



西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

MLSTAT

Machine Learning
& Statistics

第四届机器学习与统计学 (MLSTAT) 会议

会议手册

主办单位：中国现场统计研究会机器学习分会

承办单位：西南财经大学统计研究中心

西南财经大学统计与数据科学学院

协办单位：国家天元数学西南中心

中国现场统计研究会数据科学与人工智能分会

2026年7月15日-17日

四川·成都



目录

会议须知	01
会议日程概览	02
详细会议日程	03
报告详情	06
参会人员名单	19
校内地图	27
温馨提示	28

会议须知

机器学习与统计学 (MLSTAT) 会议是由中国现场统计研究会机器学习分会主办的学术会议。会议旨在促进机器学习与统计领域的国内外学者进行学术交流, 引领机器学习与统计共同交叉发展的学术文化, 推动作为数据科学与人工智能的奠基性学科的进步, 以此助力相关数字经济产业的发展。

会议相关安排通知如下:

一、会议时间地点

会议时间: 2026年7月15日-17日

会议地点: 西南财经大学柳林校区

二、会议主要议程

7月15日 下午: Tutorial

7月16日 上午: 开幕式+特邀报告

下午: 特邀报告

晚上: poster展示

7月17日 上午: 特邀报告

三、会议费用

学生代表 200元, 教师代表500元。会务组将承担会议期间的用餐, 其他费用敬请自理!

四、会务服务

李晓丽 电话: 18380123625

杨森慧 电话: 15608171281

五、其他

7月15日, 凭会议注册手机号后四位进校, 之后凭会议胸牌进校。

会议日程概览

日期	时间段	主要事宜	地点
7月15日	12:00-20:30	会议报到	西财柳林校区弘远楼一楼大厅
	13:00-17:00	Tutorial	西财柳林校区经世楼E102
	14:00-17:00	西南财经大学 统计学科发展 专家咨询会	
	17:00-19:00	晚餐（西财柳林校区五谷堂二楼C区（上电梯左边））	
7月16日	8:40-9:00	欢迎仪式&合影	弘远楼105会议室
	9:00-10:20	特邀报告（一）	
	10:20-10:40	茶歇	
	10:40-12:00	特邀报告（二）	
	12:00-13:15	午餐（三味堂二楼）	
	13:40-15:40	特邀报告（三）	弘远楼105会议室
	15:40-16:00	茶歇	
	16:00-17:20	特邀报告（四）	
	17:30-19:00	晚餐（西财柳林校区五谷堂二楼C区（上电梯左边））	
	19:30-20:00	中国现场统计研究会 机器学习分会理事会	弘远楼109会议室
	19:30-21:00	Poster展示	弘远楼一楼大厅
7月17日	9:00-10:20	特邀报告（五）	弘远楼105会议室
	10:20-10:40	茶歇	
	10:40-12:00	特邀报告（六）	
	12:00-13:15	午餐（西财柳林校区五谷堂二楼C区（上电梯左边））	

详细会议日程

会议报到、入住（7月15日，星期三）			
日期	时间段	主要事宜	地点
7月15日	12:00-20:30	会议报到	西财柳林校区弘远楼一楼大厅
	13:00-17:00	Tutorial	西财柳林校区经世楼E102
	14:00-17:00	西南财经大学统计学科 发展专家咨询会	
	17:00-19:00	晚餐（西财柳林校区五谷堂二楼C区（上电梯左边））	
学术交流（7月16日，星期四）			
地点：西南财经大学柳林校区弘远楼105会议室			
开幕式&合影 主持人：林华珍 西南财经大学			
8:40-8:45	欢迎仪式		
8:45-8:50	西南财经大学党委常委 副校长 李志生致辞		
8:50-8:55	中国现场统计研究会机器学习与统计分会理事会 理事长 张志华教授致辞		
8:55-9:00	合影		
特邀报告（一）（按姓氏拼音音序） 主持人：张志华 北京大学			
9:00-9:40	报告人：蔡占锐 香港大学 题目：A Statistical Framework for Alignment with Biased AI Feedback		
9:40-10:20	报告人：代忠祥 香港中文大学（深圳） 题目：Optimizing Agentic Workflows: From Task-Level Search to Dynamic Construction		
10:20-10:40	茶歇		

