微软亚洲研究院院刊

Matrix

NO.71
2024年 10 - 12月

微软亚洲研究院2024年度技术大展——人工智能未来之旅

迈向Z级计算：Cloud4Science范式加速科学发现进程

AI²BMD登上Nature，以量子级精度推进蛋白质动力学

01 焦点

02 前沿求索

03 文化故事

04 媒体报道

微软亚洲研究院2024年度技术大展——人工智能未来之旅

2024年，人工智能技术以惊人的速度不断推进，影响的深度和广度也在持续扩展，逐步改变着我们的生活与工作方式。在这一年中，微软亚洲研究院积极探索计算机领域的前沿技术，深度融合技术与现实应用，并始终致力于推动技术与人类社会的和谐共生。与此同时，继北京、上海、温哥华之后，微软亚洲研究院在东京也设立了新的实验室，标志着我们全球科研合作的版图进一步拓展。

岁末年初，我们特别策划了此次技术大展，借此机会与大家一同回顾2024年微软亚洲研究院在人工智能领域的前沿突破与代表性成果。让我们一起领略科技的魅力，展望未来科技的无限可能。

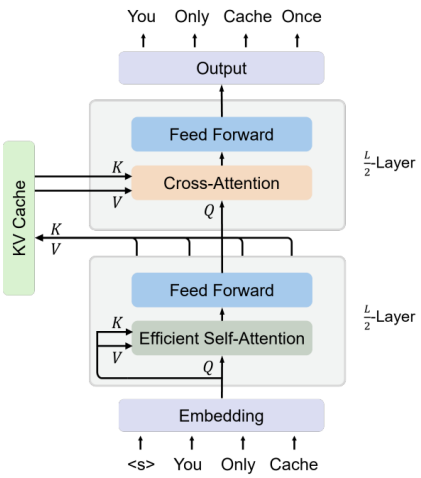
第一展区：创新前沿——探索前沿科技，奠定未来基石

欢迎来到"创新前沿"展区，在这里，你将了解到微软亚洲研究院在基础研究领域的创新突破。这些成果涵盖了模型架构、算法优化、系统网络和数据处理等多个关键领域，不仅拓展了人工智能的能力边界，也为人工智能的未来发展奠定了坚实的基础。

展台A：能力跃迁 - 模型架构创新

随着人工智能的快速发展，传统模型架构的局限性逐渐显现。本展台聚焦于新型模型架构的设计与实现，探索如何通过改变传统模型结构来实现性能飞跃。

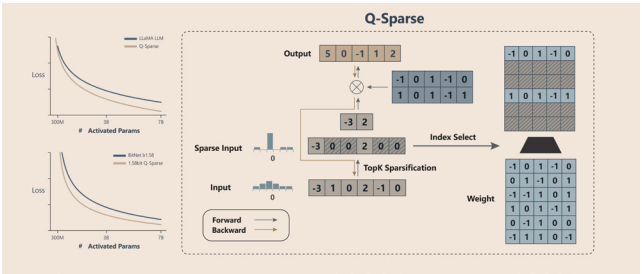
- **全新 Transformer 架构 DIFF Transformer**——通过差分注意力机制，增强对关键信息的关注以及减少对噪声的干扰，在多项语言任务中取得了显著优于Transformer模型的性能提升。
- **Decoder-Decoder架构YOCO** ——打破传统Decoder-only架构，通过自解码器和交叉解码器的独特架构，显著降低GPU内存使用，实现"模型越大，内存越省"。



展台B：效能加速 – 低比特量化与高效推理

为了使人工智能模型能够高效运行并显著降低推理成本，微软亚洲研究院深入探索了低比特量化技术和推理优化方法，希望实现低成本、高效率和低能耗的目标，从而使人工智能像水和电一样，真正成为人类社会的基础设施。

- **1比特大语言模型BitNet b1.58**——以独特的三值 {-1, 0, 1} 表示参数，在速度、内存使用、吞吐量和能耗等方面具有大幅优势，同时保持了较高的模型性能。bitnet.cpp则是用于 BitNet b1.58 等1-bit LLMs 的官方推理框架。
- **完全激活LLMs稀疏性的方法Q-Sparse**——实现大语言模型激活的完全稀疏性，与 BitNet 正交且互补，可以为大语言模型推理中的数据类型提供全面优化，降低推理阶段的计算成本和内存占用。



- **深度学习训练框架nnScaler**——通过一套并行化原语和策略限定搜索的方法来寻求最佳的并行策略组合，有效应对了当前深度学习训练效率的难题，为未来的并行策略研究提供了新的方向和工具。

展台C：认知扩展 – 长文本与外部知识处理

随着大模型在各个领域的广泛应用，让这些模型更好地利用外部知识以及处理超长上下文面临着新的挑战。本展台展示了如何提升大模型处理复杂信息的能力。

- **超长上下文窗口LongRoPE**——利用精细化非均匀位置插值和渐进式扩展策略，将大模型的上下文窗口扩展至2048k，大幅提升了长文本的处理效果。
- **RAG任务分类法**——将查询需求分为显式事实、隐式事实、可解释的推理、隐式推理四个层级，帮助大模型更准确地理解和利用专有领域知识，减少生成幻觉和虚构信息的风险，提升模型在专业领域问题处理中的可靠性。

展台D：未来基座 – 系统与基础设施革新

大模型时代，传统的计算机系统面临着前所未有的挑战与机遇，微软亚洲研究院从超级计算机系统创新、云计算和分布式系统三大方向实现了计算机系统的自我革新，为未来人工智能的发展构建基础"底座"。

- **弥合大模型低比特量化与终端部署间的鸿沟**—— 数据编译器 Ladder 和算法 T-MAC，使当前只支持对称精度计算的硬件能够直接运行混合精度矩阵乘法；新硬件架构 LUT Tensor Core 利用查找表方法，推动下一代硬件加速器原生支持混合精度矩阵乘法。
- **统一化数据库技术**—— VBase、SPFresh和OneSparse系统，为大模型提供了坚实的数据管理和查询能力，提升向量数据库在执行复杂查询时的性能。
- **Cloud4Science范式**——深度融合云计算、人工智能和高性能计算技术，重塑科学计算模式，为科学智能提供更加灵活、高效且可扩展的解决方案，加速科学发现的步伐。

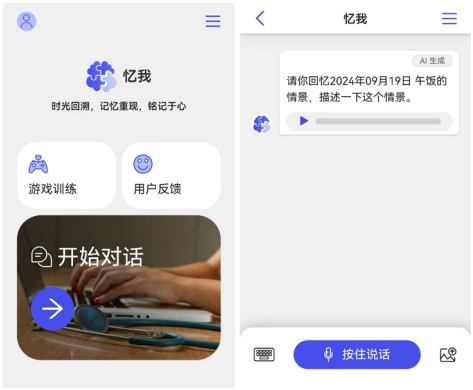
第二展区：智能视界——拓展智能边界，驱动现实变革

在"智能视界"展区，我们将展示人工智能技术如何跨越理论研究，对现实世界产生影响。从健康福祉到加速科学发现，从可持续发展到产业应用，微软亚洲研究院及其合作伙伴的创新成果正助力推动社会进步，改善人们的生活质量。

展台A：守护健康 - 提升医疗精准度与个性化护理

在医疗健康领域，早期诊断和个性化治疗对于提高患者的治疗效果与生活质量至关重要。本展台展示了如何利用多模态大模型和先进的数据分析工具，辅助医护人员与病患更早地识别疾病迹象，及时进行干预和治疗。

- **个性化认知训练框架"忆我"（ReMe）**——微软亚洲研究院与上海市精神卫生中心联合打造的认知训练工具，能够支持文字、图像、语音等多种输入输出方式，为认知障碍患者提供个性化的记忆训练体验。

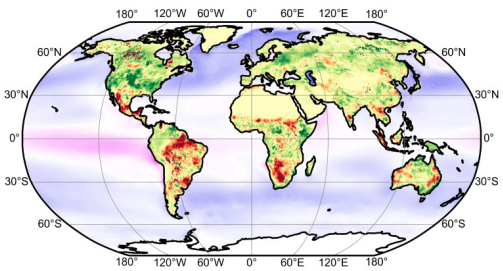


- **基于图的数据处理算法AdaMedGraph**——利用图神经网络（GNNs）构建，可自动选择重要特征来预测帕金森病患者的病情发展，为个性化治疗提供数据支持。

展台B：绿色先锋 - 可持续发展的AI解决方案

在全球气候变化日益严峻的背景下，有效管理和减少碳排放成为各国政府和企业面临的重大挑战。AI技术可以通过智能化手段优化资源管理，减少环境污染，助力实现全球气候目标。

- **电动汽车电池衰退预测模型**——微软亚洲研究院与日产汽车共同开发的全新的机器学习预测方法，将电池退化预测准确率的平均误差控制在了0.0094，为电池的高效回收提供了有力依据。
- **近实时的全球碳预算**——微软亚洲研究院联合清华大学和法国原子能署气候与环境科学实验室，基于AI技术将全球碳预算的时间从滞后1年缩短至3个月，为可持续发展相关的科学研究与政策制定提供了更加及时的数据支持。



展台C：科学探索引擎 - 上知天文下知地理

本展台展示了人工智能如何为天文学和地质学领域的科研人员带来了前所未有的便利和支持。创新的人工智能技术手段与研究模式，为科学研究提供了新的思路与方法。

- **人工智能“天文学家”Mephisto**——微软亚洲研究院联合清华大学天文系和俄亥俄州立大学开发的大语言模型智能体，可用于深入分析詹姆斯·韦布空间望远镜观测到的高红移星系，为宇宙诞生之初的“小红点”现象提供了可能的解释。
- **地质图理解基准集GeoMap-Bench和智能代理GeoMap-Agent**——首个用于地质图理解的基准集与智能代理，利用多模态大语言模型自动读取、分析、解读地质图，有望大幅提高读图效率和准确性，为地质领域相关人员提供便利。

展台D：产业赋能 – 基础模型在产业中的应用

快速发展的人工智能技术将如何在不同产业中释放其巨大的应用潜力和商业价值？微软亚洲研究院希望通过持续预训练将产业数据智能融入大语言模型中，创造出具备跨领域通用能力的新型模型，成为产业界中最有价值的任务——精准预测、高效决策和智能化工业模拟的关键工具。

- **通用金融市场模拟引擎MarS**——基于大市场模型LMM，为金融研究人员提供定制化生成式模型解决方案，构建适用于金融市场所有下游任务的生成式基座模型应用新范式，为金融行业带来更高的效率和更精准的市场洞察。
- **生成式表数据学习框架GTL**——成功将多行业数据智能相关的知识融入大语言模型，使其具备在新领域、新数据及新任务上的直接迁移和泛化能力，从而更加敏捷地响应不同的产业需求。
- **自动化研发工具 RD-Agent**——借助大语言模型的强大能力，整合数据驱动的研发系统，实现研究与开发流程的自动化，提高研发效率，促进跨领域知识迁移与创新。

第三展区：共生未来——促进人机协作，共创和谐未来

一直以来，微软亚洲研究院始终致力于利用技术创新促进实现人工智能与人类社会的和谐共生。”共生未来”展区将为你呈现人机交互领域的前沿进展，探索如何让技术更好地服务于人类，并从人类智慧中汲取灵感，以推动技术与社会的共同进步，创造更加智能、和谐和可持续的未来。

展台A：多模态生成 – 激发创意与表达

多模态技术结合了文本、图像、音频和视频等多种数据形式，为用户提供了丰富和自然的交互体验。微软亚洲研究院开发了一系列工具和技术，旨在提升了人们的创意表达能力，为创作、教育、娱乐等多个行业带来了新的可能性。

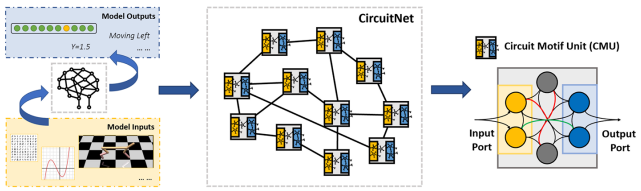
- **虚拟头部视频生成框架VASA-1**——只需一张肖像、一段音频和一些可选的信号控制，即可实时生成具有精确唇音同步、逼真面部行为和自然头部运动的逼真的说话视频，为实时交互和虚拟角色创建提供新工具。

- **语音大模型VALL-E 2**——通过重复感知采样和分组编码建模技术，突破了语音稳健性、自然度和说话人相似度方面的界限，让零样本 TTS 性能与人类水平相近，提供了更为丰富与个性化的语音交互体验。

展台B：脑启发设计 - 人工智能的进化之路

微软亚洲研究院正在探索新型人工智能架构——脑启发式人工智能，从神经元、网络层和更高级别的系统层出发，推动人工智能网络向着更低功耗、更高效率、更好性能的方向发展，同时也为具身智能发展提供了理论和方法。

- **神经回路网络 CircuitNet** ——借鉴大脑神经元局部密集和全局稀疏连接的特性，模拟包括反馈和侧向模式在内的多种神经元连接模式，在达到与其它神经网络相同性能的同时，具有相当或更少的参数。



- **用于时间序列预测任务的 SNN 框架**——充分利用脉冲神经元在处理时间序列信息上的高效性，成功实现了时间序列数据与SNN之间的时间同步，为SNN领域提供了一个既节能，又符合生物学原理的时间序列预测新方案。
- **贝叶斯行为框架**——基于变分贝叶斯方法，引入贝叶斯“意图”变量，成功连接习惯性行为和目标导向行为，为具身智能的发展提供了坚实的理论指导框架。

展台C：Societal AI – 构建负责任的智能未来

随着大模型逐渐融入人们的日常生活，其带来的重大技术和社会挑战日益凸显。微软亚洲研究院将社会责任人工智能（Societal AI）作为重点研究方向之一，通过与社会科学领域的专家紧密合作，确保人工智能沿着对社会负责的方向积极发展。

- **BaseAlign 对齐算法**——基于施瓦茨人类基本价值理论的大模型价值对齐算法框架 ，已在多元价值观场景下验证了其可行性。
- **价值观自适应框架CLAVE** ——通过结合两个互补的大语言模型，以少量人工标注数据实现高效的价值观对齐评估。
- **文化差异理解框架CultureLLM** ——利用世界价值观调查

（WVS）作为种子数据，通过语义增强生成训练数据，显著提升模型文化相关任务上的表现。

特别展区：高光时刻

这一展区汇聚了2024年微软亚洲研究院的科技高光时刻。来自学术界和产业界的至高荣誉是对微软亚洲研究院的科研人员和研究工作的肯定与认可，并将继续激励我们不断创新，引领技术变革。

时间检验奖（Test-of-Time Awards）：这里记录的是那些经受住时间考验、对计算机研究领域产生深远影响的经典成果。

- ArrayTrack: A Fine-Grained Indoor Location System 荣获 ACM MobiCom 2024时间检验奖

最佳论文奖（Best Paper Awards）：这些研究工作代表计算机研究的前沿与突破，为未来的技术发展奠定坚实的基础。

- Rho-1 荣获NeurIPS 2024最佳论文 Runner-Up 奖
- Autothrottle 荣获 NSDI 2024 杰出论文奖
- RetrievalAttention 荣获NeurIPS ENSLP 2024 最佳论文奖
- SuperBench 荣获USENIX ATC 2024 最佳论文奖
- VisEval 荣获IEEE VIS 2024 最佳论文奖
- Exploring the Feasibility of Remote Cardiac Auscultation Using Earphones 荣获 ACM MobiCom 2024 最佳论文奖
- ConvStencil 荣获 PPoPP 2024 最佳论文奖
- LayoutLM 荣获 2024国际基础科学大会 Theoretical Computer & Information Sciences - Big Data Technology 前沿科学奖

个人荣誉：向那些在科研、工程和社会贡献方面表现卓越的同事们表示祝贺。

- 郭百宁获选 2024加拿大皇家学会(RSC)院士
- 谢幸获选 2023 ACM Fellow
- 松下康之被评为 2025 IEEE Fellow
- 李东胜入选 2023年中国智能计算创新人物
- 傅建龙入选 2023年中国智能计算创新人物

尾声

感谢参观本次展览！我们希望这不仅是一次技术探索之旅，也能激发你对未来科技的的无限遐想。在科技飞速发展的今天，微软亚洲研究院将持续保持开放与合作的心态，深耕计算机基础及应用研究，与社会各界携手塑造一个更加智能、和谐的未来。

祝愿各位在2025年的科研探索中收获更多精彩与成长！期待在未来的人工智能旅程中与你相遇。

扫描二维码查看完整版
“微软亚洲研究院2024年度技术大展”

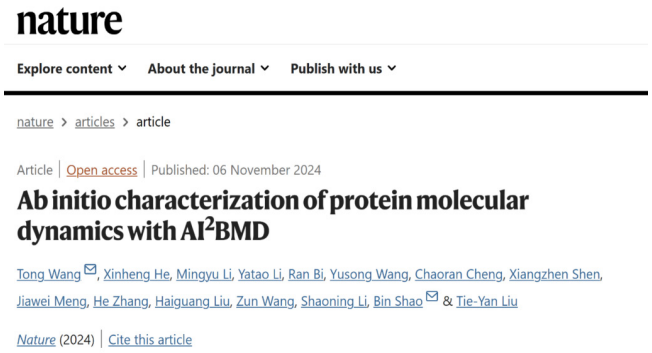
AI²BMD登上Nature，以量子级精度推进蛋白质动力学

随着人工智能在蛋白质研究中的重要性日益提升，预测静态的蛋白质晶体结构已不再是难题。然而，如何在原子级别精确刻画蛋白质动态变化仍是一项亟需解决的挑战。微软研究院科学智能中心王童研究员及其团队，历时四年研究推出的基于 AI 的分子动力学模拟系统 AI²BMD，对蛋白质等生物大分子进行量子级精度的全原子模拟，实现了比经典模拟更高的精度，和比密度泛函理论更快的速度，为包括生物分子建模等在内的生物研究提供了新的可能性。

“所有生物体的行为都可以通过原子的颤动和摆动来理解。”正如诺贝尔物理学奖得主 Richard Feynman 的名言所说，生物世界的本质是原子永不停歇的运动过程。探索生物分子的运动过程以及分子之间的相互作用，对于破译生命活动背后的机理以及设计和发现新的药物、疫苗以及生物材料都至关重要。

近年来，随着深度学习技术的发展和 GPU 算力的飞速提升，人工智能在蛋白质研究领域扮演着越来越重要的角色。2024年的诺贝尔化学奖就授予了蛋白质结构预测和蛋白质设计的研究。尽管通过计算手段预测静态的蛋白质晶体结构已经接近或达到生物学实验解析的精度，但如何利用 AI 在原子级别精确地刻画蛋白质动态变化的行为是一项仍需解决且更为困难的挑战。

日前，微软研究院科学智能中心（Microsoft Research AI for Science）王童研究员及其团队，历时四年在人工智能驱动的分子动力学模拟研究中取得重要进展，其成果已通过长文（Article）的形式在世界顶级科学杂志《自然》（《Nature》）正刊在线发表。


nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Open access | Published: 06 November 2024

Ab initio characterization of protein molecular dynamics with AI²BMD

Tong Wang ✉, Xinheng He, Mingyu Li, Yatao Li, Ran Bi, Yusong Wang, Chaoran Cheng, Xiangzhen Shen, Jiawei Meng, He Zhang, Haiguang Liu, Zun Wang, Shaoning Li, Bin Shao ✉ & Tie-Yan Liu

Nature (2024) | Cite this article

AI驱动的分子动力学模拟

分子动力学（Molecular Dynamics，简称 MD）是模拟分子和原子在真实生物细胞中运动的技术手段。动力学模拟一般以1飞秒（10-15秒）为一步模拟，通过数亿以至数千亿步的模拟，反映细胞中蛋白质分子的时空运动过程。历经半个多世纪的发展，分子动力学模拟可以分为两类：经典模拟（Classic MD Simulation）和量子模拟（Quantum Simulation）。

经典模拟以牛顿力学作为力场来驱动原子和分子的运动，具有速度快、适用性广等特点。半个多世纪以来，经典模拟被广泛应用于蛋白质等生物大分子的动态研究中，并于2013年获得诺贝尔化学奖。然而，采用牛顿力场的经典模拟，力场的准确性欠缺，且无法模拟分子成键断键等电子迁移的行为，在高精度的自由能计算、药物虚拟筛选、生物化学反应等方面捉襟见肘。

与经典模拟相对的是以密度泛函理论（Density Functional Theory，简称 DFT）为代表的量子模拟方法，该方法采用量子力学力场，对原子的运动描述可达到从头计算的精度。凭借其完备的理论基础和在计算化学领域的广泛应用，1998年密度泛函理论获得了诺贝尔奖。但由于其极高的计算代价，量子模拟既无法直接应用于蛋白质等生物大分子的研究，又无法进行长时间的模拟仿真。

如何打破经典模拟和量子模拟之间的技术瓶颈，实现对蛋白质等生物大分子量子级精度的全原子模拟，是该领域半个多世纪以来的一大挑战。

为了解决这一重大挑战，微软研究院科学智能中心的研究员们设计了基于 AI 的分子动力学模拟系统 AI²BMD (AI powered ab initio biomolecular dynamics)。该系统以从头计算的精度（即量子级的精度）高效地对各类蛋白质进行了全原子模拟仿真。这一创新在生物分子模拟中实现了一种此前标准模拟技术无法达成的权衡——比经典模拟具有更高的准确性，其计算成本虽然高于经典模拟，但计算速度领先 DFT 和其他量子力学方法数个数量级。AI²BMD 有望在生物分子建模中解锁更多新的能力，特别是在如蛋白质与药物相互作用这种需要进行高精度计算的研究过程中。

