li is currently a full professor at Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University, headed by Prof. Andrew Yao. His major research interests lie in theoretical computer science, machine learning theory, and finance. He co-authored more than 100 research papers that have been published in major computer science conferences and journals. He received the best paper awards at VLDB and ESA, best newcomer award at ICDT.

Title: Understanding the Behaviors of Large Language Models: A perspective based on Komolgorov Complexity and Shannon's Information Theory

Abstract: Large Language Models (LLMs) have demonstrated remarkable capabilities across numerous tasks, yet principled explanations for their underlying mechanisms and several phenomena, such as scaling laws, hallucinations, and related behaviors, remain elusive. In this work, we revisit the classical relationship between compression and prediction, grounded in Kolmogorov complexity and Shannon information theory, to provide deeper insights into LLM behaviors. By leveraging the Kolmogorov Structure Function and interpreting LLM compression as a two-part coding process, we offer a detailed view of how LLMs acquire and store information across increasing model and data scales -- from pervasive syntactic patterns to progressively rarer knowledge elements. Motivated by this theoretical perspective and natural assumptions inspired by Heap's and Zipf's laws, we introduce a simplified yet representative hierarchical data-generation framework called the Syntax-Knowledge model. Under the Bayesian setting, we show that prediction and compression within this model naturally lead to diverse learning and scaling behaviors of LLMs. In particular, our theoretical analysis offers intuitive and principled explanations for both data and model scaling laws, the dynamics of knowledge acquisition during training and fine-tuning, factual knowledge hallucinations in LLMs. The experimental results validate our theoretical predictions. I will also talk about some implications of our theory to the practice of training LLMs.

Weijie Su



Bio: Weijie Su is an Associate Professor at the Wharton School and, by courtesy, in the Departments of Mathematics and Computer and Information Science at the University of Pennsylvania. He is a co-director of Penn Research in Machine Learning (PRiML) Center. Prior to joining Penn, he received his Ph.D. and bachelor's degree from Stanford University in 2016 and Peking University in 2011, respectively. His research interests span the mathematical foundations of generative AI, privacy-preserving machine learning, optimization, mechanism design, and high-dimensional statistics. He serves as an associate editor of the Journal of Machine Learning Research, Operations Research, Journal of the American Statistical Association, Journal of Machine Learning, and Foundations and Trends in Statistics. He is a Fellow of the IMS. His work has been recognized with several awards, such as the Stanford Anderson Dissertation Award, NSF CAREER Award, Sloan Research Fellowship, IMS Peter Hall Prize, SIAM Early Career Prize in Data Science, ASA Noether Early Career Award, and the ICBS Frontiers of Science Award in Mathematics.

Title: Aligning Large Language Models Through the Lens of Social Choice Theory: Impossibility and Possibility

Abstract: The alignment of large language models (LLMs) with heterogeneous human preferences presents a theoretical challenge that parallels classical problems in social choice theory. As these models increasingly inform decision-making, ensuring both fairness and the preservation of preference diversity becomes crucial. This talk examines the theoretical limits of LLM alignment through the lens of social choice theory, with a focus on how human preferences can be probabilistically represented during the alignment process. We present several impossibility and possibility results that delineate the fundamental boundaries of existing approaches to LLM alignment. This work is based on joint research (arXiv:2503.10990, 2505.20627, and 2506.12350).

Qi Lei

Bio: Qi Lei is an assistant professor of Mathematics and Data Science at the Courant Institute of Mathematical Sciences and the Center for Data Science at NYU. Previously she was an associate research scholar at the ECE department of Princeton University. She received her Ph.D. from Oden Institute for Computational Engineering & Sciences at UT Austin. She visited the Institute for Advanced Study (IAS)/Princeton for the Theoretical Machine Learning Program. Before that, she was a research fellow at Simons Institute for the Foundations of Deep Learning Program. Her research aims to develop sample- and computationally efficient machine learning algorithms and bridge the theoretical and empirical gap in machine learning. Qi has received several awards, including the Outstanding Dissertation Award, National Initiative for Modeling and Simulation Graduate Research Fellowship, Computing Innovative Fellowship, and Simons-Berkeley Research Fellowship.