特邀报告 (二) 主持人: 史成春 伦敦政治经济学院	
10:40-11:20	报告人: 戴晓武 加州大学洛杉矶分校 题目: Multi-Agent Language Models: Games, Reasoning, and Alignment
11:20-12:00	报告人: 吴磊 北京大学 题目: 泛函尺度律: 大模型尺度现象的动力学解释
12:00-13:15	午餐 地点: 三味堂二楼
特邀报告 (三) 主持人: 方聪 北京大学	
13:40-14:20	报告人: 刘方辉 上海交通大学 题目: From SLT4AI to AI4SLT: Generalization under scaling, and infrastructure Lean 4
14:20-15:00	报告人: 马梓业 香港城市大学 题目: Escaping Local Minima Deterministically and Provably in Matrix Sensing: Power of Simulated Over-Parametrization
15:00-15:40	报告人: 毛小介 清华大学 题目: Contextual Linear Optimization under Full and Partial Feedback
15:40-16:00	茶歇
特邀报告 (四) 主持人: 胡天阳 香港中文大学 (深圳)	
16:00-16:40	报告人: 束俊 西安交通大学 题目: 基于Koopman理论的连续谱动力系统表示与预测方法
16:40-17:20	报告人: 周帆 上海财经大学 题目: AI for Statistics的一些最新进展
17:30-19:00	晚餐 地点: 西财柳林校区五谷堂二楼C区 (上电梯左边)
19:30-20:00	中国现场统计研究会机器学习分会理事会 主持人: 张志华 北京大学 地点: 弘远楼109会议室
19:30-21:00	Poster 地点: 弘远楼一楼大厅

学术交流 (7月17日, 星期五)

地点: 西南财经大学柳林校区弘远楼105会议室

特邀报告 (五)

主持人: 李赛 清华大学

9:00-9:40	报告人: 宋晓 佐治亚大学 题目: Evaluating biomarkers for treatment selection from reproducibility studies
9:40-10:20	报告人: 杨云斐 中山大学 题目: Approximating and learning smooth functions by ReLU neural networks

10:20-10:40	茶歇
-------------	-----------

特邀报告 (六)

主持人: 常象宇 西安交通大学

10:40-11:20	报告人: 丁杰 明尼苏达大学 题目: The Future of AI Scientists: Emerging Directions and Fundamental Challenges
11:20-12:00	报告人: 周峰 中国人民大学 题目: SpecTr-GBV: Multi-Draft Block Verification Accelerating Speculative Decoding

12:00-13:15	午餐 地点: 西财柳林校区五谷堂二楼C区 (上电梯左边)
-------------	---

报告详情

(按报告顺序排序)

A Statistical Framework for Alignment with Biased AI Feedback

报告人: 蔡占锐 (香港大学)

摘要: Modern alignment pipelines are increasingly replacing expensive human preference labels with evaluations from large language models (LLM-as-Judge). However, AI labels can be systematically biased compared to high-quality human feedback datasets. In this paper, we develop two debiased alignment methods within a general framework that accommodates heterogeneous prompt-response distributions and external human feedback sources. Debiased Direct Preference Optimization (DDPO) augments standard DPO with a residual-based correction and density-ratio reweighting to mitigate systematic bias, while retaining DPO's computational efficiency. Debiased Identity Preference Optimization (DIPO) directly estimates human preference probabilities without imposing a parametric reward model. We provide theoretical guarantees for both methods: DDPO offers a practical and computationally efficient solution for large-scale alignment, whereas DIPO serves as a robust, statistically optimal alternative that attains the semiparametric efficiency bound. Empirical studies on sentiment generation, summarization, and single-turn dialogue demonstrate that the proposed methods substantially improve alignment efficiency and recover performance close to that of an oracle trained on fully human-labeled data.

个人简介: 蔡占锐现任香港大学经管学院创新与信息管理系统助理教授。于2021年于宾夕法尼亚州立大学获得统计学博士学位，并于之后在卡内基梅隆大学进行博士后研究，以及在爱荷华州立大学统计系担任助理教授。研究兴趣包括统计在大模型中的应用，差分隐私中的统计推断，以及机器学习在统计方法中的应用等。

Optimizing Agentic Workflows: From Task-Level Search to Dynamic Construction

报告人：代忠祥（香港中文大学（深圳））

摘要： The rapid evolution of large language model (LLM) agents has demonstrated their immense potential in complex reasoning tasks, yet relying on manually designed agentic workflows fundamentally limits their adaptability and performance. In this talk, we present a trajectory of our recent research that progressively automates and refines workflow construction, moving from static to highly dynamic paradigms. We begin by introducing a task-level optimization framework that leverages bandit-guided graph evolution to discover a single, robust workflow capable of generalizing across all queries within a specific task domain. To push beyond the inherent performance ceiling of such task-level generalization, we then present a query-level orchestration method that utilizes contextual bandits to adaptively assign the optimal, customized workflow for every individual input. Finally, we address the limitations of one-shot, fixed workflow generation with Workflow-R1, a novel reinforcement learning-based approach for multi-turn workflow construction. Unlike previous methods where the workflow remains static after initial generation, Workflow-R1 optimizes group sub-sequence policies to dynamically construct and adjust the workflow online, conditioning subsequent steps directly on intermediate execution results. Together, these three works trace a clear path from static task-wide optimization to highly dynamic, execution-aware reasoning pathways, offering a comprehensive foundation for building more autonomous and resilient AI agents.