深入AI²BMD技术创新

分子动力学模拟最重要的组件之一是力场的构建。在模拟的每一步中，力场计算分子的能量和每个原子所受的力，从而驱动整个分子的运动。经典模拟采用牛顿力场，量子模拟采用量子力学力场。要构建 AI 驱动的分子动力学模拟，最大挑战是深度学习模型的泛化性，即在已知分子上训练的模型对各类未知蛋白质分子的能量和力的预测准确性。为此，研究团队设计了一种基于蛋白片段的、可泛化的分割技术，将各类蛋白质分子分割成21种通用的蛋白质片段。数据集的构建和模型的训练全都基于通用蛋白质片段进行，从而实现对各类蛋白质分子的通用解决方案。

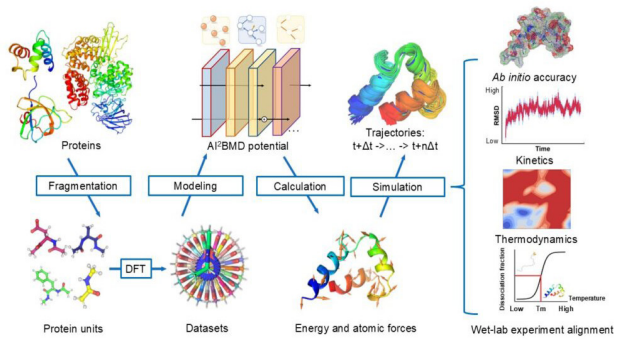


图1：AI²BMD 技术流程图

基于蛋白质通用分割方案，研究团队进一步构建了包含二千多万条数据、目前世界上最大的量子级精度蛋白质片段数据集 Protein Unit Dataset（https://github.com/microsoft/AI2BMD）。研究员们选取了此前研发的通用分子几何结构建模的网络模型 ViSNet，并在 Protein Unit Dataset 上对其进行训练，来作为 AI²BMD 的力场。考虑到分子模拟的效率问题，研究团队提出了一种全新的主从式架构（client-server），通过对 CPU 和 GPU 的动态调度，该架构可以将每步模拟时间压缩至数十毫秒量级。研究员们利用 AI²BMD 对各类蛋白的动力学和热力学进行了分析，分析结果展现了比经典模拟在蛋白质折叠自由能计算、构象空间探索等多个方面更好的结果。

生物分子模拟的技术创新

AI²BMD 在如下几个方面展示了与此前蛋白质分子经典模拟不同的创新性变化：

量子级精度：AI²BMD 通过可泛化"机器学习力场"——一种通过机器学习模型构建的原子和分子之间相互作用的模型，实现了量子级精度的全原子蛋白质动力学模拟。

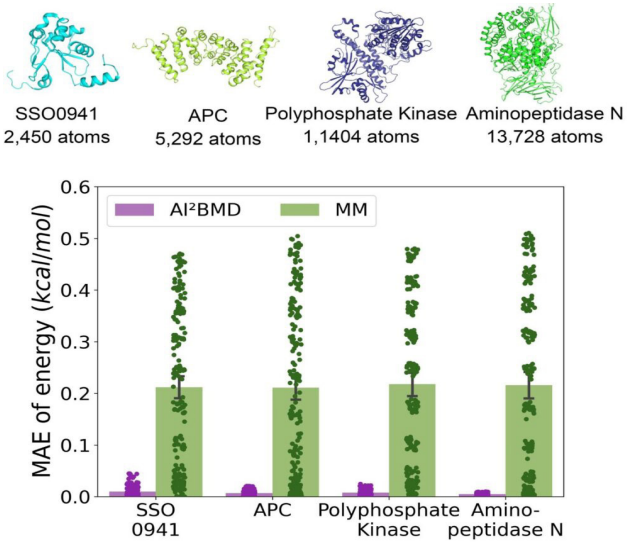
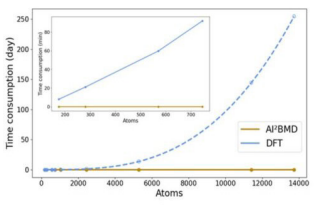


图2：AI²BMD 和经典动力学模拟对不同蛋白质能量计算的误差对比

泛化性：AI²BMD 首次解决了机器学习力场在模拟蛋白质动力学方面的泛化挑战，展示了对各种蛋白质全原子模拟的鲁棒性。

全原子模拟的兼容性：相比于结合量子模拟和经典模拟的混合模拟技术，AI²BMD 将量子级精度的计算拓展到了整个蛋白质分子上，且不需要任何关于蛋白质的先验知识。这消除了蛋白质的量子模拟和经典模拟计算之间潜在的不兼容性，并将量子模拟区域的计算速度提高了几个数量级，使全原子蛋白质的近从头计算更接近现实。因此，AI²BMD 为许多下游应用铺平了道路，并为表征复杂生物分子动力学提供了新的视角。

高效性：AI²BMD 比 DFT 和其他量子模拟的速度快几个数量级。AI²BMD 支持超过1万个原子的蛋白质的量子级精度计算，使其成为众多学科领域中最快 AI 驱动的分子动力学模拟程序之一。



Protein	Atom number of protein	Atom number of system	AI²BMD (s)	DPMD (s)	Allegro (s)
Chignolin	175	4,715	0.047	**0.040**	0.238
Trp-cage	281	6,067	**0.052**	0.055	0.322
WW domain	571	10,678	**0.070**	0.095	0.626
ABD	746	11,793	**0.085**	0.106	0.712
PACSIN 3	1,040	17,923	**0.106**	0.162	-
SSO0941	2,450	44,401	**0.213**	0.414	-
APC	5,292	54,999	**0.449**	0.580	-
Polyphosphate Kinase	11,404	97,657	**0.966**	-	-

图3：AI²BMD 与 DFT 以及其他 AI 驱动的动力学模拟软件速度的比较

构象探索的多元性：不同于经典模拟，AI²BMD 不会对键长、键角、二面角等施加任何约束。如图4，在用 AI²BMD 和经典模拟分别模拟蛋白质折叠和去折叠的过程中，AI²BMD 探索了经典模拟无法检测到的更多可能的构象空间。因此，AI²BMD 为研究药物靶标结合过程中蛋白质的柔性运动、酶催化、变构调节、内在无序蛋白等提供了更多的机会和可能。

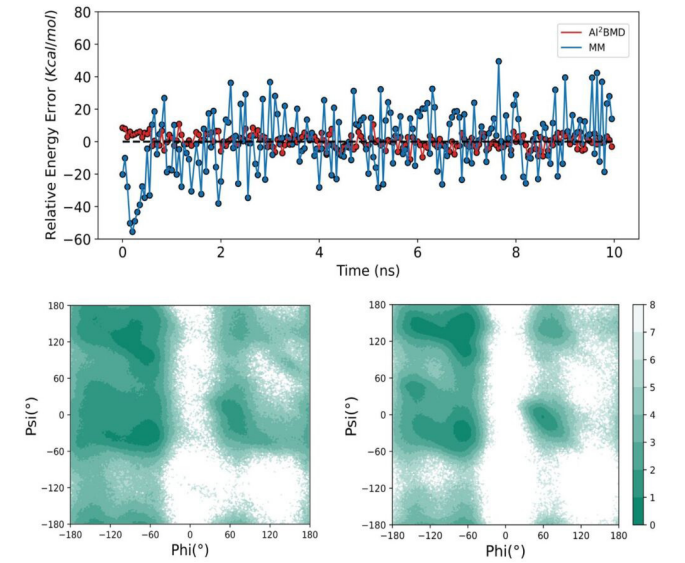


图4：AI²BMD 与经典模拟在蛋白 Chignolin 折叠过程的模拟表现

生物实验的一致性：与经典模拟和混合模拟相比，AI²BMD 在J-耦合、焓变、热容、折叠自由能、熔化温度和 pKa 等指标上都展现出了与生物学实验更高的一致性。

应用与展望

在生物分子模拟中实现量子级精度是极具挑战性的，但它在揭示生物系统的奥秘以及设计新型生物材料和药物方面具有巨大潜力。这一突破证明了 AI for Science 的远见，即利用人工智能的能力革新科学探索。AI²BMD 实现了机器学习力场在分子动力学模拟应用中准确性、稳定性和泛化性等方面的平衡，在提升能量和原子受力计算精度的同时，AI²BMD 也带来对蛋白质各类性质更为准确的计算和估计。

AI²BMD 一个关键的应用场景是药物发现中高精度的靶点蛋白和药物分子之间的结合能计算。在2023年首届全球人工智能药物开发竞赛中，AI²BMD 和其 AI 力场 ViSNet 准确识别出了与新冠病毒多个靶点相结合的潜在药物分子，在所有任务中都取得了最佳预测，赢得了冠军。

2022年，微软研究院还与全球健康药物研发中心（Global Health Drug Discovery Institute，简称 GHDDI）展开合作，将人工智能技术应用于药物设计。GHDDI 是盖茨基金会、北京市政府和清华大学联合成立的非营利机构，旨在研发用于治疗对中低收入国家（LMIC）造成严重影响的结核病和疟疾等疾病的药物。微软研究院正在与 GHDDI 密切合作，希望通过 AI²BMD 和其他人工智能技术加速药物发现过程。

AI²BMD 不仅推进了对科学问题的研究，还促进了药物发现、蛋白质设计和酶工程等领域的新的生物医学研究。利用 AI²BMD 准确、高效地表征蛋白质的动态特性正在推动科学技术创新发展，激发科学界对生物机理探索的广泛兴趣。

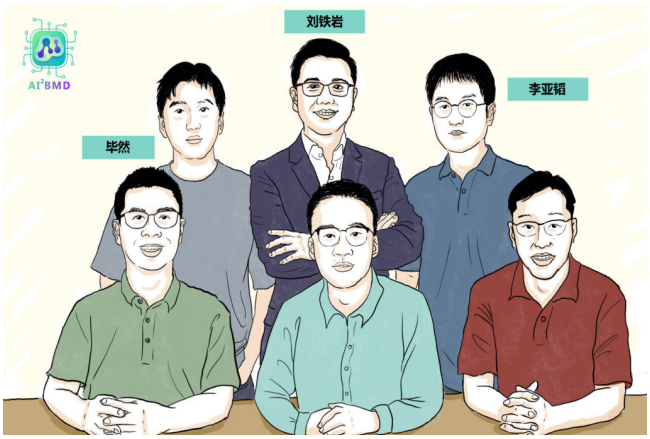


图5：AI²BMD 研究团队主要成员

相关链接：

Ab initio characterization of protein molecular dynamics with AI²BMD
https://www.nature.com/articles/s41586-024-08127-z

GitHub 链接：https://github.com/microsoft/AI2BMD

相关阅读：

对话《Nature》论文作者，揭秘AI²BMD背后的故事
AI²BMD 研究过程中面临了哪些挑战？研究团队又是如何在四年内攻克生物分子模拟领域的长期难题的？来自 AI²BMD 研究团队的微软研究院高级研究员王童、首席研究员刘海广、高级工程师毕然向大家分享了研究背后的故事。

扫描二维码
了解 AI²BMD 背后的故事

迈向Z级计算：Cloud4Science范式加速科学发现进程

传统超级计算机作为科学计算的核心支柱，在推动技术进步方面发挥了不可替代的作用，但随着科学智能时代下需求的多样化和复杂化，其扩展性和能效的局限逐渐显现。针对这一挑战，微软亚洲研究院的研究员提出了 Cloud4Science 的新范式，以云计算、人工智能和高性能计算技术的深度融合为核心，重新定义科学计算的架构，加速科学智能的研究进展。在此框架下，研究员们已对关键科学计算算法如 Stencil、FFT、SpMV 等进行了优化，并成功开发了一系列创新算法，为科学家利用云计算及人工智能平台进行科学计算和研究开辟了新的途径。相关工作已连续发表在 SC、PPoPP 等高性能计算与并行计算领域顶会，并获得了 PPoPP'24 唯一最佳论文奖。

在刚刚落幕的国际超算大会 SC'24 上，最新揭晓的戈登贝尔奖获奖应用成功突破了 E 级计算的瓶颈，标志着超级计算机应用正式迈入下一个关键阶段——万 P 级计算（每秒千亿亿次浮点运算）。作为高性能计算（HPC）的巅峰代表，超级计算机长期以来一直是推动科学和技术进步的重要力量。

科学计算作为超级计算机的核心应用领域，利用其强大的计算能力，通过数值模拟、数据分析和数学建模，旨在解决科学、工程和技术中的复杂问题，在揭示自然规律、预测未知现象以及推动技术创新中发挥着不可或缺的作用。

然而，随着科学智能（AI for science）时代的到来，超级计算机在追求更高性能的同时，也面临着一些新的挑战：

- 架构碎片化：各超算系统采用不同的硬件架构和编程模型，科学应用需要复杂的定制化适配才能运行。这不仅限制了科学应用的多样性，还难以兼顾传统科学计算与智能计算的双重需求。
- 开发难度高：科学智能时代强调多学科、多技术领域的交叉与协作。不同的超级计算机架构不仅增加了软件开发和维护的复杂度，开发者还需要不断重新学习并掌握跨领域的专业知识，阻碍了科学研究的灵活性和快速推进。
- 能耗与成本压力：当前 E 级超算每年耗电可达上亿度，未来 Z 级超算能耗可能更高。同时，系统更新换代成本巨大，应用需重新设计和部署，进一步增加了科研投入的时间和经济成本。

"传统科学计算的优势在于数值求解，通过高精度计算模拟复杂的物理过程。然而，随着问题规模的快速扩大和计算复杂度的持续攀升，单纯依赖数值求解的模式难以充分释放未来万 P 级甚至 Z 级超算的全部潜力。"微软亚洲研究院高级研究员李琨表示，"科学计算正在从传统数值求解向融合知识推理的科学智能转型。通过将高性能计算技术与未来的 Z 级算力结合，全面支撑科学智能时代对极限计算和智能推理的双向扩展需求，才会为更多突破性发现提供全新的可能性。"

Cloud4Science范式加速科学计算进程

为了应对这些挑战，微软亚洲研究院的研究员提出了 Cloud4Science 范式，希望通过融合现有的云基础设施、人工智能和高性能计算技术，重塑科学计算的模式。这一范式为传统超算范式提供了有效的补充，也为科学智能提供了一种更加灵活、高效且可扩展的解决方案。

"Cloud4Science 范式通过将科学计算任务迁移到云平台或人工智能架构上，实现了计算架构的统一，降低了科学计算的访问门槛。"微软亚洲研究院首席研究员曹婷表示，"这使得科研人员能够在单一平台上使用多种算法和应用，同时，云平台和人工智能的强大算力也将大幅提升科学计算效率，为未来的科学研究与计算应用开辟新的可能性。"

为了实现 Cloud4Science 范式，研究员们计划分两个阶段来推进：

图1: Cloud4Science融合云计算、AI 与高性能计算，驱动科学智能新未来

第一阶段是以问题为导向，从算法角度对传统科学计算进行迁移，确保这些计算任务能够在云计算或人工智能硬件架构上顺利运行。这一阶段的核心任务是将经典的科学计算算法，如 Stencil、FFT（快速傅里叶变换）、SpMV（稀疏矩阵-向量乘法）等，转换为基于矩阵乘法的计算模式，以便充分利用云计算和人工智能的强大计算能力。通过这一转化，传统科学计算算法的性

能得以显著提升，同时大幅降低了科学应用对硬件适配的复杂性，并为下一步科学计算的智能化奠定了基础。

第二阶段的目标是推动科学计算算法与人工智能的深度融合。传统的科学计算算法更注重数值计算，而科学智能则强调推理能力的提升。科学计算模型与大语言模型虽然在某些方面可以互相借鉴，但二者之间存在显著差异。科学计算模型通常包含大量的物理信息和生物信息，这些专业知识需要被有效地整合到算法设计中。因此，这一阶段的任务是设计融合传统科学计算模型与人工智能技术的创新解决方案，通过人工智能技术有效整合领域知识、生成洞见并促进科学创新，充分利用云原生和人工智能原生架构的优势，进一步推动 Cloud4Science 范式的发展。

传统科学算法向云计算与人工智能硬件无缝迁移

目前，第一阶段的研究目标已经基本完成，即实现传统科学计算算法向云计算和人工智能硬件的无缝迁移。研究员们从 Stencil 算法入手，设计了全新的算法 Jigsaw 和 ConvStencil，将 Stencil 算法向量化并重新张量化成矩阵乘法模式，使 Stencil 算法成功映射到 Tensor Core 等人工智能加速器硬件单元。随后，研究员们又引入了人工智能驱动的低秩适应（Low-Rank Approximation，LoRA）技术，进一步优化 Stencil 性能，推出了 LoRAStencil 以及融合三种经典算法的 FlashFFTStencil，这些创新让多种科学计算算法能够更高效地部署在人工智能加速单元上，实现性能的显著提升并同时降低了硬件适配的复杂性。

扩展矩阵计算边界，连接科学与 AI 硬件

为突破科学计算的性能瓶颈，研究员们提出了 ConvStencil [1]，通过将传统的科学计算算法映射为矩阵乘法，进一步扩展了矩阵计算的应用边界，为科学计算与 AI 硬件的高效协同奠定了坚实基础。基于 Stencil 算法与人工智能领域广泛应用的卷积计算模式有着相似之处，研究员们专门开发了一套针对 GPU Tensor Core 的优化算法，使得其能够充分利用 Tensor Core 强大的矩阵计算能力。通过引入布局转换与冲突消除机制，ConvStencil 不仅显著提升了科学计算与云计算及人工智能硬件的兼容性，还促进了科学计算从传统的 CPU 计算向现代 GPU 计算的顺利过渡。

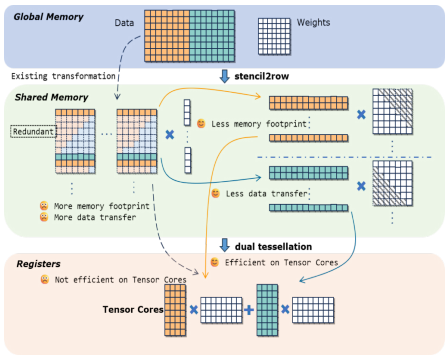


图2：基于矩阵乘法ConvStencil 计算系统（PPoPP'24 唯一最佳论文奖）

为了实现内存访问效率的大幅提高，研究员们在 ConvStencil 的基础上设计了 LoRAStencil [4]，通过融入 LoRA 技术，巧妙地结合了数据的低秩特征与计算需求。利用分解权重矩阵，优化数据的加载与复用过程，LoRAStencil 有效减少了不必要的内存访问，解决了维度残差问题。实验评估显示，LoRAStencil 相比现有技术，性能提升最高可达2.16倍。LoRAStencil 为在 Tensor Core 单元上实现高效的张量化 Stencil 计算开辟了新的途径，使其在科学计算中能发挥更大作用。

尽管 Tensor Core 单元在处理人工智能任务时表现出色，但在处理如 Stencil 这样涉及大量稀疏数据的高性能计算算法时，仍面临计算资源利用率不高和内存带宽受限的问题。为了解决这些挑战，研究员们创造性地将 Stencil、FFT 和矩阵乘法三种经典科学计算算法融为一体，提出了更为高效的 FlashFFTStencil 计算系统 [3]。实验结果证实，FlashFFTStencil 实现了无稀疏性的边界转换，其性能较现有最先进的技术平均提升了2.57倍。FlashFFTStencil 在实现了多种科学计算算法统一的同时，还成功地将这些算法与 Tensor Core 单元等先进的人工智能硬件连接起来，为科学计算的未来发展提供了新的可能性。

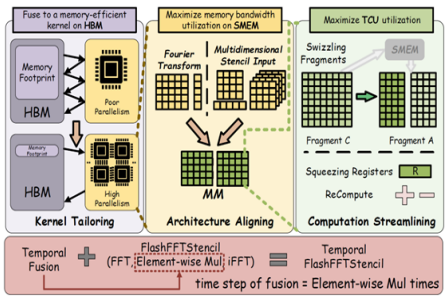


图3：基于全稠密矩阵计算的 FlashFFTStencil 系统

时空数据向量对齐，提升 CPU 计算效率

Jigsaw 算法 [5]专注于 Stencil 算法的向量化，通过采用基于通道的蝶形向量化、基于奇异值分解的维度展平（SVD-based Dimension Flattening）技术以及基于迭代的时间合并策略，有效解决了空间和时间维度上的数据对齐冲突（Data Alignment Conßict, DAC）问题，大幅提升了科学计算在 CPU 上的效率。实验结果显示，在多种测试环境中，Jigsaw 相对于当前最先进的技术平均实现了2.31倍的加速效果，适用于广泛的 Stencil 内核。

在此基础上，研究员们还对另一种重要的科学计算算法——稀疏矩阵-向量乘（Sparse Matrix-Vector Multiplication, SpMV）进行了深入优化，提出了 VNEC 算法 [6]。这是一种创新的 SpMV 存储格式，旨在优化数据局部性和向量化操作，同时缓解现有算法的局限性。VNEC 通过剔除冗余列和改进数据局部性，大幅度减少了内存访问开销，增强了向量计算的效率。实验表明，在多核处理器环境下，VNEC 在 x86 CPU 上相较于标准 MKL SpMV 例程最高实现了6.94倍（平均2.10倍）的加速，在 ARM CPU 上的加速比最高可达5.92倍（平均1.73倍）。由于 VNEC 格式转换的预处理成本较低，特别适用于实际的迭代应用场景，展现