Title: Discrepancies are Virtue: Weak-to-Strong Generalization through Lens of Intrinsic Dimension

Abstract: Weak-to-strong (W2S) generalization is a type of finetuning (FT) where a strong (large) student model is trained on pseudo-labels generated by a weak teacher. Surprisingly, W2S FT often outperforms the weak teacher. We seek to understand this phenomenon through the observation that FT often occurs in intrinsically low-dimensional spaces. Leveraging the low intrinsic dimensionality of FT, we analyze W2S in the ridgeless regression setting from a variance reduction perspective. For a strong student - weak teacher pair with sufficiently expressive low-dimensional feature subspaces Vs, Vw, we provide an exact characterization of the variance that dominates the generalization error of W2S. This unveils a virtue of discrepancy between the strong and weak models in W2S: the variance of the weak teacher is inherited by the strong student in Vs ∩ Vw, while reduced by a factor of dim(Vs)/N in the subspace of discrepancy Vw \ Vs with N pseudo-labels for W2S. Further, our analysis casts light on the sample complexities and the scaling of performance gap recovery in W2S. The analysis is supported with experiments on synthetic regression and real vision and NLP tasks.

Xinyu Zhang

Bio: Xinyu Zhang took up his current position as Professor in Academy of Mathematics and Systems Science, Chinese Academy of Sciences. Xinyu's main research interest lies in the field of statistics and econometrics. His publications to date consist of 80 refereed papers, including articles in leading journals such as the AoS, ET, JASA and JMLR.

Title: Towards Optimal Neural Networks: the Role of Sample Splitting in Hyperparameter Selection

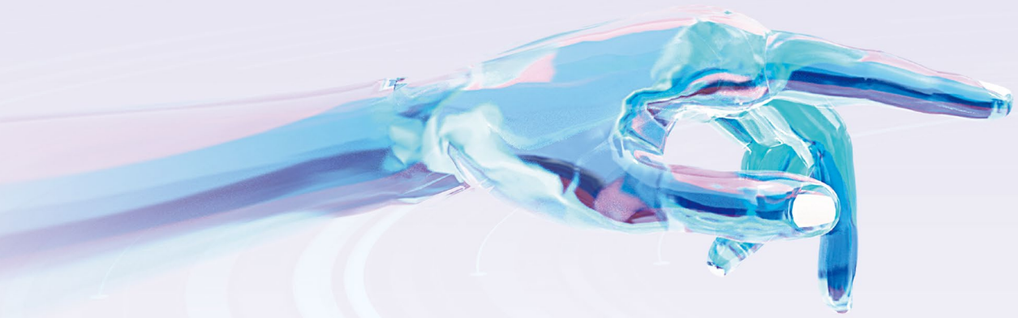
Abstract: When artificial neural networks have demonstrated exceptional practical success in a variety of domains, investigations into their theoretical characteristics, such as their approximation power, statistical properties, and generalization performance, have concurrently made significant strides. In this paper, we construct a theory for understanding the effectiveness of neural networks, which offers a perspective distinct from prior research. Specifically, we explore the rationale underlying a common practice during the construction of neural network models: sample splitting. Our findings indicate that the optimal hyperparameters derived from sample splitting can enable a neural network model that asymptotically minimizes the prediction risk. We conduct extensive experiments across different application scenarios and network architectures, and the results manifest our theory's effectiveness.

Wei Zhong

Bio: 钟威，现任厦门大学王亚南经济研究院、经济学院统计学与数据科学系教授、系主任、博士生导师。2012 年获得美国宾夕法尼亚州立大学统计学博士学位，国家优秀青年基金项目获得者（2019），国家重大人才工程领军人才（教育部，2024）。主要从事高维数据统计分析、计量经济学、统计学应用等研究。先后担任美国统计协会会刊 JASA、国际统计学会会刊 ISR 等期刊编委（AE），在 AOS、JASA、Biometrika、JMLR、JOE、JBES、AOAS、中国科学数学等国内外统计学权威期刊发表（含接收）50 余篇论文。曾获得霍英东教育基金会高等院校青年科学奖、高等学校科学研究优秀成果奖，宝钢优秀教师奖等。

Title: A Gaussian Algorithmic Stability Framework for Post-Selection Inference

Abstract: For statistical inference after model selection with the same data, the classical theory of statistical inference may not be valid, known as the double dipping problem. Zrnic & Jordan (2023) proposed an algorithmic stability method that incorporates randomization to enable post-selection corrections without computational burdens. However, it suffers from the conservativeness of the confidence intervals for complex composite algorithms due to the heavy-tailed nature of the Laplace noise. To this end, we propose a novel Gaussian Algorithmic Stability (GAS) framework for post-selection inference by introducing a new f-stability concept to obtain narrower valid confidence intervals. Different from the three tuning parameters required in Zrnic & Jordan (2023), our approach involves only two, thereby simplifying the tuning process. Theoretically, we establish the coverage guarantee of the proposed method and examine the regimes under which the proposed method yields narrower confidence intervals. In addition, we propose an aggregation approach to reduce randomness in parts which are irrelevant to inference without losing the validity of statistical inference. Numerical studies demonstrate that the proposed method has superior empirical performance.