个人简介： 代忠祥博士现任香港中文大学（深圳）数据科学学院助理教授、博士生导师，入选国家级高层次青年人才，并获评校长青年学者。主持广东省自然科学基金优秀青年项目、国家自然科学基金青年科学基金项目（C类）、深圳市自然科学基金面上项目，以及华为大模型智能体合作项目等。于2024年在麻省理工学院（MIT）担任博士后研究员，并于2021年至2023年在新加坡国立大学（NUS）从事博士后研究。他于2015年和2021年分别获得新加坡国立大学的电气工程学士学位和计算机科学博士学位。研究兴趣涵盖机器学习的理论与应用。在应用层面，专注于大语言模型，研究方向包括基于大模型的智能体、大模型个性化、大模型在线路由、基于大模型的社会模拟、大模型提示优化等；在理论层面，深入研究多臂老虎机算法。已在顶级AI会议和期刊上发表论文38余篇，其中30余篇发表于NeurIPS、ICML和ICLR，并担任NeurIPS、ICML和ICLR的领域主席（Area Chair）。

Multi-Agent Language Models: Games, Reasoning, and Alignment

报告人: 戴晓武 (加州大学洛杉矶分校)

摘要: Large Language Models (LLMs) could prone to inconsistencies and hallucinations. We introduce Peer Elicitation Games (PEG), a training-free, game-theoretic framework for aligning LLMs through a peer elicitation mechanism involving a generator and multiple discriminators instantiated from distinct base models. Discriminators interact in a peer evaluation setting, where rewards are computed using a determinant-based mutual information score that provably incentivizes truthful reporting without requiring ground-truth labels. We establish theoretical guarantees showing that each agent, via online learning, achieves sublinear regret in the sense their cumulative performance approaches that of the best fixed truthful strategy in hindsight. Moreover, we prove last-iterate convergence to a truthful Nash equilibrium, ensuring that the actual policies used by agents converge to stable and truthful behavior over time. I'll also discuss the extension of PEG to the multi-agent reasoning and inference-time alignment.

个人简介: Xiaowu Dai is an assistant professor in the Departments of Statistics and Data Science, and of Biostatistics at UCLA. Before joining UCLA, he did a postdoc at UC Berkeley working with Prof. Mike Jordan, and received a Ph.D. in Statistics at UW-Madison advised by Prof. Grace Wahba. His research focuses on statistical theory and methodology for real-world problems that blend computational, inferential, and economic considerations.

泛函尺度律：大模型尺度现象的动力学解释

报告人：吴磊（北京大学）

摘要：大模型尺度律揭示：性能随训练数据规模和计算量的增长呈现出可预测的幂律提升。这一规律深刻推动了现代人工智能的发展，但长期停留在经验观察层面，缺乏理论理解。为探究其成因，我们引入幂律核回归（power-law kernel regression）这一简化模型，作为理论原型来抽象尺度现象的关键机制。基于该模型的动力学推导，我们提出泛函尺度律（Functional Scaling Laws, FSL）：通过“内蕴时间”这一核心概念，将尺度律扩展为刻画整个训练过程的“泛函”形式，从而统一描述了不同模型规模和超参数设置下的损失演化。更进一步，FSL还揭示了“预热-稳定-退火”等常用学习率调度策略的有效性，显示出对实际大模型训练的潜在指导价值。

个人简介：吴磊，北京大学数学科学学院和国际机器学习研究中心助理教授，主要研究方向为深度学习的数理基础。2012年毕业于南开大学，获数学与应用数学学士学位；2018年毕业于北京大学，获得计算数学博士学位。2018年11月至2021年10月，先后在美国普林斯顿大学和宾夕法尼亚大学从事博士后研究工作。相关成果发表于NeurIPS、ICML、Annals of Statistics、Journal of Machine Learning Research、IEEE Transactions on Information Theory等国际顶级会议与期刊。

From SLT4AI to AI4SLT: Generalization under scaling, and infrastructure Lean 4

报告人: 刘方辉 (上海交通大学)

摘要: This talk connects two directions in modern statistical learning theory. First, I will briefly discuss SLT4AI: how generalization curves behave under suitable model capacities, e.g., norm-based capacities. Our results present generalization under scaling via deterministic equivalence, which challenges classical views of generalization, including the U-shaped bias–variance curve, double descent, and empirical scaling laws. Second, I will talk about AI4SLT by presenting our Lean 4 formalization of SLT from empirical process theory with more than 30,000 lines code under a human-AI collaborative workflow (with some design principles), including Gaussian Lipschitz concentration, Dudley's entropy integral theorem for sub-Gaussian processes, and application to localized least squares. It highlights how formal AI4SLT infrastructure can make modern learning theory reliable, reusable, and machine-checkable.