出了极高的实用价值。

Cloud4Science 范式在量子化学中的实践探索

为了验证 Cloud4Science 范式能否为科学计算带来更好的性能提升，微软亚洲研究院的研究员们与微软研究院科学智能中心（Microsoft Research AI for Science）团队合作，共同开发了一种端到端的优化编译器 EPT（Elastic Parallel Transformation）[2]。利用弹性并行转换技术，EPT 可以把传统的科学计算算法，特别是从头算量子化学计算，自动适配至 GPU 架构。因此，EPT 能够将复杂的量子化学问题分解为适合并行处理的单元，优化任务的划分粒度，并生成专为 GPU 架构优化的高效计算内核。

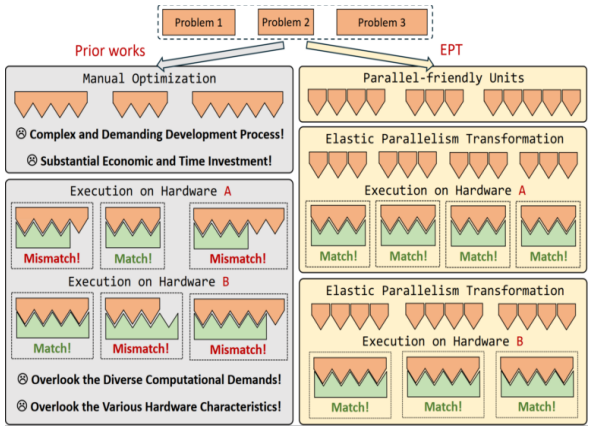


图4: 弹性并行转换（EPT）编译器系统框架图

通过在多种 GPU 硬件（如 NVIDIA V100、A6000、A100 等）上对13种具有代表性的分子进行测试，实验结果显示，EPT 在保证从头算精度的前提下，相较于现有的顶级 CPU 和 GPU 解决方案，性能分别提升了高达34.90倍和9.89倍。通过 Cloud4Science 范式，量子化学研究的计算效率和精度得到显著提升，这为加速新材料开发、药物设计和基础科学探索提供了坚实的技术基础。

Cloud4Science范式推动HPC领域变革，加速科学研究发现

在科学研究迈向智能时代的进程中，矩阵计算正逐渐成为连接传统数值计算与科学智能的关键桥梁，而 Cloud4Science 范式凭借其 Z 级计算潜力，不但为科学在时间和空间尺度上带来了质的飞跃的可能，同时也为科学计算向智能化与推理驱动方向的演进注入了动力。以量子化学为例，Cloud4Science 不仅能缩短计算周期，将复杂分子相互作用的模拟时间从数年压缩至数周甚至数天，还能通过矩阵计算与 AI 推理的融合，使得系统能够基于海量计算数据进行模式识别与智能推理，例如预测药物分子与蛋白靶点的相互作用趋势，自动发现可能的抗性突变路径。

正如个人计算机从单机时代迈入云计算时代，彻底革新了信息处理的广度与效率，未来 Cloud4Science 范式的成功应用也有望在人工智能时代为高性能科学计算带来新的变革。通过融合云计算的可扩展性、AI 的智能决策能力以及高性能计算技术，Cloud4Science 将在未来迈向 Z 级计算的过程中，实现科学计算在极限求解与智能推理两大方向的双向突破，赋予科学智能更强的灵活性、更高的效率与更广泛的可扩展性，为科学研究带来新的创新动力与发展空间。

“Cloud4Science 新范式将显著降低高性能计算基础设施的开发成本，并提升其对科研人员的易用性。尤其是对于那些资源有限的小型研究团队或初创企业而言，这一范式将赋能他们获取 E 级乃至万 P 级科学计算的潜力。这意味着更多的科研工作者可以参与到之前仅限于顶尖机构和大型企业才能涉足的前沿科学计算研究中，极大地拓宽了科学研究的边界，加速科学发现的步伐。”曹婷表示。

相关链接：

1. [PPoPP'24, [Best Paper Award]] ConvStencil: Transform Stencil Computation to Matrix Multiplication on Tensor Cores. https://doi.org/10.1145/3627535.3638476

2. [To be appeared] Matryoshka: Optimization of Dynamic Diverse Quantum Chemistry Systems via Elastic Parallelism Transformation. https://arxiv.org/abs/2412.13203

3. [PPoPP'25] FlashFFTStencil: Bridging Fast Fourier Transforms to Memory-Efficient Stencil Computations on Tensor Core Units. https://www.likun.tech/pdf/ppopp25_FlashFFTStencil.pdf

4. [SC'24] LoRAStencil: Low-Rank Adaptation of Stencil Computation on Tensor Cores. https://doi.org/10.1109/SC41406.2024.00059

5. [PPoPP'25] Jigsaw: Toward Conßict-free Vectorized Stencil Computation by Tessellating Swizzled Registers. https://www.likun.tech/pdf/ppopp25_Jigsaw.pdf

6. [IPDPS'24] VNEC: A Vectorized Non-Empty Column Format for SpMV on CPUs. https://ieeexplore.ieee.org/document/10579118

VisEval：推动自然语言生成可视化的全新评估框架

随着人工智能技术的快速发展，数据可视化日渐高效、智能。但自动化生成的图表是否可靠，成为了亟待解决的问题。微软亚洲研究院提出了 VisEval 评估框架，为这一挑战提供了解决方案，并因此荣获全球可视化领域顶尖的学术会议 IEEE VIS 2024 的最佳论文奖。通过高质量的数据集和可靠的自动化评估方法，VisEval 为数据可视化的未来发展提供了坚实的基础，助力数据可视化技术向更智能、更便捷的方向发展。

在如今这个数据驱动的时代，数据可视化已成为展示数据内在信息的重要工具之一。想象一下，若只需一句简单的指令，复杂的数据便能"化繁为简"，呈现为直观、易于理解的图表，那么既可以减轻分析数据的压力，也让数据的交流与传递变得更为轻松有趣。近期，得益于大语言模型（LLMs）的突破性进展，自动化数据可视化生成的梦想逐步实现。然而，潜在的问题也随之而来：由 LLMs 生成的可视化图表，真的可靠吗？它们是否能遵循数据可视化的最佳实践？

为了有效应对这些挑战，微软亚洲研究院推出了一套全新的评估框架——VisEval，为数据可视化生成提供了更全面、更科学的评估机制。VisEval 不仅构建了一个高质量、覆盖广泛的可视化数据集，还通过多维度的评估机制，从生成代码的有效性到图表的契合性、可读性进行了全面审查。相关论文已被全球可视化领域顶尖的学术会议 IEEE VIS 2024 评选为最佳论文。

数据可视化的智能边界

用大模型生成可视化的过程通常包括：将用户的查询（query）和数据整合到提示词（prompt）中，然后使用诸如 Matplotlib 或 Seaborn 等可视化库生成代码，最终在沙盒环境中执行这些代码以生成图表。虽然这个过程听起来简单，但实际上，现有的大语言模型在生成可视化时会面临诸多问题。

当大模型根据船舶数据生成堆叠条形图时，不同模型的表现各异：有的生成的代码无法执行，有的映射数据出错，还有的未能正确排序或图例摆放混乱。这些问题可以归纳为三大类：有效性、契合性和可读性。具体而言：有效性指图表能否成功生成并准确呈现数据；契合性是指图表是否满足用户的实际需求，例如轴、图例、数据字段等是否符合要求；可读性则考虑图表是否易于理解，例如颜色搭配和布局设计是否合理。

可靠、全面的自动化评估框架的缺乏，阻碍了人们对大模型在生成可视化时不足之处的认识。目前，可视化评估数据集普遍存在一些局限性，限制了全面深入的评估：缺少文本查询或原始数据、缺乏明确的标准答案，过于专注于狭窄领域且规模有限。

此外，现有的评估方法也存在不足。人工评估虽被视为"黄金标准"，但其耗时费力，难以大规模推广；基于规则的评估能检查数据的匹配情况，但常常忽视可视化的可读性问题；而利用大语言模型评估生成代码的方式尚未经过充分验证，其可靠性存疑。

VisEval：高质量数据集与自动化评估框架

VisEval 的提出不仅提供了一个高质量的大规模数据集，还引入了可靠且多维度的自动化评估框架，从而确保对生成的可视化可以进行全面的评估。

在数据集的构建过程中，研究员们专注于挑选无歧义、合理且无重复的查询，并为每个查询提供了准确的标准答案，来保证数据的可靠性和有效性。为了增强数据集的鲁棒性，研究员们在数据集中涵盖了多个领域和图表类型，同时排除了过于简单的查询，以确保生成的可视化具有一定的复杂性。基于 nvBench 数据集，研究员们结合大语言模型和人工专家的筛选，精心挑选出了高质量的查询。这一过程不仅兼顾了数据的质量，还有效减少了人力工作负担。此外，研究员们通过元信息（meta information）为每个查询注释了所有可接受的标准答案，并进行了数据集的重新平衡。这一系列措施都确保了 VisEval 数据集的全面性和实用性，使其在自动化数据可视化的评估中可以发挥更大的作用。最

终，VisEval 的数据集中包含了7种图表类型，超过1,000个可视化图表。

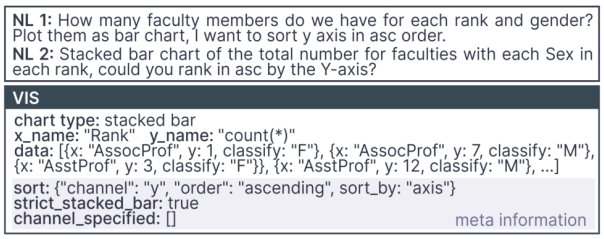


图1：数据集示例

数据集构建完成后，研究员们又开发了一个自动化评估框架。该框架如图2所示，分为三个模块，分别对生成代码的有效性（validity）、契合性（legality）和可读性（readability）进行评估。为了保证评估框架的可靠性，研究员们进行了细致的审查，重点关注数据集和评估框架的各个方面。此外，研究员们通过设计测试用例和进行定量评估，来验证评估的质量。通过这些步骤，研究员们能够确保评估框架可以准确、全面地对生成的可视化进行有效评估。

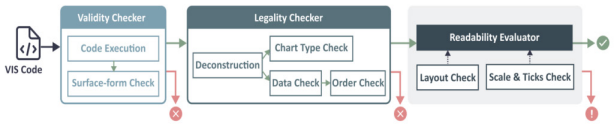


图2：自动化评估框架概览

自动化评估框架具体如下：

- 有效性检查：代码生成后，框架会执行代码，并检查是否成功生成了可视化，以确保代码的有效性。
- 契合性检查：一旦代码通过了有效性检查，契合性检查模块会提取图表类型、数据等信息，并根据数据集中注释的元信息评估图表的契合性。这一步是为了确保生成的图表能够符合用户查询的要求，并且数据映射合理。
- 可读性检查：可读性评估是框架中最具挑战性的部分。评估可读性需要考虑多方面因素，例如布局、比例和颜色等。为此，研究员们借助了 GPT-4V(ISION) 的强大能力，并将可读性评估任务分解为多个更可控的子问题。先分别对布局、比例尺和刻度进行细化评估，再将评估结果汇总，整体给出可读性的打分。研究员们的定量实验表明，VisEval 的可读性评估结果与人类偏好高度一致。

实验结果

基于 VisEval 框架，研究员们对多种模型的可视化生成能力进行了系统评估。

第一项实验测试了不同模型在使用 Matplotlib 和 Seaborn 库时的表现。如表1所示，VisEval 揭示了不同模型在生成可视化方面的显著差异。以 GPT-4 为例，在使用 Matplotlib 库生成可视化时，其质量得分为2.89（满分为5），虽然表现尚可，但依然有改进的空间。相对而言，在使用 Seaborn 库时，尽管其代码通常比 Matplotlib 更为简洁，但所有模型的得分均较低，这一结果令人意外。

Model	Library	Invalid Rate	Illegal Rate	Pass Rate	Readability Score	Quality Score
CodeLlama-7B	Matplotlib	42.95%	28.88%	28.17%	3.87	1.11
Gemini-Pro		14.35%	34.06%	51.59%	3.95	2.06
GPT-3.5		8.79%	29.42%	61.79%	3.52	2.21
GPT-4		3.29%	21.44%	**75.27%**	3.80	**2.89**
CodeLlama-7B	Seaborn	59.26%	24.25%	16.49%	3.64	0.61
Gemini-Pro		21.09%	26.82%	52.09%	3.88	2.06
GPT-3.5		9.21%	31.00%	**59.79%**	3.60	2.20
GPT-4		25.41%	15.89%	58.70%	3.87	**2.31**

表1：自动化评估框架概览

通过深入分析，研究员们发现大语言模型在可视化生成的多个阶段都容易出现错误。这些阶段包括代码编写、数据转换、可视化转换以及排序等方面。除了准确性之外，模型生成的图表在可读性方面也频繁出现问题，而这一点在以往研究中常被忽视。

在第二项实验中，研究员们探讨了不同提示词设计对模型性能的影响。其中特别分析了三种基于大语言模型的可视化生成方法：LIDA、Chat2VIS 和研究团队提出的 CoML4VIS。实验结果表明，提示词设计对模型的表现具有显著影响。由于注意到这三种方法采用了不同的表格格式，所以研究员们还进行了额外的实验。实验中研究员们保持 CoML4VIS 中其他条件不变，仅更改表格格式。如图5所示，不同的大语言模型展现出了对不同表格格式的偏好。这一发现提示研究员们，可能需要针对不同模型设计不同的提示词，以优化其表现。

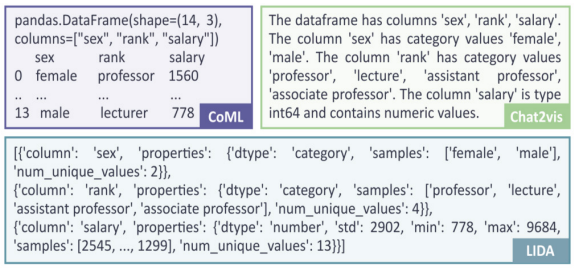


图3：不同的模型有不同的表格格式偏好

在第三项实验中，研究员们测试了无用数据表对生成可视化性能的影响。如表2所示，当给大模型的提示词中包含两张无用的数据表时，所有模型的性能均显著下降。这一结果表明，模型在处理复杂输入时容易受到干扰，强调了在选择数据时确保其相关性的重要作用。

Choice	CodeLlama-7B	Gemini-Pro	GPT-3.5	GPT-4
w/o disruption	28.17	51.59	61.69	75.27
disruption	17.44 -10.73	31.80 -19.79	54.68 -7.01	65.86 -9.41

表2：存在无用表格时对模型性能的影响

作为一种全新的可视化生成评估框架，VisEval 通过高质量的数据集和可靠的自动化评估方法，填补了现有评估体系的空白，为未来的可视化生成研究提供了重要的参考标准。微软亚洲研究院的研究员们期待，随着数据可视化技术变得更加智能和便捷，各行各业的数据驱动决策将得到更有力的支持。

相关链接：

VisEval: A Benchmark for Data Visualization in the Era of Large Language Models
https://arxiv.org/abs/2407.00981

GitHub 链接：https://github.com/microsoft/VisEval

MarS：生成式基座模型时代的通用金融市场模拟引擎

生成式基座模型（Generative Foundation Model）已成功应用在多个领域，并塑造了全新的生产范式。将这一范式与行业特有数据结合，有望构建具有行业特色的生成式基座模型。对此，微软亚洲研究院从金融场景出发，设计了大市场模型 LMM 和金融市场模拟引擎 MarS，希望帮助金融研究人员为不同场景定制生成式模型，并且构建适用于金融市场所有下游任务的生成式基座模型应用新范式，为金融领域带来更高的效率和更精准的市场洞察，在革新金融场景的同时，推动相关领域的跨越式发展。

近年来，生成式基座模型（Generative Foundation Model）在自然语言处理、图像和视频生成等领域取得了巨大成功，推动了新一轮的学术研究和产业应用浪潮，逐步为多个行业塑造出新的生产范式。生成式基座模型的强大能力离不开三个关键要素：海量且高质量的训练数据；能够将数据中的核心信息（如文本中的语义信息）进行有效的令牌化（tokenization）和序列化（sequentialization）；通过自回归（auto-regressive）的训练方式对数据进行建模，从而获得对核心信息的深刻理解和隐含的推理能力。

基于多年来在多个行业成功落地人工智能的实践经验，微软亚洲研究院的研究员们意识到，将这种范式与行业特有的核心数据相结合，可构建出具有行业独特性的生成式基座模型，进而推动相关领域的跨越式发展。

金融市场的交易订单数据就是一个典型例子。研究员们发现，金融市场订单数据具有三大重要特征：

1. 细粒度：订单作为金融场景下最基础的原子数据，能够全面、精细地刻画真实市场，结合相应的撮合规则，能够还原出市场的完整运行过程；

2. 大规模：全球交易所经过多年的电子化交易，积累了海量的交易订单数据；

3. 结构化：订单数据具有良好的结构化特性，便于进行令牌化和序列化。

这些特征使订单流数据有望成为金融市场生成式基座模型的坚实基础。基于此目标，微软亚洲研究院提出了大市场模型（Large Market Model，LMM），并设计推出了基于 LMM 的金融市场模拟引擎 MarS（Financial Market Simulation Engine Powered by Generative Foundation Model），旨在助力金融研究人员为不同场景定制生成式模型，以及构建适用于金融市场所有下游任务的生成式基座模型应用新范式，为金融行业的效率提升和精准洞察带来变革。

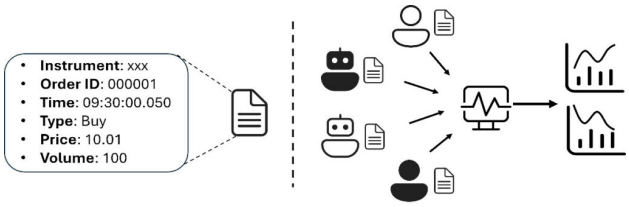


图1：股票交易市场及订单的示意

订单流信息的令牌化

作为金融市场生成式基座模型的核心要素，订单流数据不仅能精细地体现市场参与者围绕投资标的的实时博弈过程，还在不同尺度下展现出了两类独特的价值：

细粒度的市场反馈：从单一市场参与者的视角来看，每笔订单（特别是大额订单）发出后，其他市场参与者在观察到该订单后可能会调整自身决策。这种调整往往会体现在后续订单中，进而形成市场整体对该订单的反馈。这种反馈展现了金融市场价格博弈过程中的微观视角。

宏观的市场博弈过程：从整体市场的视角来看，所有市场参与者之间的复杂博弈汇聚在一起，形成了某一时间段内的市场交易特性。随着时间的推移，这种交易特性的变化记录了市场中多空双方分歧的起始、演进及最终弥合的博弈过程。

研究员们根据订单流信息的特殊价值，分别对单笔订单及其相关的订单簿，以及一段时间内的所有订单集合进行建模，进而构建了 LMM。两种不同层次的建模分别对应上述的细粒度反馈和宏观市场博弈特性，即订单模型（order model）和批量订单模型（order-batch model）。针对原始订单流数据，图2展示了服务于订单模型和批量订单模型的两种令牌化设计。这种令牌化设计使模型能够精准捕捉订单流中的微观和宏观信息，从而对金融市场的复杂动态进行更准确的建模和模拟。

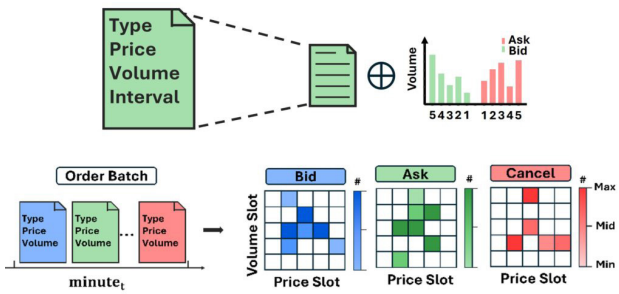


图2：针对单一订单的令牌化（上）及针对批量订单的令牌化（下）

大市场模型扩展定律，释放金融数据潜在价值

随着训练数据的增多以及模型参数的扩展，生成式基座模型的能力会持续提升，带来传统小模型难以企及的想象空间。基于上述两种令牌化方式，研究员们在 LMM 中分别设计了基于 Transformer 架构的生成式模型，并在不同规模的训练数据和参数规模下进行了测试。结果如图3所示，无论是订单模型还是批量订单模型，都表现出了显著的扩展定律（scaling law）。这意味着，在生成式基座模型的支持下，金融场景中海量的历史交易数据有望释放其长期潜藏且尚未充分发掘的巨大价值。

在 LMM 中，研究员们还融合了订单模型及批量订单模型，对不同尺度和不同博弈视角的订单流进行建模，保证了模型所生成的订单流能够体现对市场准确且深刻的理解。这不仅提升了模型的生成能力，还为市场订单数据的时序建模开辟了新路径，使 LMM 在生成市场订单流时具有更高的精准性和现实模拟能力。