Wei Lin

Bio: Wei Lin is Associate Professor with Tenure in the School of Mathematical Sciences and Center for Statistical Science at Peking University. He obtained his Ph.D. in Applied Mathematics from the University of Southern California in 2011 and was a postdoctoral researcher at the University of Pennsylvania from 2011 to 2014. His research interests include high-dimensional statistics, statistical machine learning, and causal inference, with applications to genomics, metagenomics, and environmental science. He has published papers in top journals such as Journal of the American Statistical Association, Biometrika, Biometrics, IEEE TIT, and Operations Research and top conferences such as ICML and COLT.

Title: Demystifying Neural Networks: Overparametrization, Generalization, and Feature Learning

Abstract: Neural networks have achieved remarkable successes in modern deep learning practice. Yet, their ability to generalize well even in the overparametrized regime and to learn meaningful representations from data remains controversial and mysterious in theory. In this talk, we suggest new statistical theories to elucidate the superiority of two-layer ReLU neural networks over classical machine learning methods. In the first work, we approach the problem from a nonparametric viewpoint and derive unified generalization bounds for any finite-width network, providing an intriguing explanation for the double descent phenomenon. In the second work, we consider the teacher-student setting where the data are generated from a parametric teacher network with well-separated features, and show that the student work learns these features with optimal rates, leading to performance guarantees for downstream tasks. Our proofs rely on techniques from high-dimensional statistics by exploiting the geometry of the optimal solution and a group lasso reparametrization.

Yuting Wei

Bio: Dr. Yuting Wei is an Associate Professor in the Statistics and Data Science Department at the Wharton School, University of Pennsylvania. Prior to that, Dr. Wei spent two years at Carnegie Mellon University as an assistant professor and one year at Stanford University as a Stein's Fellow. She received her Ph.D. in statistics at the University of California, Berkeley. She was the recipient of the 2025 Gottfried E. Noether Early Career Scholar Award, Google Research Scholar Award, NSF Career award, and the Erich L. Lehmann Citation from the Berkeley statistics department. Her research interests include high-dimensional and non-parametric statistics, reinforcement learning, and diffusion models.

Title: To Intrinsic Dimension and Beyond: Efficient Sampling in Diffusion Models

Abstract: The denoising diffusion probabilistic model (DDPM) has become a cornerstone of generative AI. While sharp convergence guarantees have been established for DDPM, the iteration complexity typically scales with the ambient data dimension of target distributions, leading to overly conservative theory that fails to explain its practical efficiency. This has sparked recent efforts to understand how DDPM can achieve sampling speed-ups through automatic exploitation of intrinsic low dimensionality of data. This talk explores two key scenarios: (1) For a broad class of data distributions with intrinsic dimension k, we prove that the iteration complexity of the DDPM scales nearly linearly with k, which is optimal under the KL divergence metric; (2) For mixtures of Gaussian distributions with k components, we show that DDPM learns the distribution with iteration complexity that grows only logarithmically in k. These results provide theoretical justification for the practical efficiency of diffusion models.

Shurong Zheng

Bio: 郑术蓉，东北师范大学，主要研究方向是大维随机矩阵理论及其应用，目前已经在 Annals of Statistics,JASA,Biometrika 等统计学期刊上发表跟大维随机矩阵有关的学术论文 60 余篇。现担任 Statistica Sinica, JMVA 等多个国内外学术期刊的编委。

Title: Difference between Large Statistical Model and Medium Statistical Model

Abstract: In this talk, we will show that the large statistical model will have a very different performance compared with the medium model. For example, when the sample size is fixed and the dimension of data increases (convergence regime), the power function of the log-likelihood ratio test for the covariance matrix will tend to one. Moreover, under the convergence regime, the estimated number of factors in the factor model will be more accurate. Moreover, we will give some other examples to show the difference between large statistical model and medium statistical model.

Gen Li

Bio: Gen Li is currently an assistant professor in the Department of Statistics at the Chinese University of Hong Kong. He received the Ph.D. in the Department of Electronic Engineering at Tsinghua University in 2021, and received the bachelor's degree from the Department of Electronic Engineering and Department of Mathematics at Tsinghua University in 2016. His research interests include diffusion based generative model, reinforcement learning, high-dimensional statistics, machine learning.