个人简介: 刘方辉，上海交通大学自然科学研究院与数学学院副教授，数学学院与人工智能学院博士生导师。研究方向为现代机器学习的数学理论与大模型机理分析。其主要研究工作包括函数空间视角下的机器学习理论、尺度扩展下的泛化理论，并进一步推动其在大模型微调与参数高效训练中的应用，以及机器学习理论的形式化推理系统 (AI4SLT)。近五年在 JMLR、SIAM、NeurIPS、ICML、ICLR、TPAMI 发表论文 20 余篇。2019 年博士毕业于上海交通大学，曾在 KU Leuven、EPFL 从事博士后研究，在英国 University of Warwick 担任助理教授。2023 年入选国家高层次青年人才，获 AAAI 2024 新教师奖，2025 年入选 TUM 全球访问教授计划等。担任 NeurIPS、ICLR、AISTATS 等会议领域主席。

Escaping Local Minima Deterministically and Provably in Matrix Sensing: Power of Simulated Over-Parametrization

报告人: 马梓业 (香港城市大学)

摘要: Low-rank matrix sensing is a fundamental yet challenging nonconvex problem whose optimization landscape typically contains numerous spurious local minima, making it difficult for gradient-based optimizers to converge to the global optimum. Recent work has shown that tensor over-parameterization can, in principle, convert such local minima into strict saddle points; which also serves as a theoretical footnote to why scaling can improve generalization and performance in modern machine learning models like LLMs. Motivated by this observation, we propose a new escape mechanism that simulates the landscape and escape direction of the tensor-lifted space, without resorting to actually lifting the problem, since that would be computationally intractable. In essence, we designed a mathematical framework to project over-parametrized escape directions onto the original parameter space that could guarantee a strict decrease of objective value from existing local minima. To the best of our knowledge, this represents the first deterministic framework that could escape spurious local minima with guarantee, especially without using random perturbations or heuristic estimates.

个人简介: 马梓业现为香港城市大学计算机系助理教授 (presidential assistant professor)，博士毕业于加州大学伯克利分校 (UC Berkeley) 电子工程与计算机科学系。他的主要研究方向是AI基础算法，机器学习理论，以及数学优化 (非凸优化)。他的论文主要发表在主流人工智能会议及期刊上 (NeurIPS, ICML, JMLR, AISTATS, AAAI等)，也曾多次获得口头报告 (oral) 荣誉，以及2023年的AISTATS Notable Paper。他曾获批香港研究资助局 (rgc) 的“杰出青年学者计划 (ecs)”以及国家自然科学基金 (C类)。

Contextual Linear Optimization under Full and Partial Feedback

报告人: 毛小介 (清华大学)

摘要: This talk is about Contextual Linear Optimization (CLO) across two feedback regimes, where we study the traditional two-stage Estimate-Then-Optimize (ETO) approach and the new integrated Induced Empirical Risk Minimization (IERM) framework. In the full-feedback setting, we theoretically demonstrate that under model correct specification, ETO can surprisingly achieve faster regret convergence rates than IERM by leveraging problem-specific geometric properties. In partial-feedback settings (bandit and semi-bandit), we propose a unified offline IERM framework and establish novel fast-rate guarantees. Numerical experiments on shortest path problems validate our theoretical findings across different regimes.

个人简介: 毛小介, 清华大学经济管理学院管理科学与工程系副教授, 2016年获得武汉大学经济学学士学位, 2021年获得美国康奈尔大学统计与数据科学专业博士学位。研究兴趣主要包括因果推断、数据驱动的决策方法、统计机器学习等。

基于Koopman理论的连续谱动力系统表示与预测方法

报告人：束俊（西安交通大学）

摘要：对高维时空混沌动力系统表示与预测，仍然是动力系统理论与机器学习领域中的一个基础性挑战。高维、非线性且具有连续谱结构的动力系统广泛存在于气候演化、湍流流动、复杂网络传播以及神经动力学等现实场景中。尽管目前数据驱动方法能够实现较为准确的短期预测，但在以宽频或连续谱为主导的系统中，它们往往缺乏稳定性、可解释性和可扩展性。Koopman 理论为非线性动力学的表征和预测提供了一个线性化视角，但现有方法通常依赖于有限维逼近，这在高维场景下往往导致性能退化。本报告提出一种新的神经Koopman方法，通过将可逆运动与不可逆耗散分离，实现对动力系统的结构化表示。该方法在提高长期预测精度与稳定性的同时，也有助于揭示混沌行为中哪些方面是可以被理解和学习的。

个人简介：束俊，西安交通大学数学与统计学院副教授，博导。于2016年和2023年分别获西安交通大学理学学士与理学博士学位。目前从事于机器学习基础理论与算法研究；提出基于“模拟学习方法论”（简称SLeM）的学习框架，建立起相应的统计学习理论，为机器学习自动化提供了一个可行的理论途径，并在此框架下针对训练数据自选择、样本标记自校正、网络结构自调节、表现度量自构建、优化算法自设计等机器学习自动化任务研发了系列有效的机器学习自动化核心算法与技术。已在JMLR/TPAMI/TMLR/NSR/NeurIPS/ICML等国际顶级期刊和会议发表学术论文25余篇；ESI高被引论文2篇。谷歌学术引用2000余次，单篇论文引用超1100次。主持科技部重点研发计划青年科学家项目、基金委面上项目以及多项企业横向课题等。