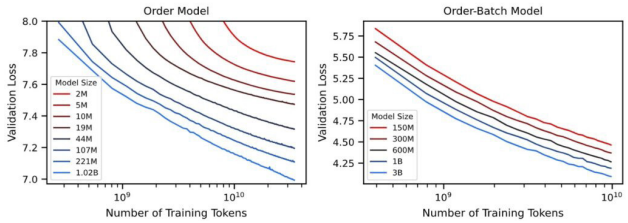


图3：订单模型及批量订单模型在不同参数规模下的扩展曲线（scaling curve）

基于大市场模型的MarS：可定制化金融场景的生成式模型

生成式基座模型一经训练完成，便能通过简单的适配应用于广泛的下游任务，并且在许多场景中超越为单一任务设计的传统小模型。为了充分发挥 LMM 对金融市场的强大建模能力，研究员们分析了各类金融场景中的潜在下游任务需求，设计并推出了基于 LMM 的金融市场模拟引擎 MarS。

MarS 不仅是一种通用的金融市场模拟工具，还为金融行业的多种下游任务提供了全新的生成式基座模型应用范式。借助 MarS，金融研究人员能够为不同的金融场景定制生成式模型解决方案，覆盖领域广泛，从市场预测、风险评估到交易策略优化等等。

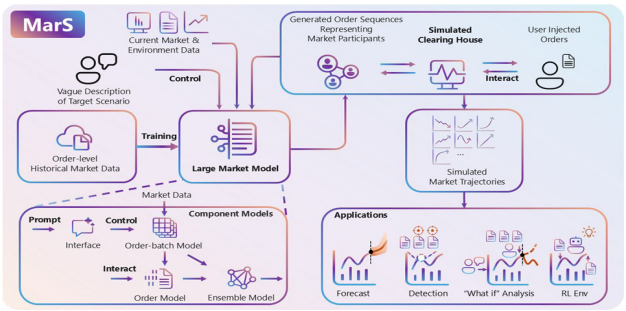


图4：金融市场模拟引擎 MarS 框架图

构建用于预测和检测任务的统一新范式

传统的金融市场解决方案往往需要专家针对不同场景和任务设计专门的算法和策略。然而，金融市场的天然动态性使得这些算法和策略必须不断调整和更新，耗费了相关人员大量的时间和精力。而在生成式基座模型时代，LMM 已经有能力对金融市场进行细致深入的建模，并且可以根据最新的市场数据进行定期

更新。因此，研究员们希望利用 LMM 强大的市场建模能力，构建一个适用于金融市场所有下游任务的"一力降十会"的新范式。

在 MarS 中，研究员们设计了模拟真实订单撮合规则的虚拟交易所（如图4右上角所示）。然后在虚拟交易所中撮合由 LMM 生成的订单流，模拟生成与这些订单流相对应的成交情况，并推演出市场的模拟轨迹（simulated market trajectories）。基于这一机制，金融场景中最常见的预测类和检测类任务便有机会在生成式基座模型的框架下，设计出全新的解决方案。

应用于预测类任务

金融市场中的预测类任务非常广泛，任何依赖于对未来市场指标进行估计的任务都属于这一范畴。当前，无论是基于经济学理论还是数据驱动的主流金融预测模型，都遵循"拟合特定场景和指标"的设计范式。这种范式的局限在于，一旦预测目标发生变化，就需要重新调整和设计模型。例如，图5展示了数据驱动模型 DeepLOB 在预测股票价格走势时的情况，通常这种情况需要分别训练5个模型来获得未来1-5分钟的走势预测。

但在 MarS 的新范式下，只需将最近的真实市场数据输入到 LMM 中持续生成未来的订单流，并在虚拟交易所中进行撮合，就能得到一条可能的未来市场轨迹。通过多次模拟，不仅能够获得未来走势的预测，还能推断出其他任何市场指标。如图5所示，基于 MarS 新范式的预测性能显著优于传统的标杆算法，为金融市场的预测类任务提供了极具吸引力的解决方案，同时也从侧面体现了 LMM 在股票市场建模方面的强大能力。

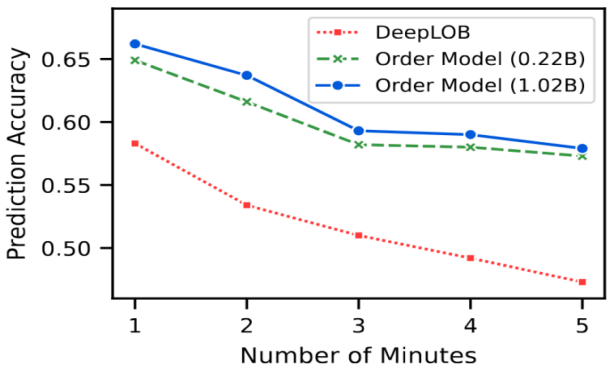


图5：以"预测股票价格未来趋势"任务为例，在金融场景预测类任务中基于 MarS 的新范式具有显著优势

应用于检测类任务

对于金融市场的监管者而言，精准有效地检测潜在的系统性风险或恶意交易行为是维护金融市场健康高效运行的重中之重。检测任务中的关键是在于找到能够区分异常情况与正常市场行为的指征。LMM 所刻画和建模的正是金融市场的通用规律，生成的订单流代表了普遍情况下的市场行为。因此，通过将真实市场轨迹与 MarS 生成的模拟轨迹进行对比，就有机会获得传统方法难以察觉的异常指征。

图6展示了一次真实的恶意市场操纵行为的前、中、后三个时段内，模拟市场轨迹与真实市场轨迹之间的 Spread 分布差异（Spread 指最优买卖价格之间的差值，可用于反映资产的实时流动性，Spread 大意味着流动性较差）。可以明显观察到，在监管机构披露的恶意操纵的时间段内，模拟市场轨迹与真实市场轨迹之间的相关程度显著降低，这是一个有助于监管机构更高效查处恶意市场操纵行为的重要指征。这一类依赖于对微观市场行为有高质量建模的监管指征，在没有高质量的订单流生成的基座模型之前是难以获得的。利用这种方法，许多金融市场的检测类任务将可以通过对比模拟市场与真实市场，找到高效的检测指征。

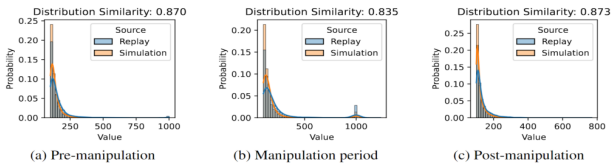


图6：模拟市场轨迹与真实市场轨迹在 Spread 分布下的相关度在已被证实的市场操纵的前、中、后的差异，操纵进行中的相关度显著降低，有潜力成为一项检测市场操纵嫌疑的指标。

重新定义金融科技新场景

生成式模型能够通过简单的描述生成符合特定条件的内容。在模拟引擎 MarS 中，研究员们设计了一种方案，可以根据自然语言描述生成特定市场状况下的订单流。考虑到金融场景的极端市场状态具有特殊的研究意义，研究员们特意设计了一套基于层次扩散模型的订单流调控信号生成系统，从而保证即便是在生成真实世界中罕见的极端市场情况下，例如股灾、熔断等，也能生成若干有区别又有高保真度的调控信号。通过这些细粒度的高保真调控信号，MarS 得以将宏观的市场描述转化为对微观订单流的精细调控，实现精准的订单流生成。

此外，由于市场参与者的所有意图和行为最终都会通过订单这一形式表达和交互，所以对市场的研究本质上是对订单及其交互行为的深入分析。MarS 内置了能撮合任意合法订单流的虚拟交易所，通过研究员们开发的机制可以使外部交互订单能够无缝插入由 LMM 生成的订单流中，同时确保后续订单流仍保持高保真度，真实地反映这些交互订单所带来的影响。通过观察外部交互订单对模拟市场的影响，相关研究人员能够在生成式模型的帮助下，收集到以往只能靠投入巨大财力才能获取的珍贵数据。

通过结合 MarS 的可调控订单流生成能力和对交互订单的真实反馈，研究员们发现 LMM 不仅为主流的预测和检测任务提供了一种新的统一范式，更有希望重新定义金融科技的研究方向、应用技术、市场探索以及理解市场的方式。为此，研究员们尝试将原本只能在实验室环境中构想的两类应用场景带入现实——"假设……会怎样"的分析（"What If" Analysis）以及为强化学习等算法提供接近真实金融市场的数字孪生训练和测试环境。

“如果……会怎样？”类的分析研究任务

“在不同市场环境下，不同规模的交易订单的投入会对市场产生怎样的影响？”这一假设性问题对金融市场非常重要。但传统研究方法依赖于收集真实订单交易的市场反馈以及诸多经验总结和假设，成本高昂，致使相关研究进展缓慢。而生成式模型为这一问题的解决提供了突破性的契机。

图7左上展示了在 MarS 市场模拟中，一批买入订单如何影响资产价格轨迹并演化出不同的市场轨迹。图7右上则展示了 MarS 模拟不同交易策略的市场影响曲线，它们与传统研究总结出的真实市场模式几乎一致，证明了 MarS 在替代传统高成本研究方法上的巨大潜力，也间接证明了 MarS 在模拟和刻画订单间复杂市场行为方面的强大建模能力。

更进一步，研究员们利用 MarS 的低成本优势，通过模拟市场轨迹的大量数据，借助常微分方程（ODE）构建了较为准确的市场影响模型。图7左下展示了通过 ODE 得到的市场影响公式，图7右下则展示了该公式的高可解释性。

研究员们相信，借助 LMM 对金融市场的准确建模，以及 MarS 对可控生成和交互式订单的支持，金融场景中的“假设……会怎样？”类研究问题将迎来快速的发展和显著进步。

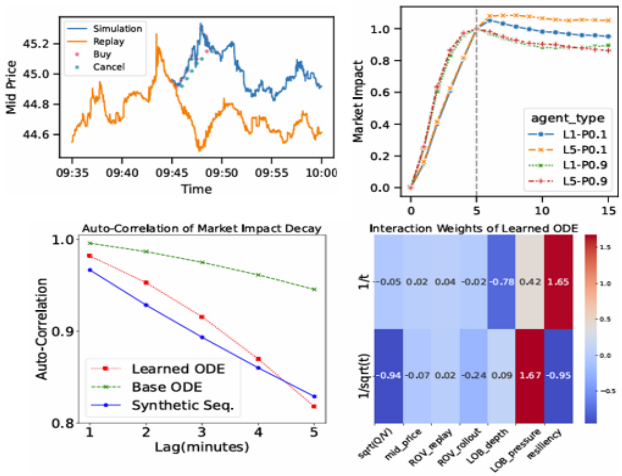


图7：基于 MarS 的“订单的市场影响”问题的研究成果示例

金融市场中强化学习算法的训练环境

过去，强化学习等自主探索和优化算法只能在实验室环境下运行，在实际场景中的应用受限。这类算法依赖于在模拟环境中进行结果评估和优化决策。然而，金融市场的行为和决策往往表现为订单流的变化，进而影响市场。如果训练所依赖的模拟环境不能够准确模拟市场影响，且无法根据算法的行为/决策的改变及时调整反馈，那么在模拟环境中表现良好的算法在实际场景中可能无法达到预期。此外，由于强化学习算法需要自主探索和调优，如果模拟环境只能模拟常规场景而无法刻画现实中的极端情况，那么训练得到的算法在实际应用时可能会在极端场景下出现极其不佳的表现。

MarS 的高保真调控生成能力和对外部交互订单的实时反馈，为强化学习算法在金融市场下游任务中的应用提供了更广阔的空间。图8展示了在 MarS 模拟引擎中从头开始训练交易智能体的过程。研究表明，随着市场反馈的不断更新，强化学习算法在真实模拟环境中逐步学会了更优的交易策略，并获得了令人满意的回测结果。这一成果显示了 MarS 作为强化学习训练环境的潜力，将可以为金融市场算法的自主优化提供有力支持。

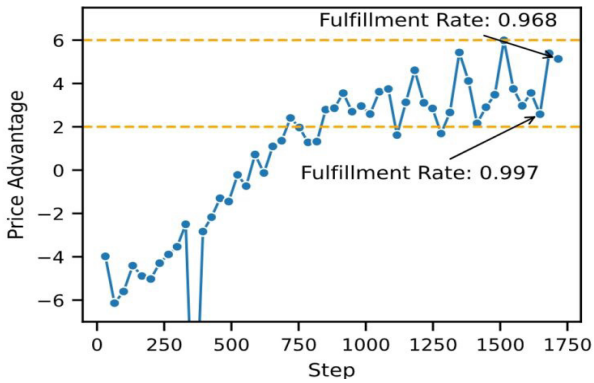


图8：以 MarS 为环境训练的强化学习交易代理的训练表现。在训练过程中，代理的性能显著提升，展示了 MarS 在帮助训练强大且面向真实市场的强化学习算法的能力。

生成式基座模型推动各行业迈向智能化、自动化和精准化的新高度

随着基于生成式基座模型的新范式不断完善，金融市场相关领域的各类下游任务有望通过适配这一新范式实现性能提升和突破。更重要的是，这一新范式具有普适性。未来，其他拥有海量且复杂核心数据的垂直领域，包括医疗健康、能源、物流和制造业等，也具备开发行业特有生成式基座模型的潜力。例如，能源领域可以利用历史能源消耗和价格波动等数据，建立智能化的能源分配和预测系统。

生成式基座模型的广泛应用将不仅仅推动金融市场相关技术的飞跃，也为其他数据密集型领域提供了全新的研究方向和解决方案。伴随新范式的发展，未来会有更多行业和领域在生成式人工智能的支持下，迈向智能化、自动化和精准化的新高度。注：本文中提到的微软亚洲研究院在金融领域的研究属于科研探索性质，旨在推动科学进步，并为金融领域的研究和应用提供理论和技术支持。所有研究均严格遵守微软负责任的人工智能流程的指导，并遵循公平、包容、可靠性与安全性、透明、隐私与保障、负责的原则。文中提及的技术和方法目前仍处于研究和开发阶段，尚未形成商业产品或服务，亦不构成任何金融解决方案。我们建议读者在做出金融决策时，咨询具有合法资质的金融专业机构和从业者。

相关链接：

MG-TSD: Multi-Granularity Time Series Diffusion Models with Guided Learning Process
https://arxiv.org/abs/2403.05751

Controllable Financial Market Generation with Diffusion Guided Meta Agent
https://arxiv.org/abs/2408.12991

MarS: a Financial Market Simulation Engine Powered by Generative Foundation Model
https://arxiv.org/abs/2409.07486

GitHub链接：
https://mars-lmm.github.io
https://github.com/microsoft/MarS

如何泛化AI的深度推理能力？

大语言模型的推理能力一直是人工智能领域的研究热点，但传统依赖大规模数据和参数扩展的预训练方式在提升模型推理能力上逐渐遇到了瓶颈。微软亚洲研究院的最新研究关键计划步骤学习 CPL（Critical Plan Step Learning），旨在将强化学习扩展到更广泛、更复杂的问题场景，并取得了突破性进展。CPL 通过在自我生成的高层次抽象计划上进行强化学习，不仅提升了模型在数学推理任务上的表现，还在多个跨领域推理任务上展现出了卓越的泛化能力。

通过大规模预训练，大语言模型（LLMs）在自然语言处理、数学推理等任务中取得了显著进展。然而，传统依赖大规模数据和参数进行扩展的预训练方式在提升模型推理能力上面临着新的挑战。科研人员相信 LLMs 的能力不止于此，在后训练阶段学习人类的慢思考，特别是通过在自我生成的数据上进行大规模强化学习（Reinforcement Learning, RL）训练，对问题进行全面的分析和推理是实现 LLMs 智能的关键。这种新的强化学习范式在复杂任务上的推理表现取得了前所未有的突破，也为未来的超级智能奠定了技术基础。

但想将强化学习扩展到更广泛、更复杂的问题场景仍面临诸多挑战。传统的强化学习方法，由于其特定的动作空间限制了模型的泛化能力，因此，如何将强化学习对推理能力的提升泛化到更广泛的推理任务丞待解决。此外，与传统强化学习不同，LLMs 的复杂问题推理涵盖了巨大的搜索空间。考虑到 LLMs 推理的高成本，如何在如此庞大的搜索空间中高效且有效地找到正确的解决方案也是一个关键挑战。

为了解决这些问题，微软亚洲研究院的研究员们推出了关键计划步骤学习 CPL （Critical Plan Step Learning），作为扩展强化学习以开发通用推理模型的重要一步。具体来说，该方法提出了一种在高层次抽象计划的动作空间中进行搜索的新方法，以提升模型的泛化能力。通过使用蒙特卡洛树搜索（Monte Carlo Tree Search, MCTS）探索多样化的计划步骤，并引入步骤级优势偏好优化算法（Step-level Advantage Preference Optimization, Step-APO）来学习解题中的关键步骤，CPL 帮助模型聚焦于推理过程中的重要决策，显著提升了模型的推理能力和泛化性。

CPL 不仅在数学推理任务中表现突出，还在 HumanEval、GPQA、ARC-C 等跨领域推理基准任务上取得了优异成绩，为大语言模型提升推理能力的泛化性开辟了新的方向。这一工作成功提高了模型在多种跨领域任务中的迁移表现，并为未来强化学习扩展方向的研究做出了一定的贡献。

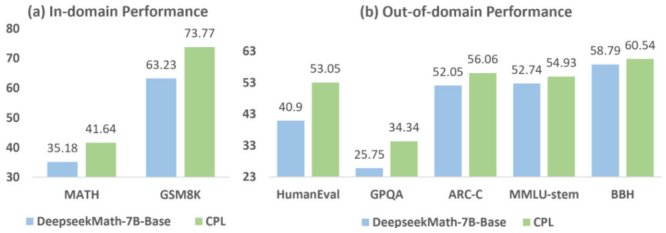


图1: CPL 结果概述：(a) 领域内表现；(b) 领域外表现。

在抽象计划中寻找推理的钥匙

CPL 的核心由两个关键步骤组成——计划搜索和关键计划步骤学习。这些方法旨在有效提升模型在多步推理任务中的推理能力和泛化能力。

研究员们提出了基于计划的 MCTS（Plan-based MCTS），主要在高层次抽象计划的搜索中帮助 LLMs 探索多样化的解题策略，以应对广阔的搜索空间。不同于传统方法只专注于具体执行步骤的探索，CPL 还强调解决问题的整体思路和策略探索，从而帮助其学习更通用的任务无关技能，显著提升模型在各种推理任务中的泛化能力。具体来说，CPL 通过逐步创建解决问题的计划，并最终给出完整的解答；MCTS 迭代构建出计划树，并生成高质量的计划步骤监督信号。此外在 MCTS 中，研究员们会使用价值模型来评估每个部分推理路径的预期回报，这使得模型在面对复杂推理任务时能够有效选择最优路径，最大限度地减少无效搜索的影响。

步骤级优势偏好优化（Step-APO）是 CPL 的第二个关键组成部分，旨在学习和强化推理过程中的关键步骤。Step-APO 建立在直接偏好优化（Direct Preference Optimization, DPO）的基础之上，进一步结合了 MCTS 中获得的步骤级优势估计，用于更精细地比较不同步骤之间的偏好。传统的偏好学习方法通常通过标记第一个错误步骤为不优来指导模型，但这种启发式方法限制了模型对自生成数据的充分利用。相比之下，Step-APO 可以为每个推理步骤计算其相对于同层其他步骤的"优势值"，从而使得模型能够识别出对最终推理结果至关重要的步骤，并赋予这些关键步骤更高的优化权重。通过这种方式，模型能够更好地识别并强化关键步骤，实现更高效的计划优化与泛化。

结合计划搜索与关键步骤学习让 CPL 有效实现了模型推理能力的全面提升。实验结果表明，即便只在 GSM8K 和 MATH 数据集上进行训练，CPL 依然在多个跨领域推理任务中取得了明显的性能提升，例如 HumanEval、GPQA 和 ARC-C 等跨领域基准任务，充分验证了该方法在大模型推理与泛化方面的显著优势。

CPL的实验性能

研究员们在多个推理基准上对 CPL 方法进行了详尽的实验评估，以验证其在推理能力和泛化性上的表现。整个实验评估从以下三个方面展开：域内推理任务的表现、跨领域推理任务的泛化性，以及不同优化策略的效果。

Model	In domain		Out-of-Domain				
	MATH	GSM8K	HumanEval	ARC-C	GQPA	BBH	MMLU-stem
DeepseekMath-Base	35.18	63.23	40.90	52.05	25.75	58.79	52.74
Self-Explore-MATH	37.86	**78.39***	41.46	54.01	33.83	60.04	54.04
AlphaMath	-	-	49.39	53.41	33.33	56.63	55.31
CPL(Round1 SFT)	36.30	63.79	42.68	54.44	28.78	59.68	54.58
CPL(Round1 Step-APO)	40.56	71.06	46.34	55.55	31.31	60.18	55.15
CPL(Round2 SFT)	39.16	69.75	48.78	54.95	29.79	59.93	**55.44**
CPL-final	**41.64**	73.77	**53.05**	**56.06**	**34.34**	**60.54**	54.93