Title: Faster Convergence and Acceleration for Diffusion-Based Generative Models

Abstract: Diffusion models, which generate new data instances by learning to reverse a Markov diffusion process from noise, have become a cornerstone in contemporary generative modeling. While their practical power has now been widely recognized, the theoretical underpinnings remain underdeveloped. Particularly, despite the recent surge of interest in accelerating sampling speed, convergence theory for these acceleration techniques remains limited. In this talk, I will first introduce an acceleration sampling scheme for stochastic samplers that provably improves the iteration complexity under minimal assumptions. The second part focuses on diffusion-based language models, whose ability to generate tokens in parallel significantly accelerates sampling relative to traditional autoregressive methods. Adopting an information-theoretic lens, we establish a sharp convergence theory for diffusion language models, thereby providing the first rigorous justification of both their efficiency and fundamental limits.

Yang Yuan

Bio: Yang Yuan is now an assistant professor at IIIS/CollegeAI, Tsinghua. He is also a PI at Shanghai Qizhi Institute and Shanghai AI Lab. He finished his undergraduate study at Peking University in 2012. Afterwards, he received his PhD at Cornell University in 2018, advised by Professor Robert Kleinberg. Before joining Tsinghua, he spent one year at MIT Institute for Foundations of Data Science (MIFODS) as a postdoc researcher. He works on AI+Healthcare, AI Theory and Applied Category Theory.

Title: Structural AI: From Pretraining Abstractions to Formal Reasoning Structures

Abstract: In this talk, we explore how structural perspectives offer new insights into the foundations and capabilities of AI systems. We first show that even under ideal conditions—perfect pretraining, infinite data, and unlimited compute—the effectiveness of prompt-based learning is fundamentally limited to representable tasks. Fine-tuning, by contrast, can theoretically recover any task within the category defined by the pretext task. Building on this foundation, we turn to reasoning in large language models. We introduce conceptual unfolding—a method of structurally expanding task definitions—to enhance task understanding. By using Coq to generate formal, semantically rich prompts, we improve model performance on a mathematical reasoning task from 33% to 52%. This suggests that structural clarity can improve AI reasoning.

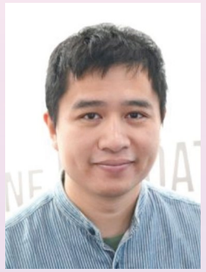
Chi Jin

Bio: Chi Jin is an Assistant Professor of Electrical and Computer Engineering at Princeton University. He received his Ph.D. in Computer Science from UC Berkeley, advised by Michael I. Jordan. His research focuses on building intelligent agents capable of complex strategy, reasoning, and planning. His group has made key contributions to the mathematical foundations of machine learning, especially in nonconvex optimization, reinforcement learning, and game theory/multi-agent systems. Recently, his work has expanded to enhancing LLM reasoning and developing LLM-based agents for mathematics and games. His group's open-source model, Goedel-prover, achieves state-of-the-art performance in formal reasoning. He is a recipient of the NSF CAREER Award, Sloan Research Fellowship, and Keyes/Emerson Faculty Advancement Award.

Title: Goedel-Prover: A Frontier Model for Open-Source Automated Theorem Proving

Abstract: We introduce Goedel-Prover, an open-source Large Language Model (LLM) that achieves the state-of-the-art performance in automated formal proof generation for mathematical problems. A key challenge in training LLMs for formal reasoning is the scarcity of formal data. To address this, we train formalizers to translate a substantial corpus of mathematical problems from natural language into formal language (Lean 4) in various styles, producing 1.64 million syntactically correct and content-accurate formal statements. We then train a prover iteratively, alternating between generating verified proofs from statements and refining the model using these proofs. Our model outperforms all existing open-source models for whole-proof generation across multiple benchmarks. On the miniF2F benchmark (Pass@32), our model attains a 57.6% success rate, surpassing the previous best open-source model by 7.6%. On PutnamBench, it successfully solves 7 problems (Pass@512), ranking first on the leaderboard. Furthermore, it generates 29.7K formal proofs for Lean Workbook problems, nearly doubling the 15.7K produced by earlier works.

Difan Zou



Bio: Dr.Difan Zou is an assistant professor in computer science department and institute of data science at HKU. He has received his PhD degree in Department of Computer Science, University of California, Los Angeles (UCLA). His research interests are broadly in machine learning, deep learning theory, graph learning, mechanism interpretation, and interdisciplinary research between AI and other subjects. His research is published in top-tier machine learning conferences (ICML, NeurIPS, COLT, ICLR) and journal papers (IEEE Trans., JMLR, Nature Comm., PNAS, etc.). He serves as an area chair/senior PC member for NeurIPS, ICML and AAAI, and PC members for ICLR, COLT, etc.