AI for Statistics的一些最新进展

报告人：周帆（上海财经大学）

摘要：本报告将介绍我们近期在AI for Statistics方面的一系列探索与进展。围绕“大模型的统计推理能力”这一核心问题，我们从底层数据、知识建模与辅助研究三个层面展开研究。首先，我们构建了StatEval——首个面向统计学的综合性问答与推理的综合性数据集和评测基准，系统覆盖从本科与研究生基础知识到前沿科研级证明问题，填补了现有大模型基础数据与评测中统计学维度长期缺失的空白。特别的，我们设计了一个具有人机协同验证的多智能系统TRACE，能够在确保学术严谨性的同时，自动化完成结构抽取、逻辑解析和多难度题目生成。基于StatEval，我们提出了一种针对复杂证明的自适应过程化评分流水线，支持对推理能力进行细粒度的逐步评估。此外，利用StatEval进行检索增强生成（RAG）和领域微调，可明显提升推理能力。在此基础上，我们还开发了一个可以自动化进行研究级统计定理证明的智能体StatProver，能够辅助研究者构建证明思路与框架，处理科研级别的统计定理证明时频繁出现的逻辑断裂与公式幻觉问题，实现从统计学命题到完整证明的全自动生成与自我纠错。

个人简介：周帆，上海财经大学统计与数据科学学院副教授，国家级青年人才，博士毕业于美国北卡罗来纳大学教堂山分校，现担任统计学顶刊JASA的副主编与人工智能顶会NeurIPS等的领域主席。研究兴趣包括深度学习，强化学习的算法与理论，大模型，因果推断，在包括JASA, JMLR, NeurIPS, ICML, ICLR等统计学，机器学习顶刊和顶会上发表一作通讯文章数十篇，曾获泛华统计协会国际会议新研究者奖,UNC James E. Grizzle杰出校友奖和Barry H. Margolin Award.

Evaluating biomarkers for treatment selection from reproducibility studies

报告人: 宋晓 (佐治亚大学)

摘要: We consider evaluating biomarkers for treatment selection under assay modification. Survival outcome, treatment, and Affymetrix gene expression data were obtained from cancer patients. Consider migrating a gene expression biomarker to the Illumina platform. We propose an approach that allows a quick evaluation of the migrated biomarker with only a reproducibility study needed to compare the two platforms, achieved by treating the original biomarker as an error-contaminated observation of the migrated biomarker. We adopt a nonparametric regression model to characterize the relationship between the event rate and the biomarker and obtain an optimal marker-based treatment regime. Using B-spline approximation for nonparametric regression functions, we estimate the treatment regime via the SIMEX approach. SIMEX is a simple and general method that can handle measurement errors in complex cases. Through establishing the uniform convergence rate of the SIMEX estimator for the nonparametric regression functions, we derive the consistency and asymptotic normality of the empirical estimator of the biomarker performance measure. The approach is assessed by simulation studies and demonstrated through application to lung cancer data.

个人简介: 宋晓, 佐治亚大学流行病和生物统计系教授, 本科和硕士毕业于北京大学数学系, 博士毕业于北卡罗莱纳州立大学统计系, 曾任华盛顿大学(University of Washington) 公共卫生学院生物统计系研究助理教授, 主要研究兴趣包括生存分析、测量误差和缺失数据、非参数方法、高维数据和图像数据、以及基于生物标志物的医学检验预测与精准医疗。

Approximating and learning smooth functions by ReLU neural networks

报告人: 杨云斐(中山大学)

摘要: In this talk, we will discuss some recent progresses on the approximation and learning theory of ReLU neural networks. We divide the approximation theory of neural networks into two parts according to the methods. In the first method, we approximate smooth functions by piecewise polynomials and then construct neural networks to approximate these piecewise polynomials. Using constructive approximation, one can derive optimal approximation rates in terms of the width and depth. In the second method, we consider the function space of shallow neural networks, which is called Barron space, and use random approximation method to derive approximation bounds in this space. By studying the relation between the Barron space and the smooth function spaces, we can characterize the approximation error of shallow neural networks by the width and certain norm of the weights. As an application, we will discuss how these approximation results can be used in nonparametric regression problems. In particular, we will show that least squares estimations based on deep or shallow neural networks can achieve minimax optimal rates of convergence for learning smooth function classes in various settings.

个人简介: 杨云斐, 中山大学数学学院(珠海)副教授, 2022年博士毕业于香港科技大学。研究兴趣包括机器学习、应用调和分析和逼近论。目前的研究工作主要集中于神经网络的逼近与泛化理论。相关研究成果发表于Journal of Machine Learning Research、Applied and Computational Harmonic Analysis、Constructive Approximation、NeurIPS等期刊和会议。

The Future of AI Scientists: Emerging Directions and Fundamental Challenges

报告人：丁杰（明尼苏达大学）

摘要：Modern AI systems are rapidly evolving from passive tools into agentic systems that can reason, learn, collaborate, and interact with human scientists in the research process. This talk discusses emerging directions in AI scientist systems, drawing inspiration from human-centered principles such as lifelong learning, teaching, collaboration, objective alignment, and auditing. The goal is to identify fundamental research challenges and stimulate new perspectives on building steerable and productive AI systems in complex real-world AI-for-Science settings.