表1: 领域内和领域外推理任务的主要结果

首先，研究员们在数学推理任务 MATH 和 GSM8K 数据集上评估了 CPL 在域内推理任务中的表现。MATH 数据集包含5000个复杂的竞赛级问题，旨在评估模型对高难度数学推理的能力，而 GSM8K 数据集则包含1320道小学数学题目，用以测试基础算术和推理技能。与基准模型 DeepSeekMath-Base 相比，CPL 在这两个数据集上均展示了显著的性能提升。在 MATH 数据集上，CPL 最终取得了41.64%的准确率，相较于基准提升了6.5%；在 GSM8K 上，CPL 的准确率达到了73.77%，比基准提升了10.5%。这表明 CPL 通过高层次计划的学习和关键步骤优化，有效增强了模型在数学推理方面的能力。

其次，研究员们重点评估了 CPL 在五个跨领域推理任务上的泛化能力，分别是 HumanEval、ARC-C、GPQA、BBH 和 MMLU-STEM。这些任务涵盖代码生成、科学知识问答、生物与化学领域的复杂推理等，旨在验证 CPL 在应对不同领域推理任务时的适应性。结果表明，相较于学习特定任务解决方案的模型在跨领域推理任务上表现不佳（如 AlphaMath 在 BBH 任务上出现了2.2%的性能下降），CPL 极大提升了跨领域推理任务的性能。例如，在代码生成的 HumanEval 任务中，CPL 相比基准提升了12.2%；在科学问答的 GPQA 上，CPL 的表现提升了8.6%；在 ARC-C 和 BBH 等任务中，CPL 也展现了优于基准模型的表现。对比基于执行步骤的搜索方法如 AlphaMath 和 Self-Explore，CPL 通过高层次计划的探索与学习，显著增强了模型的泛化能力，使其在面对不同类型的推理任务时表现得更加稳定和优越。

最后，研究员们对比了不同优化策略的效果，包括传统的实例级 DPO（Instance-DPO）、步骤级 DPO（Step-DPO）和 CPL 所使用的步骤级优势偏好优化（Step-APO）。实验结果表明，Step-DPO 相较于 Instance-DPO 在部分任务上有小幅度的性能提升，而 CPL 的 Step-APO 通过更精细的关键步骤学习大大提升了模型的推理性能。在跨领域任务中，Step-APO 的优化效果尤为明显。这说明通过有效识别和学习推理中的关键步骤，CPL 能够显著增强模型的泛化性和推理能力。

	MATH	GSM8K	HumanEval	ARC-C	GQPA	BBH	MMLU-stem
SFT	36.30	63.79	42.68	54.44	28.78	59.68	54.58
Instance-DPO	37.72	69.29	43.90	54.61	24.24	60.13	54.42
Step-DPO	37.89	69.83	42.68	54.44	25.25	59.44	54.68
Step-APO	**40.56**	**71.06**	**48.78**	**55.55**	**31.31**	**60.18**	**55.15**

表2: Step-APO 的优势

构建更智能的通用推理模型

扩展强化学习（Scaling RL）以开发一个通用的推理模型依然是一个开放且重要的研究课题。研究员们提出的 CPL 利用高层次抽象计划的搜索以及关键步骤的优势偏好优化，增强了 LLMs 在推理任务上的泛化能力。

未来，研究员们将增加计划策略多样性，并且结合测试时搜索（test-time search）进一步提升模型的整体推理效果。研究员们也将继续探索如何有效学习推理中的关键步骤，以实现更高效的搜索，为构建更加智能化、具备通用推理能力的语言模型奠定坚实基础。

相关链接：

CPL: Critical Plan Step Learning Boosts LLM Generalization in Reasoning Tasks
https://arxiv.org/pdf/2409.08642

近实时的全球碳预算，揭示2023年陆地碳汇能力锐减

2023年12月，“全球碳项目”（Global Carbon Project，GCP）发布了《2023年全球碳预算》（Global Carbon Budget 2023）报告。然而，该报告仅覆盖至2022年底的碳预算监测，时间滞后长达一年。作为制定与实施“双碳”策略的关键参考，这种长时间的延迟使得碳预算结果难以提供良好的参考依据。针对延迟问题，微软亚洲研究院联合清华大学和法国原子能署气候与环境科学实验室（Laboratoire des Sciences du Climat et de l'Environnement），基于人工智能技术开发了一种创新的方法。通过采用自上而下和自下而上的估算方式，该方法可以对地表与二氧化碳的交换进行评估，成功将全球碳预算时间从以往的滞后一年缩短至三个月。这一成果将为环境保护与可持续发展相关的科学研究和政策制定提供更加及时的数据支持，助力推动全球环境治理与生态文明建设。

工业革命以来，化石燃料的燃烧和土地利用方式的转变，尤其是对森林的砍伐，成为了大气二氧化碳升高的主要推手。尽管陆地植被和海洋作为自然界的主要碳汇，吸收了部分二氧化碳，但排放量的激增已远远超出了它们每年的吸收极限，导致大气中的二氧化碳浓度不断攀升，引发全球变暖和极端天气。在这一严峻背景下，碳预算估算成为了实现碳中和的关键。

碳预算是指对全球碳循环中碳源和碳汇的评估，它综合了化石燃料和水泥排放、土地利用和土地利用变化相关的排放和清除、海洋和自然陆地的二氧化碳来源与吸收等证据，从而衡量大气中二氧化碳浓度的变化。准确且及时的碳预算对理解和应对全球气候变化具有决定性意义。在全球气候和环境挑战日益加剧的今天，监测碳汇与碳排放变得至关重要。特别是在全球各国积极推进碳达峰和碳中和策略的当下，碳预算已成为相关科学研究和制定可持续发展政策的基础。

为了支持全球的可持续发展，微软亚洲研究院制定了“基于人工智能的近实时全球碳预算（ANGCB）”计划，旨在更有效地利用海洋、陆地等自然环境条件来捕捉二氧化碳，以实现全球实时碳预测的目标。

通过与清华大学和法国原子能署气候与环境科学实验室（Laboratoire des Sciences du Climat et de l'Environnement）紧密合作，微软亚洲研究院利用人工智能技术开发了一种创新的综合方法，成功将全球碳预算的时间延迟从一至两年缩短至三个月。这一成果不仅为相关领域的科学研究提供了更及时的数据支持，也为碳减排和碳汇政策的制定者提供了更迅速的数据反馈。

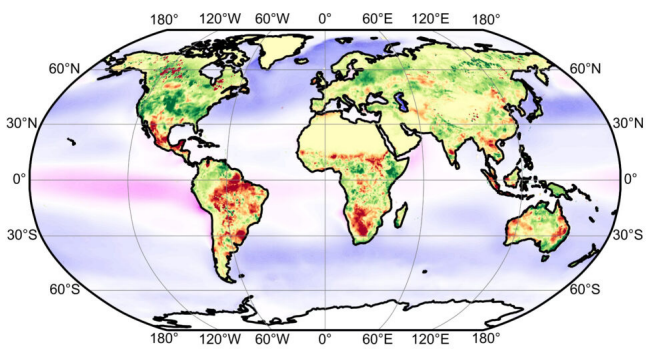


图1：基于人工智能的近实时全球碳预算（ANGCB）

传统碳预算方法存在明显的滞后性

传统的碳预算包含了采用数值模拟的方法，这种方法虽然可以模拟复杂的地球系统过程，但由于涉及的计算量巨大且数据更新速度慢，通常存在一至两年的滞后。以"全球碳项目"（Global Carbon Project，GCP）发布的最新的《2023年全球碳预算》（Global Carbon Budget 2023）报告为例，该报告于2023年12月发表，但其结果仅能覆盖到2022年底，未能包含2023年的信息，碳预算时间滞后长达一年。

而2023年，全球遭遇了多起重大环境事件，对气候产生了显著影响。例如，北美地区森林火灾频发；自2020年以来一直较为活跃的拉尼娜现象在2023年6月转变为中等强度的厄尔尼诺现象；根据 GRACE 卫星的观测，2023年北半球大部分地区的陆地水储量出现下降，这可能导致植物面临水分胁迫，即因土壤缺水而抑制植物生长；亚马逊热带雨林地区从2023年6月到11月遭受了极端干旱，而热带非洲地区则比正常年份更为湿润。

这些关键的气候变化均未被《2023年全球碳预算》所覆盖，这种滞后性不仅影响了对气候变化趋势的准确性判断，也延缓了应对气候变化采取行动的时机。

近实时的人工智能全球碳预算方法

造成滞后的原因主要在于，传统的数值模拟方法是由滞后一年的气候数据驱动的，不能完全适应当前的碳预算需求。若使用近实时更新的气候再分析数据来驱动模型，那么数值模拟方法无法直接兼容新数据，所以结果的准确性会有较大差异。此外，"全球碳项目"采用了十余个陆地和海洋模型，每个模型都来自不同的实验室或机构。这意味着需要协调多个组织同时推进同一项目，进一步加剧了碳预算的延迟。

对于自下而上的海洋碳汇预算，研究员们结合海洋生物地球化学知识和数据驱动模型，设计了全新的机器学习仿真器。"此前，海洋碳预算主要有两种方法。一种是数值模拟，但存在滞后性；另一种是依靠海洋中航行的船只，通过将船只底部传感器收集的数据与卫星监测数据结合，构建机器学习模型，但航测数据是由散点观测扩展至整个全球观测，这种以点代面的方法无法保证结果的准确性。" 微软亚洲研究院应用科学家桂晓凡介绍道。

基于这两种方法，微软亚洲研究院构建了新的机器学习仿真器，使其能够实现近实时的更新。截至目前，仅使用这一个海洋数据驱动的模型，就已经可以完成覆盖全球每个区域的近实时海洋碳预算。

而对于自上而下的陆地碳预算，微软亚洲研究院也在积极尝试利用人工智能技术探索全球陆地碳预算的有效路径。

及时的碳预算为可持续发展策略提供理论依据

微软亚洲研究院近实时碳预算模型的分析结果显示，2023年陆地碳汇能力降至2003年以来的最低点。通常在夏季达到碳汇峰值的北部陆地在2023年的表现低于预期，中欧、西俄罗斯、中美洲也出现了异常的碳排放源。而2023年的海洋碳汇能力则比2022年有所增加，特别是在太平洋和部分南大西洋区域。这一变化主要归因于拉尼娜现象的消退和厄尔尼诺现象的发展，减少了热带太平洋的二氧化碳排放来源，同时高海表温度又降低了东北大西洋的碳汇能力。

总体来看，在2023年，化石燃料的排放估计为102亿吨碳，而碳汇却难以跟上碳排放增长的步伐。受拉尼娜向厄尔尼诺转变导致的极端天气事件频发影响，虽然海洋吸收了26亿吨碳，但陆地的吸收量却急剧下降，仅为1.4亿吨碳，导致大气中二氧化碳浓度水平持续上升。若没有及早进行碳预算，那么人们将无法及时发现这些变化。设想一下，如果未来十年全球持续变暖，并像2023年那样对陆地碳汇产生负面影响，自然碳汇将可能失去一半目前吸收人类活动产生的二氧化碳的能力。

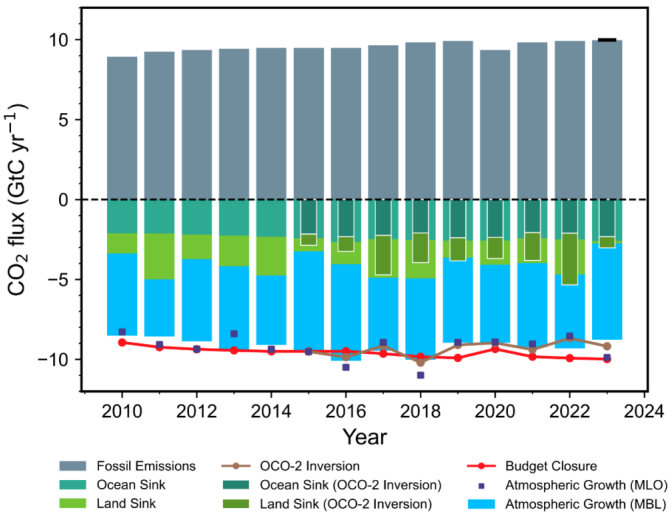


图2：2010年至2023年的碳预算

这表明，减少化石燃料排放以及保护自然碳汇是应对全球气候变暖的重要举措。气温每一度的上升都非常重要，因为温度的上升和碳汇的减弱将决定地球与人类的未来。保护气候、恢复碳平衡刻不容缓，全球社会都应该立即采取行动。

"碳预算时间的缩短意味着政府和决策者可以更快地掌握碳排放和碳汇的最新情况，从而更有效地制定和调整相关政策，确保人类社会的可持续发展。利用与微软亚洲研究院合作构建的近实时碳预算方法，相关机构能够迅速评估政策的实施效果，而不必等待一两年才能看到结果。这种及时反馈对于优化政策执行和提高政策响应的速度至关重要。"法国原子能署气候与环境科学实验室研究员 Philippe Ciais 表示。

跨领域合作是AI在专业领域发挥作用的关键

在当今科学研究的前沿领域，跨学科合作已成为推动创新和解决复杂问题的重要途径。基于人工智能的近实时碳预算方法，就是微软亚洲研究院、清华大学和法国原子能署气候与环境科学实验室共同合作的成果，其中结合了环境科学、地球系统科学、生态学、大气科学和人工智能等领域的专业知识与技术。

“在跨学科合作中，我们深刻体会到不同学科间知识互补的重要性。法国原子能署气候与环境科学实验室提供的宝贵数据，为我们的研究奠定了坚实基础，而人工智能的创新应用让实时监测碳汇与碳排放成为可能。环境科学、地球系统科学、生态学和大气科学的复杂性，需要人工智能领域的科研人员深入理解和融会贯通，而环境科学领域的专家也需要掌握人工智能技术的前沿发展及应用潜力。这种知识的融合与共同学习，为我们实现了人工智能技术在环境监测中的有效应用，也充分展现了跨学科合作在解决全球性问题中的核心价值。”微软亚洲研究院资深首席研究员边江说。

随着技术和数据的不断完善，微软亚洲研究院将整合海洋与陆地碳预算模型，进一步推进 ANGCB 计划。为此，研究员们也将持续提升模型的性能和效率，希望通过碳预算的提前预测，为全球气候变化的研究和相关政策的制定提供更及时准确的数据支持，助力推动全球环境治理与生态文明建设。

相关链接：

Low latency carbon budget analysis reveals a large decline of the land carbon sink in 2023
https://doi.org/10.1093/nsr/nwae367

人工智能“天文学家”能否帮助人类理解宇宙？

在广袤无垠的宇宙中，存在着无数类型各异的天体。借助现代技术，人们能够获取这些天体的丰富信息，包括形状、光谱、坐标、红移、引力透镜、爆发时变等大量数据，进而探究宇宙起源与演变的奥秘。但传统科学技术已难以应对海量数据的处理需求，这限制了天文学研究的进一步发展。

为了帮助天文学家分析遥远星系历经百亿年旅程到达太空望远镜的测光数据，微软亚洲研究院联合清华大学天文系以及俄亥俄州立大学（The Ohio State University）开发了大语言模型智能体 Mephisto。Mephisto 以自然语言形式存储知识库，进而学习、分析相关天文学问题，为天文学家提供了新的研究思路，也为初学者提供了有益参考。

天文学起源于人类仰望星空时对未知的好奇心。作为历史最悠久的学科之一，天文学曾多次引领人类文明的科学革命。在 AI for Science（科学智能）发展如火如荼的当下，基于大语言模型的科学智能体是否也能协助天文学家探索宇宙，发现未知？

天文学是一个相对“小众”的学科，除去诺贝尔物理学奖带来的高光时刻，天文学家大部分时间都隐匿于大众的视野之外。如果用一句话来概括天文学家的核心任务，那就是为宇宙中的各种观测现象——小到每一个氢原子，大到整个可观测宇宙——寻找一个合理的“解释”。

天文学与其他学科有两大根本区别：

- 宇宙中天体的物理条件过于极端，导致天文学的研究对象无法在实验室中开展控制变量实验，因此其理论框架中的大部分内容都存在争议，即便是一个小问题往往也存在几种甚至数十种在一段时间内无法验证的理论；
- 天文学研究的成果通常不具有直接的现实意义。与蛋白质合成和材料发现不同，天文学研究成果通常是作为一种可解释的白盒模型来推动其他学科的发展，例如太阳的谱线观测直接推动了量子力学的诞生，广义相对论的首次验证源自对水星进动的观测等。

这两个特点使得天文学研究对可解释性的要求极为严格，这也限制了黑盒人工智能模型在天文学核心问题中的广泛应用。然而，大语言模型通过在大量文本中进行预训练，不仅掌握了丰富的天文学基础知识，还拥有强大的逻辑推理能力，使其能够构建因果模型来解释观测现象。

在此背景下，微软亚洲研究院与清华大学天文系以及俄亥俄州立大学（The Ohio State University）的研究员们联合开发了大语言模型智能体 Mephisto，并首次将其用于对由詹姆斯·韦布空间望远镜（James Webb Space Telescope, JWST）观测到的高红移星系的深入分析，为宇宙诞生之初的"小红点（Little Red Dots）"提供了可能的解释，开创了将大语言模型作为逻辑推断引擎进行科学发现的新范式。

小贴士：高红移星系是指那些距离地球非常遥远的星系，它们发出的光在到达地球的过程中，由于宇宙的膨胀，波长被拉长，导致光谱向红色端移动，这种现象称为红移。红移的数值（z）表示星系远离地球的速度与光速的比值，红移值越高，星系离我们越远，也越古老。

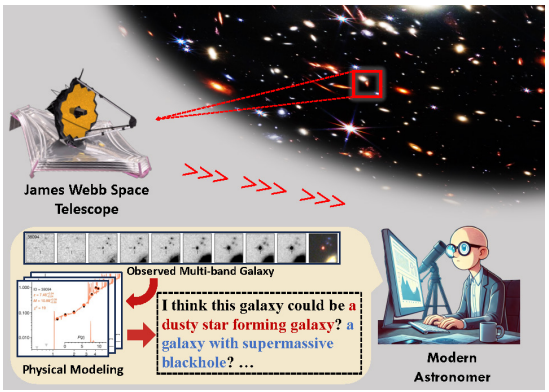


图1：人类天文学家的工作模式：由空间望远镜对成千上万的星系进行观测，天文学家从中发现"有趣"的源并试图使用一系列物理模型对其进行解释。（图像来源：NASA, I. LABBE）

大语言模型智能体Mephisto助力星系数据分析

在传统天文学研究中，分析单个星系的物理性质是每一个天文学研究生新生的必修课。研究人员需要对星系形成理论有深入的了解，并对大量观测数据进行分析，才能建立足够扎实的专业技能。即便对于那些已经拥有丰富专业知识和经验的研究人员而言，仔细探究一个星系的性质、排除各种假设，也是一个耗时的"体力活"。

而大语言模型智能体 Mephisto 则可以帮助天文学家分析那些经过百亿年旅程才到达空间望远镜的遥远星系的测光数据。Mephisto 能够基于给定的测光数据提出相应的星系物理模型，并与一个名为 Code Investigating GALaxy Emission (CIGALE) 的星系光谱模拟程序交互，评估当前物理模型与实际观测数据的差异，分析其中可能的仪器系统误差或者物理模型的不适用性，同时通过不断调整星系物理模型的假设与参数先验，为观测数据寻找若干种可能的解释。

因拥有以自然语言形式存储的知识库和存储模块（memory），这使得 Mephisto 能够从之前的尝试中进行学习，避免重复失败的路径。其知识库包含了与专业相关的技能知识，并且可以在与人类天文学家的交互以及对实际观测数据的互动中，利用强化学习提升自身能力。

Mephisto 提取的知识库具有现实的物理意义，反映了各种物理模型在不同情况下的优势和局限性，为人类天文学家提供了新的研究思路，也为初学者提供了有益的参考。通过模仿科学家的思考方式，Mephisto 将提出假设与不断优化的过程形式化为一个树搜索框架，保证了最终的科学结论来自于整个推理过程的深入分析。

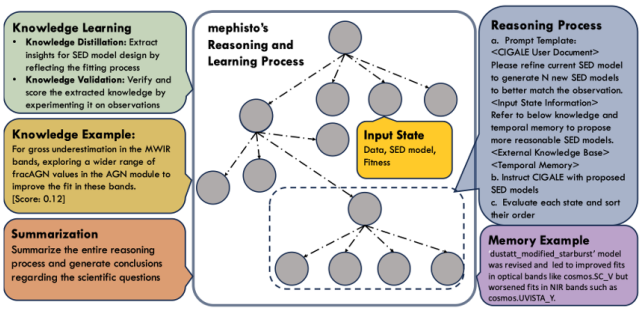


图2：Mephisto的知识库与存储模块

分析JWST LRD最新数据，Mephisto提出更好的物理模型和假设

研究员们在多样化的数据以及前沿科学问题上的测试证明了，Mephisto 可以在持续的搜索中不断提出更加符合观测数据的物理模型，并在这一过程中利用改善机制（self-reßection），学习更多的星系物理知识，进而提出更精准的假设。

通过提供星系在不同波段发出的光流量大小数据，Mephisto 从一个基础模型出发，持续探索和改善，发现了与当前星系观测数据更吻合的解释。在这一探索过程中，Mephisto 不仅逐渐完善了当前观测的可能的假设空间，还验证了科学结论对模型选择的鲁棒性。天文学家可以根据这些报告制定更新的观测计划，修正理论模型，逐步拓展科学的边界。