Title: On the Mechanism Interpretability of LLM for Fine-tuning and Reasoning

Abstract: While Reinforcement Learning (RL) and Fine-Tuning demonstrably enhance Large Language Model (LLM) capabilities, particularly in reasoning and task adaptation, the underlying mechanisms remain poorly understood. This talk integrates insights from two complementary studies to advance mechanistic interpretability. First, we dissect Reinforcement Learning with Verifiable Rewards (RLVR), revealing its core benefit lies in optimizing the selection of existing high-success-rate reasoning patterns, with theoretical convergence analyses showing distinct dynamics for strong versus weak initial models (mitigated by prior supervised fine-tuning). Second, we employ circuit analysis to interpret fine-tuning mechanisms, uncovering that circuits undergo significant edge changes rather than merely adding components, contrasting prior findings. Leveraging this, we develop a circuit-aware LoRA method, improving performance over standard LoRA by 2.46%. Furthermore, we explore combining circuits for compositional tasks. Together, these studies provide novel theoretical and empirical insights: RL enhances reasoning primarily through pattern selection, while fine-tuning fundamentally rewires circuit connections. This deeper understanding informs the design of more effective and interpretable adaptation strategies for LLMs.

Qiang Sun

Bio: Qiang Sun is a Professor in the Department of Statistical Sciences and the Department of Computer Science at the University of Toronto (UofT), and a visiting faculty member at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), where he leads the NeXAIS Lab. His research lies at the intersection of statistics and AI, with a focus on trustworthy AI, efficient generative AI (GenAI), and the foundations of AGI. His work is often inspired by real-world challenges in technology, finance, and science. Prior to joining UofT, he was an Associate Research Scholar at Princeton University. He received his Ph.D. in Statistics from the University of North Carolina at Chapel Hill (UNC-CH) and his B.S. in SCGY from the University of Science and Technology of China (USTC). In addition to his academic roles, Qiang serves as an Associate Editor for the Electronic Journal of Statistics (EJS), an Area Chair for several leading machine learning conferences, and a member of the Executive Committee of the BAAI Young Scientist Club. He is a recipient of the James E. Grizzle Distinguished Alumni Award from UNC-CH and the Connaught New Researcher Award, and has delivered a series of invited plenary talks.

Title: Making AI trustworthy

Abstract: Large-scale decision-making in high-stakes domains, such as recommendation systems and algorithmic trading, is increasingly powered by deep learning. However, these models often exhibit fragility in the face of both structured and random noise. How can we make them truly trustworthy? While explainability is frequently touted as a pathway to trustworthy AI, we argue that it is neither sufficient nor necessary. Instead, from a statistical perspective, we emphasize the importance of adaptivity, the ability of algorithms to consistently generalize to future scenarios under varying conditions. A trustworthy algorithm should demonstrate stable performance across diverse environments. In this talk, I will share recent research motivated by collaborations with industry, highlighting practical strategies for building robust, adaptive systems. We will conclude with key takeaways.

Xiaodong Yan

Bio: 严晓东，西安交通大学数学与统计学院教授，博士生导师，入选国家级青年人才项目和校内青拔 A 类支持计划 , 研究方向为统计决策、统计推断和统计优化等，学术成果发表在著名期刊 JRSSB, AOS, JASA, JOE 以及人工智能顶级会议 ICML, AAAI, AISTATS 等。在“高等教育出版社”以独立主编出版了《机器学习》、《数据科学实践基础 - 基于 R》两部教材。

Title: Statistical Detection for Silent Data Corruption during Large-scale Model Training

Abstract: During large-scale model training, Silent Data Corruption (SDC) caused by hardware failures has emerged as a core challenge threatening model robustness and safety. SDC arises stochastically from factors such as chip transistor defects, voltage fluctuations, and temperature variations. Its concealment and latency make it difficult for traditional hardware detection methods, software redundancy techniques, and algorithm-level solutions to meet practical application standards in terms of coverage, computational overhead, and false alarm rate. To capture the system state transitions induced by SDC, this paper formulates the problem as Change-point Detection and constructs a novel test statistic based on strategic central limit theorem. This approach achieves high correct detection rate (>95%) with an ultra-low false positive rate (<10^-8).. Simulation studies provide supporting evidence for the algorithm's performance, demonstrating its robustness with limited samples. In addition, an empirical analysis was added based on Huawei's large model training scenario to further validate the effectiveness of the method.

RMTA
CHINA
中国现场统计研究会
随机矩阵理论与应用分会

清華大學 统计与数据科学系
Department of Statistics and Data Science, Tsinghua University