个人简介：丁杰，明尼苏达大学统计学院副教授，电子工程系、计算机系客座教授，AI中心核心成员，明尼苏达大学AI健康医疗工作小组联合组长，本科毕业于清华大学，博士毕业于哈佛大学。其研究位于人工智能、统计学、信息科学的交叉领域，当前重点关注可扩展与可信赖的人工智能、AI for Science，以及人机协作系统。曾任 Amazon Scholar，与 Amazon Alexa 及 AGI 团队合作开展联邦学习与多模态大模型相关研究。其工作获得美国国家科学基金会 CAREER Award、ARO Young Investigator Award、Cisco Research Award、Meta Research Award 及多项最佳论文奖。他还发起了明尼苏达大学首个面向全校的人工智能开源课程。近年来，他进一步关注AI智能体系统的工程化与落地实践，探索如何让科研人员或各领域专家在短时间内构建可控的AI智能体，并围绕其联合创立的AI-for-Science平台MorphMind及相关开源项目WorldSeed开展系统性探索。

SpecTr-GBV: Multi-Draft Block Verification Accelerating Speculative Decoding

报告人: 周峰 (中国人民大学)

摘要: Autoregressive language models achieve state-of-the-art performance across a wide range of natural language processing tasks, but suffer from high inference latency due to their sequential decoding nature. Speculative decoding (SD) mitigates this by employing a lightweight draft model to propose candidate tokens, which are selectively verified by a larger target model. While existing methods either adopt multi-draft strategies to increase acceptance rates or block verification techniques to jointly verify multiple tokens, they remain limited by treating these improvements in isolation. In this work, we propose SpecTr-GBV, a novel SD method that unifies multi-draft and greedy block verification (GBV) into a single framework. By formulating the verification step as an optimal transport problem over draft and target token blocks, SpecTr-GBV improves both theoretical efficiency and empirical performance. We theoretically prove that SpecTr-GBV achieves the optimal expected number of accepted tokens for any fixed number of draft sequences, and this bound improves as the number of drafts increases. Empirically, we evaluate SpecTr-GBV across five datasets and four baselines. Our method achieves superior speedup and significantly higher block efficiency while preserving output quality. In addition, we perform comprehensive ablation studies to evaluate the impact of various hyperparameters in the model.

个人简介: 周峰, 中国人民大学统计学院副教授, 中国人民大学“杰出青年学者”, 主要研究领域包括统计机器学习、贝叶斯方法、随机过程、大模型推理加速等, 主持国家自然科学基金青年项目、面上项目, 在JMLR、STCO、ICML、NeurIPS、ICLR、AAAI、KDD等国际期刊和会议上发表论文40余篇, 担任NeurIPS、ICLR、AISTATS等国际会议领域主席, 国际期刊《Statistics and Computing》副主编, 《Transactions on Machine Learning Research》执行编辑, 《Journal of Machine Learning Research》编委, 中国商业统计学会人工智能分会副秘书长、全国工业统计学教学研究会青年统计学家协会第二届理事会理事、IEEE高级会员。

参会人员名单

(按单位字母音序排序)

序号	姓名	工作单位
1	崔彤	安徽财经大学
2	刘德志	安徽财经大学
3	杨海媚	安徽财经大学
4	白皓文	北方民族大学
5	邓祥幸	北方民族大学
6	丁维福	北方民族大学
7	冯娜	北方民族大学
8	高文焕	北方民族大学
9	韩海龙	北方民族大学
10	侯丁匀	北方民族大学
11	蒋文蒙	北方民族大学
12	李俊杰	北方民族大学
13	李任盈	北方民族大学
14	刘婧雅	北方民族大学
15	鲁荆泰	北方民族大学
16	马烨	北方民族大学
17	倪嘉琪	北方民族大学
18	乔荣钰	北方民族大学
19	田宏艳	北方民族大学
20	王博雅	北方民族大学
21	王法	北方民族大学
22	王钱	北方民族大学
23	王倩	北方民族大学
24	吴丽子	北方民族大学
25	姚球	北方民族大学
26	于航志	北方民族大学
27	余娜	北方民族大学
28	朱廷	北方民族大学
29	朱欣怡	北方民族大学
30	方聪	北京大学
31	林伟	北京大学
32	苗旺	北京大学
33	吴磊	北京大学