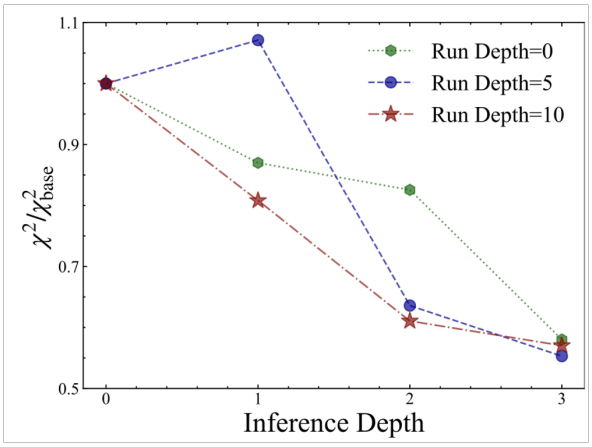


图3：Mephisto 所提出的物理假设随着推理深度（inference depth）以及学习深度（run depth）的演化过程。图中 Y 轴为模型和数据的拟合优度，可以发现随着推理深度与学习深度的提高，Mephisto 可以逐渐提出更好的假设。

在处理前沿科学问题方面，例如 JWST 观测到的小红点——一类可能彻底颠覆天文学家对宇宙认知的天体时，Mephisto 也表现出了专业研究人员甚至更高的水平。"Mephisto 能够全面探索所有关于'小红点'的潜在假设，帮助天文学家更深入地理解这些超出现有理论框架的天体的物理内涵，进而可能带来全新的科学发现。"高红移星系研究专家、清华大学天文系副系主任蔡峥教授评价道。

如图4所示，在星系恒星质量、尘埃消光与是否存在超大质量黑洞的三维坐标系下，Mephisto 充分遍历了所有可能的物理假设，得出的结论与人类天文学家（图中红点所示）相似，甚至更加完备。这类星系在早期宇宙中大量存在，极大地挑战了目前的宇宙学理论。Mephisto 作为人工智能助手，能够持续挖掘此类数据，帮助人类天文学家拓展宇宙认知的边界。

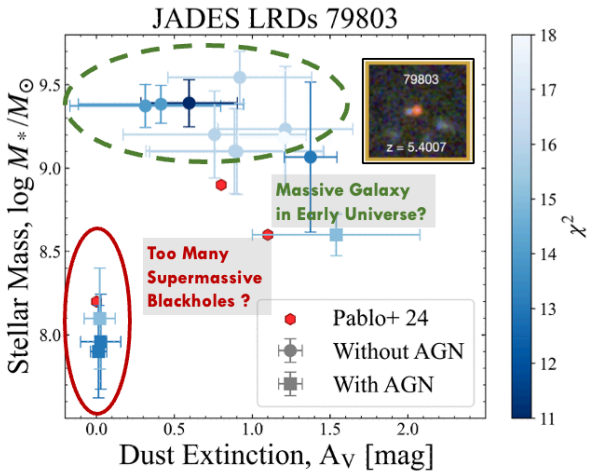


图4：Mephisto 在 JWST LRD 的最新观测数据 JADES ID 79803（一个宇宙形成12.7亿年时的早期星系）上的表现与人类天文学家相似，甚至更加完备。目前天文学界的两种主流解释为：一个充满尘埃的恒星形成星系，或是一个拥有超大质量黑洞的缺乏尘埃的星系。

AI引领天文学研究新范式，与天文学家携手开启科研新图景

"传统上，天文学家只能通过某些启发式的标准快速筛选观测数据，只有最有潜力的天体才会得到专家的深入分析，而大多数星系都未曾被详细研究。大语言模型智能体 Mephisto 的出现改变了这一局面，它让我们在数据爆炸的今天也能够对观测到的数十亿个星系进行深入分析，帮助研究那些行为与现有物理学理论不符的异常天体。这一技术正加速推动我们突破科学知识的边界。"来自俄亥俄州立大学天文系的丁源森教授评论道。

相比传统的人工智能应用，Mephisto 革新了天文学家与人工智能的交互方式。天文学家们现在可以通过自然语言直接与人工智能进行交流，将他们的领域知识和需求直接反馈给人工智能，无需进行反复且成本高昂的训练过程，人工智能也能够将发现以自然语言的形式反馈给天文学家。这种以自然语言表达的知识可以在不同的星系光谱模拟程序和大语言模型之间迁移，无需重复训练。

Mephisto 的推理过程严格遵循目前的星系形成理论，实现了白盒的求解过程，这与天文学追求的可解释性完美契合，意味着 Mephisto 可以无缝融入到当前的科学研究范式中。更重要的是，Mephisto 具备自主学习和持续进化的能力，能够在分析大量数据的过程中不断学习，同时避免了人类科学共同体可能存在的偏见，从而提出尚未被充分考虑的假设，进一步拓宽人类科学家的认知边界。

作为天文学家的人工智能助手，Mephisto 能够在超级计算机上夜以继日、不知疲倦地挖掘那些尚未被充分研究的星系测光数据，并将有趣的发现反馈给人类专家。同时，天文学爱好者也可以借助 Mephisto 更多地参与到天文学研究中。未来，将大语言模型作为逻辑推理引擎，实现科学分析自动化，这一新范式将深入天文学研究的各个领域，持续激发天文学领域的创新活力。

相关链接：

Interpreting Multi-band Galaxy Observations with Large Language Model-Based Agents
https://arxiv.org/pdf/2409.14807

Rho-1：基于选择token建模的预训练方法

在自然语言处理领域，预训练语言模型常因大规模噪声数据而面临挑战。对此，微软亚洲研究院的研究员们提出了一种新型的基于选择 token 建模的预训练方法。该方法通过选择性语言建模（Selective Language Modeling, SLM）策略，精准筛选出对模型训练有价值的 token，有效提升了数据效率和模型性能。这一突破不仅优化了模型训练过程，也为自然语言处理技术的进一步发展提供了新思路。本篇论文在 NeurIPS 2024 上荣获最佳论文 Runner-Up 奖。

传统预训练方法中的token级挑战

现有大模型基于大批量文本语料进行预训练，在各类文本生成、文本理解和文本逻辑推理等任务上表现突出。然而，预训练过程中从各种来源获取的原始语料存在大量噪声，因此科研人员经常采用一些质量过滤方法对原始语料进行过滤，使其可以用于模型预训练。例如，文档级（document-level）过滤可以去除一些干扰文档，进一步地还可以在行级（line-level）过滤单个文档中的噪声，从而得到高质量的语料，用于预训练。

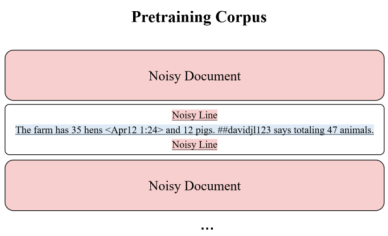


图1：语料清洗示意图

在以往的方法中，过滤出来的高质量语料输入到以因果语言建模（Causal Language Modeling）方式的模型当中，计算每个 token 的损失并平均后求梯度，然后更新模型的参数。然而，当使用这种 next-token prediction 的形式对完整的句子序列进行建模时，可能忽略一些 token 级（token-level）的内容。

比如看到图2具体的案例："The farm has 35 hens and 12 pigs. ##davidjl123 says totaling 47 animals."，其中包含像""这样的时间信息，以及"##davidjl123"这样的用户 id。这种 token 级的噪声在语料中较为常见，这些细粒度的噪声很难通过以前采用的文档级和行级过滤察觉。即使语料质量较高，中间仍然可能存在一些高度不确定性的 token，例如在图2的案例中，通过"12 pigs"和"35 hens"可以推测农场里一共有"47 animals"，但是在没有先验条件和上下文的前提下，让语言模型学习准确预测农场里有几只猪是困难的。

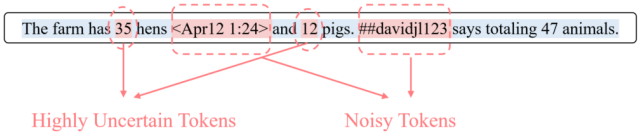


图2：Token 级噪声样例图

然而这些高度不确定的 token 和噪声 token 在因果语言建模中没有区分，以平均的权重参与到最终的模型更新中，会使语言模型感到困惑。

选择性语言建模（SLM）方法

为了进一步研究 token 级对于模型训练的影响，研究员们对语言模型预训练过程做了 token 损失（token loss）的动态分析。研究员们使用了 15B 的 OpenWebMath 语料来训练 Tinyllama-1B 模型，而且在每训练 1B 的 token 后于验证集上评估所有 token 的损失。通过获取所有检查点在验证集上的 token 损失数据，研究员们为验证集中的每个 token 拟合了损失的变化趋势，并重点关注训练初期和末期的 token 损失，以及训练前后的损失差值。基于训练前后损失差值和整体token平均的损失，研究员们将验证集中的 token 分为四类：H→H、L→H、H→L、L→L，分别代表 token 损失在训练过程中动态变化的趋势。（具体的划分依据如图3所示。）

$$\text{Token's Loss } (l_0, l_1, \ldots, l_n),\ f(a,b) = \text{minimize} \sum_{i=0}^{n} (l_i - (ax_i + b))^2$$

$$\mathcal{L}_{\text{start}} = b,\ \mathcal{L}_{\text{end}} = an + b,\ \Delta\mathcal{L} = \mathcal{L}_{\text{end}} - \mathcal{L}_{\text{start}},\ \mathcal{L}_{mean} = \frac{1}{n}\sum_{i=0}^{n} ax_i + b$$

Token Types:

- **H → H：** $(-0.2 \le \Delta\mathcal{L} \le 0.2 \text{ and } l_n > \mathcal{L}_{mean})$
- **L → H：** $(\Delta\mathcal{L} > 0.2)$
- **H → L：** $(\Delta\mathcal{L} < -0.2)$
- **L → L：** $(-0.2 \le \Delta\mathcal{L} \le 0.2 \text{ and } l_n \le \mathcal{L}_{mean})$

图3：Token 类别划分依据

H→H 代表一直保持较高损失的 token，L→L 则一直保持较低的损失。L→H 代表损失上升的 token，而 H→L 是最常见的变化趋势——token 损失下降。从图4(a)中可以看到，仅26%的 token 属于 H→L 类别，大多数 token 的损失变化不大，甚至有12%的 token 的损失呈现上升趋势。当研究员们随机采样 H→H 和 L→L 类别的 token 并单独观察其损失曲线（图4(b)和图4(c)）

时，可以发现这些 token 在训练过程中处于反复波动的状态，可能影响模型的收敛速度。因此，研究员们认为如果有一种方法可以合理选择适合学习且更有用的 token，让其参与训练，将可以减少噪声并提升模型的数据效率。

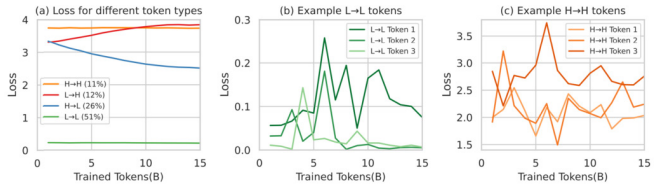


图4：Token 级损失的动态示意图

基于此，研究员们提出了选择性语言建模（Selective Language Modeling, SLM），在保证原有输入序列的情况下，通过在损失端裁剪模型所需的 token 损失，来选择有用的 token，如图5所示。

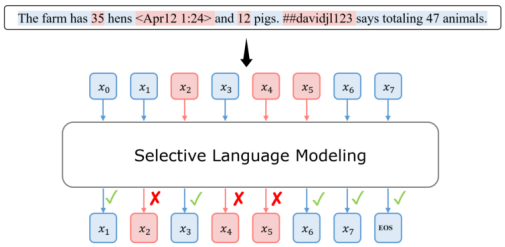


图5：选择性语言建模示意图

如何选择有用的token?

首先，需要有一个高质量的语料库。第一步，使用传统的因果语言建模损失在高质量的语料库上训练一个参考模型（reference model）来建模高质量 token 的分布。第二步，用训练好的参考模型在离线阶段对预训练语料中的每个 token 打分，最终得分由 Token Scoring 公式计算得到。这样的打分方式不会在实际的训练过程中引入额外的时间开销。第三步，用打好分的预训练语料训练模型，对每个 token 的分数排序后，取 topk% 作为选择的 token，对应 Token Selection 的公式。最后，通过 SLM 方式训练的损失公式迭代更新模型。

图6显示，在相同的数据集中，通过 SLM 方式训练的模型比直接训练的模型有效提高了数据效率，加快了收敛速度。

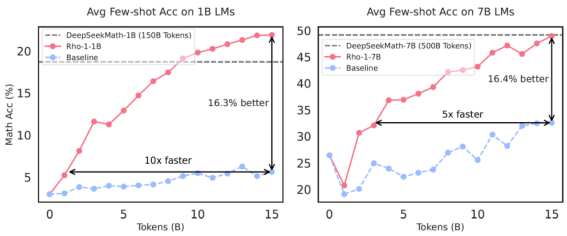


图6：选择性语言建模的数据效率

实验结果与应用

研究员们在数学领域进行了实验。实验中，研究员们使用 14B OpenWebMath 语料，分别在 Tinyllama-1B 和 Mistral-7 上继续预训练该模型，并采用 SLM 训练方式，选择比例分别为60%和70%。使用 SLM 训练的 Rho-1 Math 1B 和 7B 模型相较于直接继续预训练，性能分别提升了16%和10%。

为了进一步验证预训练的结果，研究员们基于上述训练的基础模型进行了推理微调对比实验。该模型在数学领域上取得了与 DeepSeekmath7B 相当的成绩，在数学基准上的准确率均高于 50%。值得注意的是，在预训练过程中，该模型仅使用了 14B 的 OpenWebMath，远少于 DeepSeekMath 使用的 120B 数学相关语料，进一步证明了使用 SLM 训练的数据效率。

研究员们也在通用领域（general domain）上进行了类似的实验，采用包含总计 80B token 的预训练语料对 Tinyllama1B 进行继续预训练。Rho-1 在各项通用基准测试中平均提升了约6%。

同时，本篇论文还探讨了在缺乏高质量语料作为参考的情况下，SLM 是否能够正常运作。如图7所示，可以直接将预训练好的基础模型作为参考模型进行自我参考（self-reference）迭代。在 Tinyllama 上继续预训练以验证自我参考的可行性。仅通过一轮迭代，SLM 就可以显著提升模型性能，这在模型的自我提升方面具有重要意义。

图7：自我参考流程示意图

总而言之，该研究表明并非所有 token 在语言模型预训练过程中都是同等重要的。通过基于 token 级数据筛选的 SLM 建模方式能够极大提高模型的数据效率。这种 token 级的思路不仅适用于预训练，还可以应用于微调、强化学习、多模态等领域。此外，选择 token 的形式可以多种多样，要根据具体场景及需求确定不同的 token 选择方法。研究员们希望未来能够出现更多有效的 token 选择策略和重新加权策略。

相关链接：

Rho-1: Not All Tokens Are What You Need
https://arxiv.org/abs/2404.07965

Github链接：https://github.com/microsoft/rho

NeurIPS 2024微软亚洲研究院论文合集

2024年12月，全球最负盛名的人工智能盛会之一 NeurIPS 大会在加拿大温哥华举办。在 NeurIPS 2024 大会上，来自微软亚洲研究院的多篇论文成功入选。其研究领域涵盖：提升大语言模型的能力、优化大语言模型及机器学习算法、负责任的人工智能、生成式AI与扩散模型、多模态与跨模态学习，以及领域特定的基础模型。

增强和提高大语言模型（LLMs）的能力与效率是推动人工智能技术进步的关键。在第一期 NeurIPS 2024 精选论文解读中，大家将了解到微软亚洲研究院的研究员们不仅通过提升 LLMs 的逻辑推理、鲁棒性和组合能力来拓宽其应用边界，从而应对日益复杂的任务，同时，也在探索提高速度和优化资源利用率的方法，使 LLMs 更实用、更易于被广泛采用。

扫描二维码查看文章

如今，生成式 AI 和扩散模型正成为 AI 内容创作的中坚力量。在第二期 NeurIPS 2024 精选论文解读中，大家将了解到微软亚洲研究院的研究员们如何提升生成式 AI 与扩散模型的效率以及多功能性，从而使其在多样化的应用场景中更加强大、稳健。与此同时，为了确保 AI 系统与人类的价值观和社会规范保持一致，研究员们还开发了可评估人工智能风险并推动伦理实践的工具，希望加强人工智能的治理与可信度。

扫描二维码查看文章

随着人工智能技术的不断进步，多模态和跨模态学习已成为AI领域的重要发展方向之一。在第三期 NeurIPS 2024 精选论文解读中，大家将了解到微软亚洲研究院的研究员们如何通过开发创新框架，加强不同信息模态间的协同作用，从而提升 AI 系统的有效性。同时，为了满足特定行业的精准需求，研究员们也开始定制领域特定的基础模型，以更好地捕捉行业知识，提高 AI 在各领域的精确度，为实现更精准、更个性化的解决方案提供了可能。

扫描二维码查看文章

一个基于3D虚拟形象的、用于口语到手语翻译的基线方法(ECCV 2024)

微软亚洲研究院开发了一个将口语翻译成手语的系统 Spoken2Sign。该模型由文本到手语翻译器、手语连接器和渲染模块三个组件构成,希望可以为听障人士和健听人士之间更便捷、更包容的交流做出贡献。

扩散模型是几何评估器:使用预训练扩散先验进行单图像三维编辑(ECCV 2024)

Diff3DEdit无需进行微调和额外的训练,巧妙地利用预训练的图像扩散模型所提供的先验知识,就可以实现单图像的三维编辑。

FontStudio:用于生成字体特效的形状自适应扩散模型(ECCV 2024)

FontStudio 框架通过形状自适应扩散模型(SDM)和形状自适应风格迁移(SAET),可以有效提高非矩形画布上的内容生成质量并保证风格的一致性。

通向精确视觉文本生成的定制化文本编码器Glyph-ByT5(ECCV 2024)

微软亚洲研究院通过开发 Glyph-ByT5 及融合方法,证实定制化文本编码器在解决扩散模型中视觉文本渲染问题上的可行性和必要性。

用于视频编解码器的长期上下文获取(ECCV 2024)

仅依赖短期上下文的做法限制了 NVC 在降低时间冗余方面的潜力。对此,研究员们提出了DCVC-LCG丰富上下文的多样性,助力提升重构质量并抑制误差传播。

RodinHD:基于扩散模型的高保真3D数字化身生成(ECCV 2024)

RodinHD是一种创新的数据调度策略和权重正则化技术。该方法有效地增强了共享解码器对细节渲染的能力。

扫描二维码查看文章

自我进化实现Rust自动形式化证明

微软亚洲研究院通过让大语言模型自我进化的方式,实现了 Rust 代码上的自动形式化证明。自我进化框架能够迭代式地合成更高质量的数据,并用于训练更强的能够生成形式化证明代码的模型。

基于图模式的理解基准测试

研究员们测试了7 种大语言模型在图模式识别中的能力,并总结出了一些重要发现。这些发现为未来基于大语言模型的图模型开发以及提升其对复杂逻辑结构的理解,提供了重要指导。

IGOR: 通过学习统一的动作表示空间让机械臂模仿人类动作

图像目标表示IGOR通过结合 LAM 和 World Model,成功地将一个视频中的物体运动"迁移"到了其他视频中,开辟了通过大量人类和机器人视频预训练学习动作表示并泛化到不同任务和智能体的新范式。

扫描二维码查看文章

对话松下康之：以具身智能突破人工智能与物理世界的边界

2024年10月，松下康之（Yasuyuki Matsushita）博士在离开近十年之后重返微软亚洲研究院。再次加入，松下康之有了一个全新的身份——微软亚洲研究院（东京）负责人。在此之前，他曾于2003年至2015年间在微软亚洲研究院任职，主要研究方向包括计算机视觉、机器学习和优化等，随后他转赴大阪大学担任教授。

作为微软研究院全球化战略的一部分，微软亚洲研究院东京实验室的成立将进一步巩固微软研究院在亚太地区的科研投入。为此，我们与松下康之进行了一次深入的对话，回顾了他的职业生涯，探讨了科学技术的发展演变，并展望了微软亚洲研究院（东京）令人期待的新机遇。

微软亚洲研究院（东京）负责人松下康之（Yasuyuki Matsushita）博士

重返微软亚洲研究院

Q：欢迎您回到微软亚洲研究院，并肩负起建设东京实验室的重任。您曾在微软亚洲研究院的北京实验室工作了十二年，之后又转战学术界担任教授。是什么促使您在十年后选择回归微软亚洲研究院？

松下康之：微软亚洲研究院一直走在科学研究的最前沿，特别是在当前的人工智能时代。今年初，我得知了研究院计划扩展其研究网络，包括在东京设立新的实验室，这对我来说是一个令人激动的新机遇，让我有机会在本地进行具有深远影响的研究，也能在国际舞台上施展才干。而且，微软在人工智能领域处于领先地位，这也是我重新参与其中的一个最佳时机。

我相信，在微软亚洲研究院这个世界级的科研平台上，我能够为人工智能这个蓬勃发展的领域注入新的能量。参与到人工智能的发展中并贡献一份力量，也让我感到倍感兴奋。

微软亚洲研究院过去十年的变与不变

Q：在回来的这一个多月里，从您的角度看，过去十年微软亚洲研究院有哪些变与不变？

松下康之：一个最直接的感受就是员工使用的工具和资源有了显著的变化。我目前还在熟悉这些先进的数字化系统，它们极大地提高了我们的工作效率，并促进了团队间的协作。这十年间，微软不仅推动了其他公司的数字化转型，自身也经历了深刻的变革。

除此之外，微软亚洲研究院的许多独特之处依然如故，比如持续为人才的成长营造创新和协作的文化与环境。研究院始终拥有并吸引着对科学研究充满热情的优秀人才。微软亚洲研究院最大的优势之一就是开放与协作的精神，这一点在和众多高校与研究机构建立的长期合作伙伴关系中也得到了体现。这种合作促进了跨地区、跨文化和跨学科的交流，激发了创新，并推动了产业的发展。对卓越的不懈追求一直是微软亚洲研究院的核心特质，这一点始终未变。

对微软亚洲研究院（东京）的规划

Q：随着微软亚洲研究院在温哥华、东京、新加坡和香港等地区的扩展，您作为微软亚洲研究院（东京）的负责人对东京实验室有哪些规划？微软亚洲研究院（东京）的建立将如何为亚太地区的创新生态做出贡献？

松下康之：当前我的首要任务是确保东京实验室的发展与微软研究院的使命保持一致，即推动科学技术的进步，造福人类。东京实验室的研究将与日本社会经济发展的重点相契合，并特别关注具身智能（embodied AI）、社会福祉与神经科学、社会责任人工智能（societal AI），以及产业创新等领域。这些领域的研究工作旨在解决当前社会面临的紧迫挑战，并将推动人工智能技术

的发展，使其惠及整个社会。

我们始终坚持开放的研究实践，通过发布并开源研究成果和工具，我们希望我们的工作能够使更广泛的行业从中受益，并为丰富全球知识库做出贡献。我们的目标是分享那些能够推动全社会进步和创新的深刻见解。

加速下一代人才成长

Q：人才成长与发展是微软研究院使命与文化的核心。微软亚洲研究院（东京）正在寻找具备哪些特质的人才？东京实验室将如何加大力度培养下一代科技创新人才？

松下康之：作为微软的一部分，我们的关键优势在于能够将研究与现实世界的应用紧密结合。这种研究与实践的桥梁可以确保我们的技术创新能够带来更加有意义且有益的成果，直接影响社会发展。

在招募新人时，我们寻找的是具有自驱力、强烈好奇心以及对解决社会挑战充满热情的伙伴。对复杂问题背后的"为什么"有强烈的求知欲是我们最看重的特质之一。虽然技术专长至关重要，但我们相信，致力于解决社会问题可以激发更多创造力并促进更多有意义的进步。而这种好奇心和使命感的结合更能激发创新，并推动微软亚洲研究院向前发展。

培养下一代科技创新人才也是微软亚洲研究院（东京）的核心愿景之一。我们将延续微软亚洲研究院过往成功孵化的人才培养项目，包括联合研究计划、访问学者项目和实习生项目等。这些项目为青年研究人员和学生提供了宝贵的实践经验，能够帮助他们掌握必要的科研技能，加深对复杂技术挑战的理解。

我们致力于营造一个有利于人才成长、协同合作并为全球科技发展做出贡献的科研环境。通过将技术创新与现实世界的需求紧密结合，我们希望可以激励新一代人才不断突破极限，推动社会进步。

计算机视觉领域的快速发展

Q：十年前您主要专注于光度学和视频分析领域的研究，能否分享一些当时的关键成果？您认为像人工智能这样的新兴技术对计算机视觉领域有什么影响？

松下康之：十年前我的研究重点集中在计算机视觉领域，特别是用于三维（3D）重建的光度学和提升视频质量的视频分析。在那个时期一个比较有影响力的项目是我们开发了一台能够捕捉高分辨率 3D 信息的十亿级像素相机。这台相机在敦煌莫高窟的文化保护中发挥了重要作用，它以前所未有的精度对敦煌壁画和佛像等文化遗产进行数字化保护。

另一个值得一提的项目是视频稳定技术，该技术作为媒体基础的一部分被集成到了 Windows 7 操作系统中。这项技术通过校正不必要的相机抖动提升了视频质量，实现了更流畅、专业的视频输出效果。在当时，能够实时处理并提升视频稳定性的算法是一项具有开创性的研究进展。

此后，深度学习、大规模数据集和复杂的神经网络架构的引入将计算机视觉推向了新的高度。曾经我们认为困难的任务，如目标检测、识别和分割，现在借助人工智能技术已经变得司空见惯。网络架构、学习策略的持续创新，以及增强的数据集，正在进一步拓展科学技术的边界。当前，一个令人兴奋的趋势是人工智能在现实世界交互场景中的应用，而这催生了具身智能这一新兴的研究领域，这也是我目前工作的一个重点。

理解具身智能：超越机器人技术

Q：您目前的研究方向主要是具身智能，那么请问具身智能是什么？它与我们熟悉的机器人技术有何不同？

松下康之：具身智能超越了传统机器人技术的范畴。机器人通常配有执行器，专门设计来完成特定的任务，而具身智能则侧重于开发能够执行复杂任务并在物理和虚拟环境中理解和交互的智能系统。过去，机器人技术和人工智能是两个相对独立的领域。具身智能则是这两个领域的融合，它可以将人工智能技术与能够在真实世界中感知、行动和学习的物理实体集成在一起。

本质上，具身智能是一个跨学科领域，它涉及机器人控制、强化学习、空间感知、人机交互和推理等多个方面。例如，具身智能拥有推断因果关系的能力，可以理解没有支撑的笔记本电脑会因重力而坠落。这种交互和认知能力源于对物理世界的接触和理解，因此，具身智能是一个令人兴奋且多面的研究领域。

正是由于具身智能的高度复杂性，没有任何单一组织能够独自覆盖其发展的所有方面。我们非常期望与工业界和学术机构合作，通过结合他们的专业知识和我们在人工智能领域的优势，共同推动具身智能领域的发展。

给计算机视觉与人工智能年轻科研人的建议

Q：您在学术界和工业界都有着深厚的研究经验。以教育者和科研人的双重身份，您对那些有志于投身于计算机视觉和人工智能研究的年轻人有什么建议？

松下康之：对于想要从事计算机视觉和人工智能领域研究的年轻人来说，扎实的数学基础和计算机科学知识是必不可少的。

即使在研究课题与技术演进日新月异的今天，这些核心技能依然至关重要，例如对梯度、雅可比矩阵和向量空间等基础数学概念的深刻理解。掌握了这些原理，无论编程语言还是开发平台如何更迭，你都能游刃有余。

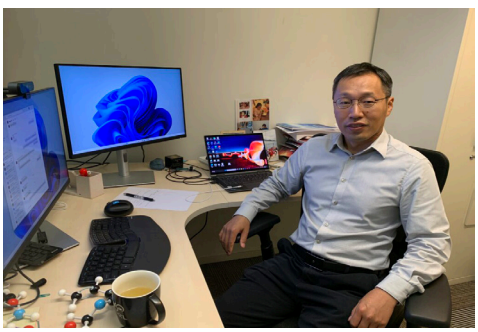
另外，持续学习的能力同样不可或缺，因为研究领域时刻都在发生变化。十年前，深度学习远不如今天这样重要，但现在它已经成为人工智能领域的基石。微软特别强调培养成长型思维，即要灵活适应变化、对新技术持开放态度，并随着行业的发展不断调整。新入行的研究者需要培养快速掌握新技能的能力，同时不断巩固基础知识，这种适应能力是在科研领域长期发展和取得成功的关键。

刘海广：发挥“生物多样性”法则的力量，寻找科学新答案

刘海广是一位典型的跨学科研究者，他的学术足迹遍布物理、生物、计算机科学和人工智能等多个领域，在跨学科的学习与研究中不断探索和突破。如今，作为微软研究院科学智能中心（Microsoft Research AI for Science）的首席研究员，刘海广在跨学科研究中取得了哪些成果？他和团队又将如何利用人工智能技术来加速科学研究进程，并推动科研成果向实际应用转化？

从清华大学转学至香港浸会大学，为什么会做出这个选择？熟悉刘海广的人都会对此感到好奇。“性格所致。我一直渴望探索未知，见识不同的世界。人生本来就很短暂，我们应该坚定地追随内心的信念。特别是当你对某件事充满热情时，无论选择哪条路，最终都会朝着内心的目标前进。所以无论是留在清华大学还是浸会大学，我最终依然会投身于科学研究，现在看来都是殊途同归的。”刘海广说道。

刘海广始终坚持“心之所想，力之所及”的做事准则，他坚信科学研究的目标是造福于人。2022年，带着“让科学研究成果在现实世界落地实用”的想法，刘海广从纯学术研究领域来到了微软研究院。在加入微软之前，他在学术道路上不断探索跨学科研究的深度与广度，而微软研究院则为他提供了一个更具创新性和影响力的平台。如今，作为微软研究院科学智能中心（Microsoft Research AI for Science）的首席研究员，刘海广正借助前沿的人工智能技术，加速科学研究进程，推动科研成果向实际应用转化，逐步将自己的愿景变为现实。



微软研究院科学智能中心首席研究员刘海广

探索科学领域的隐藏角落

刘海广可谓是跨学科研究的资深探索者。1999年，刘海广被清华大学录取，由此开启了他以物理学为起点的学术生涯。在清华的学习生涯刚刚开始，一次转变的机会就出现在了刘海广的面前——香港浸会大学面向内地高校招生。渴望“拓宽视野”的刘海广抓住了这次机会，转而在香港浸会大学继续他的物理专业学习，师从著名统计物理学家汤雷翰教授。

但不同的是，这一次他选择了一个融合了物理学与计算机科学的交叉专业，同时学习物理系和计算机系两个学科的核心课程。最后的毕业论文研究则将计算方法应用于了生物物理体系中，特别是针对蛋白质结构与动力学的研究。“这一专业融合了计算机、物理、生物学三个学科，让我自然而然地跨越了物理学，进入了计算机科学与生物学的交叉领域。”刘海广说。

大学毕业后，刘海广选择出国深造，在美国加州大学戴维斯分校获得了博士学位。在这一时期，他的研究侧重于运用物理思维来抽象生物学机制，并通过计算的方法模拟生物分子的动力学变化过程。随着对交叉学科研究的深入，刘海广发现这种跨学科研究存在一些局限性。

首先，受当时计算能力的限制，即使是用当时最先进的超级计算机也无法精确描述上万个原子组成的蛋白质分子体系，所以只能将其重要特征抽象为物理模型，但这在科学研究中不够严谨；其次，从计算物理角度出发的跨生物学科研究，对生物体系的抽象描述由于模型自身的缺陷，即便拥有强大的计算资源能够进行模拟，获得的结果的置信度和可靠性也不能替代实验测量。

因此，在博士毕业后的职业选择上，刘海广明确了目标：致力于跨学科研究。彼时，美国劳伦斯伯克利国家实验室正在研发一种方法利用 X 光探测生物分子结构和动力学变化，急需相应的计算方法解读数据，这正是刘海广能够发挥专长的机会。在伯克利国家实验室，他不仅运用计算方法将复杂的生物数据转化为有意义、有价值的模型，而且实验室的生物实验也加深了他对跨学科知识的理解。在这里，刘海广真正实现了从纯粹的计算研究向与实验相结合的交叉领域研究的转变。

（可能是）第一个在美国SLAC国家实验室吃螃蟹的人

沿着这一研究方向，刘海广又在亚利桑那州立大学物理系/BioXFEL 中心完成了博士后工作。在取得了一系列成果之后，他选择了回国发展，加入北京计算科学研究中心，组建了自己的科研团队，与国内外大学和研究机构合作，通过整合多种实验手段和计算方法，深入探索生物分子的结构、动力学和功能，持续在计算机与生物学领域进行跨学科的研究与创新。

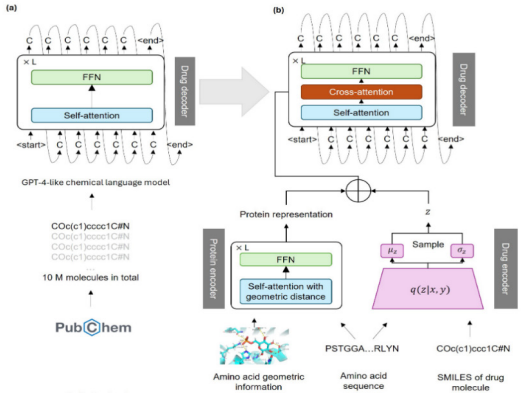
做实用主义的科学研究

随着计算能力的提升和人工智能技术的发展，计算方法与跨领域研究的结合日益紧密，显著加快了科学发现的步伐。在2015至2022年的七年时间里，刘海广带领团队取得了一系列突破性成果。然而，这些成果往往只停留在论文阶段。"我更希望将这些研究成果转化为实用工具，让更多人受益。但在纯粹的学术界，这些研究很难直接应用于实践。"刘海广说。

"微软研究院提供了一个既能进行科学技术研究，又能紧密对接工业需求的绝佳平台，让我们可以专注于科学研究的探索、技术产品的转化及其实际应用。例如，科学智能中心的一些研究成果已经转化为工具，并在微软 Azure 云平台上部署，使更多人能够受益于创新技术。"这便是刘海广眼中的微软研究院，也是他选择加入的原因。

目前，微软研究院科学智能中心的研究方向涵盖了面向自然科学的人工智能模型、基于人工智能的科学计算和科学发现等多个领域，具体包括药物研发、生命科学、材料科学等。刘海广的研究重点集中在与人类健康紧密相关的药物研发领域：基于靶点蛋白质的化合物设计以及对蛋白质本身的深入研究。

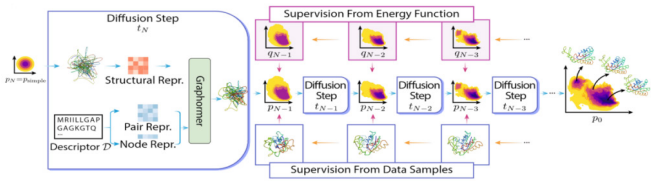
针对特定靶点蛋白的化合物设计，刘海广及团队开发了化合物研究工具 TamGen，相关论文已被《自然-通讯》（Nature Communications）杂志接收。与通过计算筛选预定义化合物库的虚拟筛选方法相比，基于靶点蛋白结构的从头药物设计能提供新的候选药物。TamGen 集成了 Transformer 靶标感知分子生成器的生成解决方案，能够直接从零开始为特定靶点生成候选药物，扩大了现有化合物的筛选范围，为新药发现提供了更多可能性。



TamGen 示意图

对于蛋白质本身的研究，刘海广及团队专注于蛋白质结构及其动力学性质，以及氨基酸突变性质的探索。在药物研发中，对靶点蛋白的深入理解至关重要，因为一旦靶点蛋白发生突变，可能导致药物效果降低甚至完全失效。传统的分子动力学模拟方法在分子结构采样分析时存在局限，往往只能捕捉到分子的初始状态，无法全面描述其在不同条件下的行为。

针对这一问题，刘海广和同事们开发了用于分子结构平衡分布预测的深度学习框架 Distributional Graphormer（DiG）。这项创新技术能够快速生成多样化的分子构象，为从单一结构预测到平衡分布预测的突破奠定了基础。DiG 的开发不仅为分子科学开辟了新的研究方向，也为药物设计和材料科学领域提供了新的工具。相关研究成果已发表在《自然-机器智能》(Nature Machine Intelligence)杂志上。



Distributional Graphormer 示意图

此外，作为一名拥有丰富跨学科研究经验的研究员，刘海广积极致力于推动不同领域间的合作。微软研究院科学智能中心会定期举办科学论坛，邀请全球的专家学者分享领域知识和前沿研究成果，促进学术界与工业界的交流与合作，加速科技成果的落地应用。

2024年5月，科学智能中心与西湖大学联合举办了一场学术

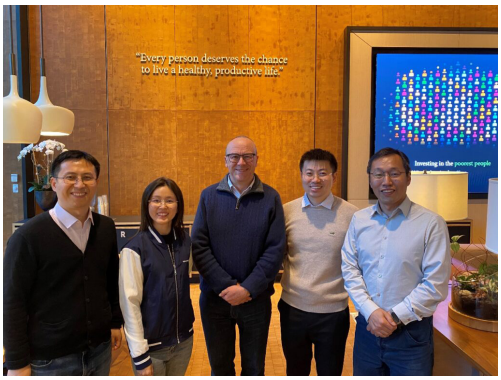
研讨会，双方分享了在药物研发、生命科学和材料科学领域的最新成果。“我们还与全球健康药物研发中心（GHDDI）合作，利用 TamGen 平台为肺结核和冠状病毒等全球性传染病设计高效的新候选药物，为治疗这些疾病提供了新的希望。”刘海广介绍道。

AI for Science：加速跨学科研究的智能引擎

在 AI for Science 的研究中，人工智能的最大优势在于其提升效率和知识提取的能力。微软研究院科学智能中心运用机器学习技术，显著加快了药物研究和材料发现的计算速度，实现了数十甚至数百倍的提升。“在我读博期间，一次仿真模拟需要两三个月才能完成，现在一两天就能得出结果。”刘海广说，“在信息爆炸的今天，人工智能和大语言模型还能帮助我们更快地发现和理解其他领域的知识，并定期更新我们的知识库。人工智能像一位博学多才的助手，能够触类旁通，这对于跨学科和跨领域的研究至关重要。”

刘海广认为，人工智能技术在生物研究的各个环节都能发挥重要作用，促进我们对分子结构和相关数据的深入解读。以眼睛感光蛋白——视黄素蛋白的研究为例，这种蛋白位于细胞膜上，能在光照下利用光能完成对离子或者信号的传输。研究视黄素蛋白的工作原理需要依赖多种实验方法，包括结构生物学、光谱学、分子动力学模拟和高精度超快的显微拍摄技术等。而现在，人工智能可以应用于这些研究的每个环节，提升计算效率，加速科学发现的进程。

与此同时，人工智能作为一项融合了多学科和多领域知识的技术，无论是在其自身发展过程中，还是在与各行各业的跨界融合中，都迫切需要跨领域人才的支持。微软研究院汇聚了来自不同领域和背景的顶尖人才，并与众多高校和企业保持深入合作，不仅推动了创新技术在现实世界中的应用，更有助于推动跨领域的科学研究。



刘海广（右一）与同事们访问盖茨基金会

“这里的研究员充满活力且富有个性，我们可以随时交流和辩论，这种多样性和包容性是一个优秀研究机构不可或缺的特质。这就如同细菌群落需要保持多样性一样，如果群落仅由单一类型的细菌构成，当环境变得不利时，很容易导致整个群落的消亡。相反，生物多样性的存在意味着只有部分类型的细菌会受到影响，而细菌群落整体仍能存续。正是具备这种韧性和多样性，微软研究院才得以在过去的三十多年中一直稳步发展，并始终保持着创新的活力。”刘海广说。

相关链接：

Distributional Graphormer：从分子结构预测到平衡分布预测
https://www.microsoft.com/en-us/research/articles/distributional-graphormer/

TamGen: Target-aware Molecule Generation for Drug Design Using a Chemical Language Model
https://www.biorxiv.org/content/10.1101/2024.01.08.574635v2.full.pdf

机器之心 | 简单而强大：DIFF Transformer降噪式学习，开启模型架构新思路

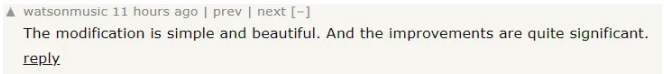
Transformer 模型对大语言模型以及人工智能发展所带来的革命性意义不言而喻。近期，微软亚洲研究院提出了一种全新的 Transformer 架构 DIFF Transformer（差分 Transformer）。通过差分注意力机制，DIFF Transformer 能够增强对关键信息的关注，同时减少对噪声的干扰，从而在多项语言任务中取得了显著优于 Transformer 模型的性能提升。

DIFF Transformer 与此前微软亚洲研究院发布的 BitNet（b1.58）、Q-Sparse 和 YOCO 等工作，正交且互补。研究员们致力于从基础研究角度为大语言模型的发展带来变革，为大语言模型的理论研究以及未来的实际应用带来更多新的可能性。

Transformer 的强大实力已经在诸多大语言模型（LLMs）上得到了证明，但该架构远非完美，也有很多研究者致力于改进这一架构，比如机器之心曾报道过的 Reformer 和 Infini-Transformer。

今天我们又将介绍另一种新型 Transformer 架构：Differential Transformer（差分 Transformer，简称 DIFF Transformer）。该架构来自微软亚洲研究院和清华大学，有四位共一作者：叶天竺、董力、夏雨晴、孙宇涛。

在 Hacker News 及 Twitter 等社交网络上，该论文都反响热烈，有网友表示差分 Transformer 提出的改进简单又美丽，而带来的提升又非常显著。



甚至已有开发者做出了差分 Transformer 的轻量实现！

那么差分 Transformer 弥补了原生 Transformer 的哪些问题呢？Transformer 往往会过度关注不相关的上下文，该团队将此称为注意力噪声（attention noise）。而差分 Transformer 则能放大对答案范围的注意力并消除噪音，从而增强上下文建模的能力。这就要用到该团队新提出的差分注意力机制（differential attention mechanism）了。

差分注意力机制可以消除注意力噪声，鼓励模型重点关注关键信息。该方法类似于电气工程中的降噪耳机和差分放大器。

下面我们就来详细了解一下差分 Transformer 的设计思路。

差分 Transformer

差分 Transformer 是一种用于序列建模的基础模型架构。为了方便说明，他们使用了仅解码器（decoder-only）模型作为示例来描述该架构。

该模型堆叠了 L 个 Diff Transformer 层。给定一个输入序列 x，将输入嵌入打包成 X^0。输入会被进一步上下文化来获得输出 X^L。每一层都由两个模块组成：一个差分注意力模块和之后的前向网络模块。