序号	姓名	工作单位
34	张志华	北京大学
35	李秉翰	北京航空航天大学
36	张慧铭	北京航空航天大学
37	郭旭	北京师范大学
38	吴京桐	北京师范大学
39	刘瑞平	北京信息科技大学
40	夏思薇	成都理工大学
41	左思程	成都理工大学
42	高晓红	楚雄师范学院
43	刘盛	电子科技大学
44	王鹏飞	东北财经大学
45	Xiaofei Wang	东北师范大学
46	刘洋	东北师范大学
47	孙锡吉	东北师范大学
48	玄永屹	东北师范大学
49	郑术蓉	东北师范大学
50	周国玲	东北师范大学
51	虞祯	对外经济贸易大学
52	甘雨露	福建师范大学
53	李小英	福建师范大学
54	连莹颖	福建师范大学
55	罗龙辉	福建师范大学
56	王家琛	福建师范大学
57	王科皓	福建师范大学
58	卫盼怡	福建师范大学
59	熊恒嘉	福建师范大学
60	朱颖	福建师范大学
61	乐沅	福州大学
62	王维贤	广西师范大学
63	陈翌熙	国防科技大学
64	韩世龙	国防科技大学
65	梁伟轩	国防科技大学
66	黄松	哈尔滨工业大学
67	吴仁洪	汉阳大学(韩国)
68	代银	合肥大学
69	李宁	合肥大学
70	刘旭亮	合肥大学

序号	姓名	工作单位
71	牛勇	合肥大学
72	王江丽	合肥大学
73	许守俊	合肥大学
74	闫晓辉	合肥大学
75	杜雲泰	河南大学
76	杨翔宇	河南大学
77	刘新乐	河南理工大学
78	孙明阳	湖州师范大学
79	孔令才	华北电力大学 (保定)
80	於州	华东师范大学
81	葛文秀	华南师范大学
82	李德旺	惠州学院
83	戴晓武	加州大学洛杉矶分校
84	Zerui Tao	理化学研究所 (RIKEN)
85	史成春	伦敦政治经济学院 (LSE)
86	叶锴	伦敦政治经济学院 (LSE)
87	Xi Chen	迈阿密大学 (University of Miami)
88	丁杰	明尼苏达大学
89	刘檬缘	南京师范大学泰州学院
90	卫雨婷	南京信息工程大学
91	高源	南开大学
92	叶源	内蒙古大学
93	刘志宏	普渡大学
94	孙伟	普渡大学
95	孙丹丹	青海大学
96	王金鹏	青海大学
97	从鑫	清华大学
98	高子文	清华大学
99	李慧慧	清华大学
100	李建	清华大学
101	李赛	清华大学
102	毛小介	清华大学
103	叶子怡	清华大学
104	周弘毅	清华大学
105	侯志强	山东财经大学
106	荆文君	山西财经大学
107	Yuan Zhiying	山西大学

序号	姓名	工作单位
108	安柳红	山西大学
109	曹琼	山西大学
110	陈丽妮	山西大学
111	冯梦菲	山西大学
112	胡文静	山西大学
113	贾盈盈	山西大学
114	雒冯凯	山西大学
115	穆昱玫	山西大学
116	张心怡	山西大学
117	赵馨悦	山西大学
118	秦璐	山西师范大学
119	林鎏燚	汕头大学
120	钟家富	汕头大学
121	张良宇	上海财经大学
122	张智恒	上海财经大学
123	周帆	上海财经大学
124	柯蓉	上海对外经贸大学
125	王昭文	上海对外经贸大学
126	尤梦莹	上海对外经贸大学
127	刘方辉	上海交通大学
128	张明娟	上海立信会计金融学院
129	曾子欧	四川大学
130	陈欣	四川大学
131	刘婧睿	四川大学
132	刘伟	四川大学
133	彭小凡	四川大学
134	王子谦	四川大学
135	巫岚	四川大学
136	夏稚坤	四川大学
137	杨芹淼	四川大学
138	张朝瑞	四川大学
139	张晓婧	四川大学华西第四医院
140	何叶	四川师范大学
141	马兰	宿州学院
142	李辉	塔里木大学
143	田芸瑞	塔里木大学
144	赵铭	塔里木大学

序号	姓名	工作单位
145	孙思颖	天津商业大学
146	焦雨领	武汉大学
147	Hao Di	西安交通大学
148	常象宇	西安交通大学
149	范子言	西安交通大学
150	付尧	西安交通大学
151	龚铁梁	西安交通大学
152	蒋云轩	西安交通大学
153	李明芮	西安交通大学
154	束俊	西安交通大学
155	张红英	西安交通大学
156	张泽扬	西安交通大学
157	王嘉欣	西北大学
158	冯慧芳	西北师范大学
159	张建东	西北师范大学
160	wuyunyun	西南财经大学
161	杨婧怡	西南财经大学
162	李晓洁	西南财经大学天府学院
163	彭丽娟	西南交通大学
164	王一帆	西南交通大学
165	Jiang Huilin	西南民族大学
166	柴英明	西南民族大学
167	吕郭庆	西南民族大学
168	马梓业	香港城市大学
169	蔡占锐	香港大学
170	邹荻凡	香港大学
171	代忠祥	香港中文大学 (深圳)
172	胡天阳	香港中文大学 (深圳)
173	桑铭键	新疆财经大学
174	阿依努尔·艾斯图拉	新疆工业学院
175	黄芮	新疆工业学院
176	刘晓芸	新疆工业学院
177	马宏良	新疆工业学院
178	余福宏	新疆工业学院
179	张宇申	新疆工业学院
180	赵鑫	新疆工业学院
181	Yuyang Yao	犹他大学 (University of Utah)