相比于 Transformer，差分 Transformer 的主要差别在于使用差分注意力替换了传统的 softmax 注意力，同时保持整体宏观布局不变。此外，他们也参考 LLaMA 采用了 pre-RMSNorm 和 SwiGLU 这两项改进措施。

差分注意力

差分注意力机制的作用是将查询、键和值向量映射成输出。这里使用查询和键向量来计算注意力分数，然后计算值向量的加权和。

此处的关键设计是使用一对 softmax 函数来消除注意力分数的噪声。具体来说，给定输入 X，首先将它们投射成查询、键和值 Q_1、Q_2、K_1、K_2、V。然后差分注意力算子 DiffAttn（·）通过以下方式计算输出：

$$[Q_1; Q_2] = XW^Q, \quad [K_1; K_2] = XW^K, \quad V = XW^V$$
$$\text{DiffAttn}(X) = (\text{softmax}(\frac{Q_1 K_1^T}{\sqrt{d}}) - \lambda \, \text{softmax}(\frac{Q_2 K_2^T}{\sqrt{d}}))V \tag{1}$$

其中 W^Q、W^K 、W^V 是参数，λ 是可学习的标量。为了同步学习动态，将标量 λ 重新参数化为：

$$\lambda = \exp(\lambda_{\mathbf{q_1}} \cdot \lambda_{\mathbf{k_1}}) - \exp(\lambda_{\mathbf{q_2}} \cdot \lambda_{\mathbf{k_2}}) + \lambda_{\text{init}} \tag{2}$$

其中 λ_q1、λ_k1、λ_q2、λ_k2 是可学习的向量，λ_init ∈ (0, 1) 是用于初始化 λ 的常数。该团队通过经验发现，设置 λ_init = 0.8 − 0.6 × exp (−0.3 · (l − 1)) 在实践中效果很好，其中 l ∈ [1, L] 表示层索引。它在实验中被用作默认策略。

他们也探索了另一种初始化策略：对所有层使用相同的 λ_init（例如0.8）。如后面消融研究所示，使用不同的初始化策略时，性能相对稳健。

差分注意力利用两个 softmax 注意力函数之间的差来消除注意力噪声。这个想法类似于电气工程中提出的差分放大器，其中两个信号之间的差用作输出，这样就可以消除输入的共模噪声。此外，降噪耳机的设计也基于类似的想法。

- 多头差分注意力机制

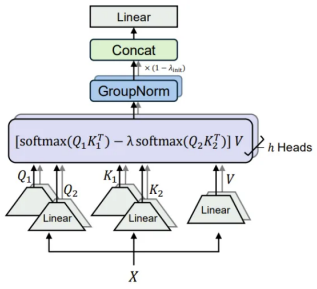
该团队也为差分注意力使用了多头机制。令 h 表示注意力头的数量。他们对各个头使用不同的投影矩阵 W^Q_i 、W^K_i 、W^V_i ，i ∈ [1, h]。标量 λ 在同一层内的头之间共享。然后对头输出执行归一化，并投射成最终结果，如下所示：

$$\begin{aligned} \text{head}_i &= \text{DiffAttn}(X; W_i^Q, W_i^K, W_i^V, \lambda) \\ \overline{\text{head}_i} &= (1 - \lambda_{\text{init}}) \cdot \text{LN}(\text{head}_i) \\ \text{MultiHead}(X) &= \text{Concat}(\overline{\text{head}_1}, \cdots, \overline{\text{head}_h}) W^O \end{aligned} \tag{3}$$

其中 λ_init 是 (2) 式中的常数标量，W^O 是可学习的投影矩阵，LN (·) 是对每个头使用 RMSNorm，Concat (·) 的作用是沿通道维度将头连接在一起。这里使用一个固定乘数 (1 − λ_init) 作为 LN (·) 的缩放尺度，以使梯度与 Transformer 对齐。

- 逐头归一化

下图使用了 GroupNorm (·) 来强调 LN (·) 独立应用于每个 head。由于差分注意力往往具有更稀疏的模式，因此头之间的统计信息更加多样化。为了改进梯度的统计情况，LN (·) 算子会在连接操作之前对每个头进行归一化。




```
def DiffAttn(X, W_q, W_k, W_v, λ):
    Q1, Q2 = split(X @ W_q)
    K1, K2 = split(X @ W_k)
    V = X @ W_v
    # Qi, Ki: [b, n, d]; V: [b, n, 2d]
    s = 1 / sqrt(d)
    A1 = Q1 @ K1.transpose(-1, -2) * s
    A2 = Q2 @ K2.transpose(-1, -2) * s
    return
      (softmax(A1) - λ softmax(A2)) @ V

def MultiHead(X, W_q, W_k, W_v, W_o, λ):
    O = GroupNorm([DiffAttn(X, W_qi, W_ki,
          W_vi, λ) for i in range(h)])
    O = O * (1 - λ_init)
    return Concat(O) @ W_o
```

整体架构

其整体架构会堆叠 L 层，其中每层包含一个多头差分注意力模块和一个前向网络模块。如此，便可将差分 Transformer 层描述为：

$$Y^l = \text{MultiHead}(\text{LN}(X^l)) + X^l \tag{4}$$

$$X^{l+1} = \text{SwiGLU}(\text{LN}(Y^l)) + Y^l \tag{5}$$

其中 LN (·) 是 RMSNorm，SwiGLU (X) = (swish (XW^G) ⊙ XW_1) W_2，且 W^G、W_1、W_2 是可学习的矩阵。

实验

该团队从以下角度评估了差分 Transformer 在 LLMs 中的应用，包括对比评估、应用评估和消融研究。这里我们仅关注实验结果，更多实验过程请访问原论文。

语言建模评估

该团队评估了差分 Transformer 的语言建模能力。为此，他们使用 1T token 训练了一个 3B 大小的差分 Transformer 语言模型，并与之前的 Transformer 语言模型做了比较。

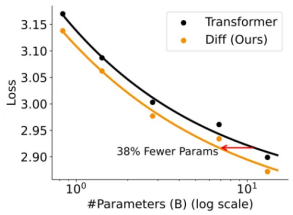
结果见下表，其中报告的是在 LM Eval Harness 基准上的零样本结果。

Model	ARC-C	ARC-E	BoolQ	HellaSwag	OBQA	PIQA	WinoGrande	Avg
Training with 1T tokens								
OpenLLaMA-3B-v2 [13]	33.9	67.6	65.7	70.0	26.0	76.7	62.9	57.5
StableLM-base-alpha-3B-v2 [39]	32.4	67.3	64.6	68.6	26.4	76.0	62.1	56.8
StableLM-3B-4E1T [40]	—	66.6	—	—	—	**76.8**	63.2	—
DIFF-3B	**37.8**	**72.9**	**69.0**	**71.4**	**29.0**	**76.8**	**67.1**	**60.6**

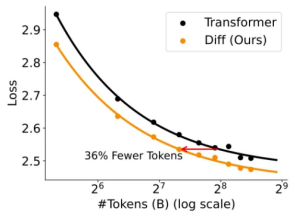
可以看到，3B 规模下，差分 Transformer 语言模型的表现优于之前的 Transformer 语言模型。此外，实验表明差分 Transformer 在多种任务上胜过 Transformer，详见原论文附录。

与 Transformer 的可扩展性比较

该团队也比较了新旧 Transformer 的可扩展性。其中 a 比较了模型规模方面的可扩展性，而 b 则是训练 token 数量方面的可扩展性。



(a) Scaling model size ranging from 830M to 13B.

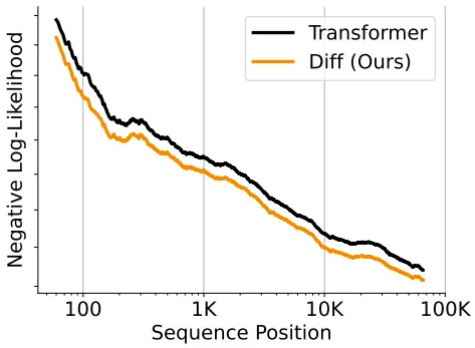


(b) Scaling number of training tokens for 3B models.

可以看到，在这两个方面，差分 Transformer 的可扩展性均优于常规 Transformer：仅需后者65%左右的模型大小或训练 token 数量就能达到相媲美的性能。

长上下文评估

当 3B 模型上下文长度增长至 64K，模型的表现又如何呢？又使用另外 1.5B token 训练了 3B 版本的检查点模型之后，该团队发现随着上下文长度的增加，累积平均负对数似然（NLL）持续下降。差分 Transformer 得到的 NLL 值低于常规 Transformer。见图4，这样的结果表明，差分 Transformer 可以有效地利用不断增加的上下文。



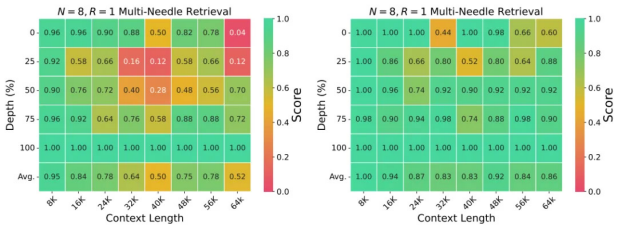
关键信息检索

为了检验差分 Transformer 检索关键信息的能力，该团队执行了 Needle-In-A-Haystack（草堆找针）测试。

下表给出了 4K 上下文长度的情况，其中 N 是针的数量，R 是查询引用的数量。可以看到，差分 Transformer 的多针检索准确度高于常规 Transformer，尤其是当针数量较多时，差分 Transformer 的优势会更加明显。

Model	$N=1$ $R=1$	$N=2$ $R=2$	$N=4$ $R=2$	$N=6$ $R=2$
Transformer	**1.00**	0.85	0.62	0.55
DIFF	**1.00**	**0.92**	**0.84**	**0.85**

那么当上下文长度提升至 64K 时，又会如何呢？结果见图5，这里使用的上下文长度在 8K 到 64K 之间，使用了 N = 8 和 R = 1 的设置。



(a) Transformer. (b) DIFF Transformer.

可以看到，在不同的上下文长度下，差分 Transformer 能够保持相对稳定的性能。而当上下文长度越来越大时，常规 Transformer 的性能会逐渐下降。

另外，下表展示了分配给关键信息检索任务的答案范围和噪声上下文的注意力分数。该分数可代表模型保留有用信息、抵抗注意力噪声的能力。

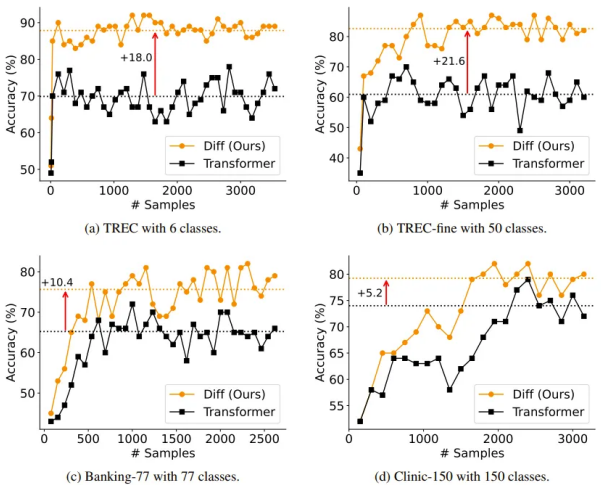
Model	Attention to Answer ↑ 0%	25%	50%	75%	100%	Attention Noise ↓ 0%	25%	50%	75%	100%
Transformer	0.03	0.03	0.03	0.07	0.09	0.51	0.54	0.52	0.49	0.49
DIFF	**0.27**	**0.30**	**0.31**	**0.32**	**0.40**	**0.01**	**0.02**	**0.02**	**0.02**	**0.01**

可以看到，相比于常规 Transformer，差分 Transformer 能为答案范围分配更高的注意力分数，同时为注意力噪声分配更低的注意力分数。

上下文学习能力评估

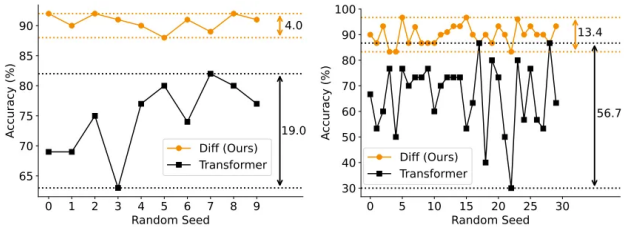
该团队从两个角度评估模型的上下文学习能力，包括多样本分类和上下文学习的稳健性。

下图展示了新旧 Transformer 模型的多样本分类结果。结果表明，在不同的数据集和不同演示样本数量上，差分 Transformer 均稳定地优于 Transformer。此外，差分 Transformer 的平均准确度优势很明显，从5.2%到21.6%不等。



(a) TREC with 6 classes. (b) TREC-fine with 50 classes.
(c) Banking-77 with 77 classes. (d) Clinic-150 with 150 classes.

下图则展示了两种模型的上下文学习稳健性结果。该分析基于 TREC 数据集，并且采用了两种提示词格式：示例随机排列和按类别交替排列。



(a) Examples are randomly arranged. (b) Examples are arranged alternately by class.

在这两种设置下，差分 Transformer 的性能方差要小得多。结果表明，新方法在上下文学习任务中更为稳健。相比之下，Transformer 容易受到顺序排列的影响，导致最佳结果与最差结果之间差距巨大。

上下文幻觉评估

该团队基于文本摘要和问答任务评估了模型的上下文幻觉现象。

Model	XSum	CNN/DM	MultiNews
Transformer	0.44	0.32	0.42
DIFF	**0.53**	**0.41**	**0.61**

(a) Accuracy (i.e., free of hallucinations) on text summarization datasets.

Model	Qasper	HotpotQA	2WikiMQA
Transformer	0.28	0.36	0.29
DIFF	**0.39**	**0.46**	**0.36**

(b) Accuracy (i.e., free of hallucinations) on question answering datasets.

可以看到，相比于常规 Transformer，差分 Transformer 在摘要和问答任务上的上下文幻觉更低。该团队表示，原因可能是差分 Transformer 能更好地关注任务所需的基本信息，而不是无关上下文。

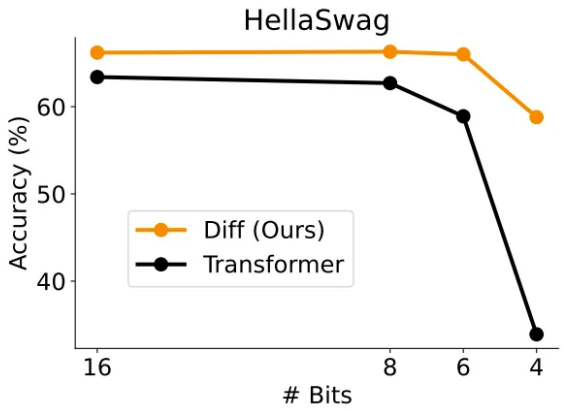
激活异常值分析

在 LLMs 中，一部分激活值明显大于大多数激活值的现象被称为激活异常值（activation outliers）。异常值导致训练和推理过程中模型量化困难。实验表明差分 Transformer 可以降低激活异常值的幅度，从而可能实现更低的量化位宽。

下表展示了两个训练得到 Transformer 和差分 Transformer 模型的激活值统计情况。这里分析了两种类型的激活，包括注意力 logit（即 pre-softmax 激活）和隐藏状态（即层输出）。可以看到，尽管中位数相似，但与 Transformer 相比，差分 Transformer 的较大激活值要低得多。这表明新方法产生的激活异常值较少。

Model	Activation Type	Top-1	Top-2	Top-3	Top-10	Top-100	Median
Transformer	Attention Logits	318.0	308.2	304.9	284.7	251.5	5.4
DIFF	Attention Logits	38.8	38.8	37.3	32.0	27.4	3.3
Transformer	Hidden States	3608.6	3607.4	3603.6	3552.1	2448.2	0.6
DIFF	Hidden States	1688.2	1672.5	1672.1	1624.3	740.9	1.2

下图则展示了将注意力 logit 量化到更低位的情况。这里使用的方案是：使用 absmax 量化的动态后训练量化。其中，16位配置表示未经量化的原始结果。模型逐步量化为8位、6位和4位。这里报告的是在 HellaSwag 上的零样本准确度，但该团队也指出在其它数据集上也有类似表现。

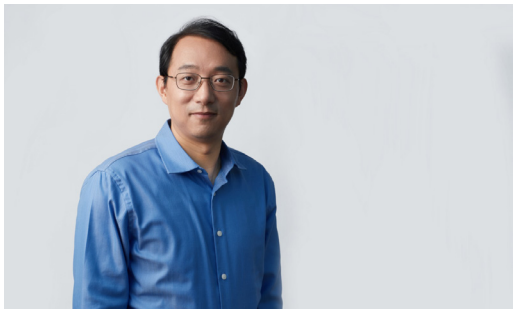


从图中可知，即使降低位宽，差分 Transformer 也能保持较高性能。相较之下，常规 Transformer 的准确度在6位和4位量化时会显著下降。这一结果表明，差分 Transformer 本身就能缓解注意力分数中的激活异常值问题，从而可为低位 FlashAttention 的实现提供新机会。

最后，该团队进行了消融实验，证明各个新设计的有效性。·

相关链接：

Differential Transformer
https://arxiv.org/pdf/2410.05258



周礼栋

微软全球资深副总裁
微软亚太研发集团首席科学家
微软亚洲研究院院长

如今，我们正处在孕育新一代计算范式的关键节点。在不久的将来，虚拟世界和现实世界的边界会不断消弭，计算会像电力一样无处不在。新的计算范式将赋能人类生活和工作的方方面面，给各行各业带来颠覆性的变革，也将催生众多新的机遇。

面对科技发展的新浪潮，微软亚洲研究院将践行所有利于激发创新力的原则，持续致力于营造多元、包容、自由、平等、开放、可持续的研究氛围和科研协作环境，让各种具有创造性的想法、观点和创意，在微软亚洲研究院这个“化学反应池”中交流、碰撞、提炼和升华，使创新的星星之火形成燎原之势。同时，我们也将保持积极开放的姿态，与国内外各界伙伴携手，共同推动技术进步，实现人类社会的可持续发展。

关于微软亚洲研究院

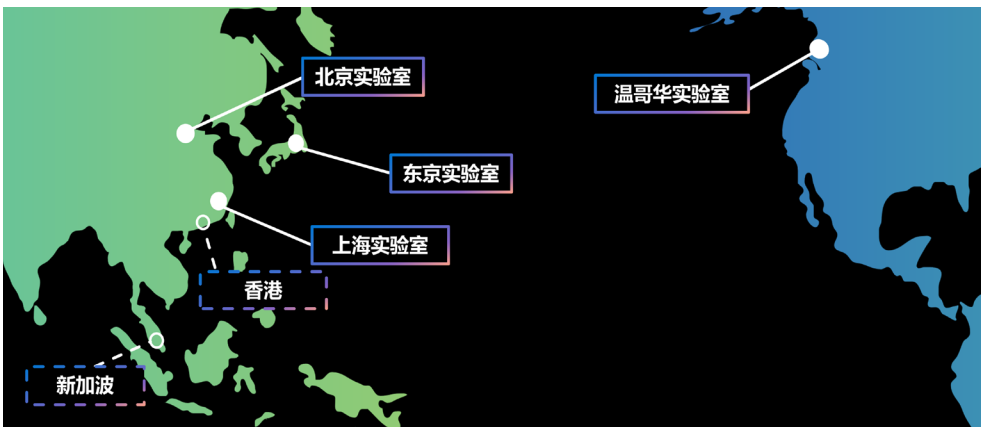
微软亚洲研究院成立于1998年，是微软公司在亚太地区设立的研究机构，在北京、上海、温哥华、东京、新加坡和香港设有实验室及研究岗位，研究方向涵盖计算基础创新、下一代智能交互、多维感知与通信、人工智能与社会福祉、科学发现与行业赋能等。通过来自世界各地不同学科和背景的多元人才的鼎力合作，微软亚洲研究院已经发展成为世界一流的计算机基础及应用研究机构。多年来，从微软亚洲研究院诞生的新技术层出不穷，对微软公司的产品创新以及全球范围的科技发展产生了深远的影响。

作为微软研究院全球体系的一员，微软亚洲研究院拥有广阔的国际视野，同时融合了东西方创新文化的精髓。秉持开放合作的理念，微软亚洲研究院始终与高校和科研机构开展持久而有效的合作，推动跨地区、跨文化和跨学科的交流，激发创新潜力，促进行业发展。

微软亚洲研究院倡导对技术进步怀有远大抱负，推崇富于冒险的极客创新精神，鼓励研究人员拓展研究的深度与广度，跨越计算机领域的界限，把视野拓展到解决具有广泛社会意义的问题上，为未来的计算新范式奠定基础，并为AI和人类发展创造更美好的未来。

扫描二维码观看视频介绍

微软亚洲研究院实验室分布



Microsoft

微信　　*知乎*

电话：86-10-59178888
网址：https://www.microsoft.com/en-us/research/lab/microsoft-research-asia-zh-cn/
微博：http://t.sina.com.cn/msra