序号	姓名	工作单位
182	李好奇	长江师范大学
183	顾润哲	浙江大学
184	凌晨	浙江大学
185	卢昊宸	浙江大学
186	施玥贝	浙江大学
187	赵冠岚	浙江大学
188	Yujia Sun	浙江农林大学
189	董冉	中国海洋大学
190	龚识瑾	中国科学技术大学
191	李文慧	中国科学院数学与系统科学研究院
192	张婉	中国科学院数学与系统科学研究院
193	Han Wang	中国农业大学 (China Agricultural University)
194	韩雨泽	中国人民大学
195	刘蕾	中国人民大学
196	刘勇	中国人民大学
197	宋昌城	中国人民大学
198	尹建鑫	中国人民大学
199	章明	中国人民大学
200	周峰	中国人民大学
201	黄加罗	中山大学
202	杨云斐	中山大学
203	张维佳	重庆三峡科技大学
204	丁麒皓	重庆邮电大学
205	郭瑞莲	重庆邮电大学
206	明静	重庆邮电大学
207	潘宇	重庆邮电大学
208	邵亚斌	重庆邮电大学
209	王兰	重庆邮电大学
210	许航格	重庆邮电大学
211	周鹏	自然资源部第二海洋研究所
212	宋晓	佐治亚大学
213	李志生	西南财经大学
214	林华珍	西南财经大学
215	董春	西南财经大学
216	郭建军	西南财经大学
217	邹先云	西南财经大学
218	马丹	西南财经大学

序号	姓名	工作单位
219	兰伟	西南财经大学
220	曹占涛	西南财经大学
221	陈雪蓉	西南财经大学
222	成青	西南财经大学
223	戴琳琳	西南财经大学
224	戴明伟	西南财经大学
225	邓蔚	西南财经大学
226	郭斌	西南财经大学
227	李可	西南财经大学
228	李伊	西南财经大学
229	刘斌	西南财经大学
230	刘耀午	西南财经大学
231	马铁丰	西南财经大学
232	马昀蓓	西南财经大学
233	余睿	西南财经大学
234	王粼入	西南财经大学
235	吴量	西南财经大学
236	阎瑶	西南财经大学
237	张晨琳	西南财经大学
238	张佛德	西南财经大学
239	张术林	西南财经大学
240	周凡吟	西南财经大学
241	周岭	西南财经大学
242	唐诗昀	西南财经大学
243	李晓丽	西南财经大学
244	杨森慧	西南财经大学
245	陈铄雳	西南财经大学
246	付康	西南财经大学
247	陈露	西南财经大学
248	程文锐	西南财经大学
249	程喆	西南财经大学
250	邓苏桐	西南财经大学
251	龚梓阳	西南财经大学
252	郭雅丹	西南财经大学
253	何九军	西南财经大学
254	李佳佳	西南财经大学
255	李宪华	西南财经大学

序号	姓名	工作单位
256	梁俊康	西南财经大学
257	刘浩然	西南财经大学
258	马靖凯	西南财经大学
259	马一丹	西南财经大学
260	聂刘泉	西南财经大学
261	申博延	西南财经大学
262	孙珂祎	西南财经大学
263	谭植耀	西南财经大学
264	汤嘉仪	西南财经大学
265	唐立	西南财经大学
266	王蛟龙	西南财经大学
267	王凌锐	西南财经大学
268	王敏	西南财经大学
269	王星	西南财经大学
270	向鹏程	西南财经大学
271	杨昌富	西南财经大学
272	杨丹丹	西南财经大学
273	殷潇怡	西南财经大学
274	张其瑞	西南财经大学
275	张天昊	西南财经大学
276	张雅婧	西南财经大学
277	赵京京	西南财经大学
278	周井然	西南财经大学
279	左永宝	西南财经大学
280	才林	西南财经大学
281	陈龙	西南财经大学
282	陈燃	西南财经大学
283	陈湑	西南财经大学
284	郭立云	西南财经大学
285	李慧玲	西南财经大学
286	刘郅暘	西南财经大学
287	倪艺庭	西南财经大学
288	相慧	西南财经大学
289	肖亚萍	西南财经大学
290	余澜	西南财经大学
291	张舒宇	西南财经大学
292	周正	西南财经大学

校内地图



温馨提示

参会交通指引

校区地址：成都市温江区柳台大道 555 号

地铁到站提示：

- ▶ 4 号线万盛站 C 口：步行至校区南门 / 西门
- ▶ 4 号线杨柳河站 D 口：步行至校区东门（杨柳河站在万盛站前一站）

出发地	最优地铁路线	预估时长	打车参考
天府机场	19 号线→凤溪河→4 号线万盛	100min	90km/1.5h
双流机场	19 号线→凤溪河→4 号线万盛	70min	24km/50min
成都东站	2 号线→中医大省医院→4 号线万盛	75min	43km/50min
成都南站	7 号线内环→文化宫→4 号线万盛	70min	33km/55min

