

AI4S 团队以新思路解决模型数据稀缺难题

“真题有限的情况下,学生可以通过做模拟题提高成绩,类似的逻辑在基础模型训练上也成立。”针对部分垂直领域基础模型训练中的高质量数据稀缺难题,计算机科学技术学院颜波教授团队创新性地提出了一种实现数据高利用率的基础模型训练框架,利用可控生成式人工智能(AIGC)生成合成数据,并采用“真实数据+合成数据”的混合数据模式训练基础模型。依托该框架训练的医疗基础模型性能表现优秀。

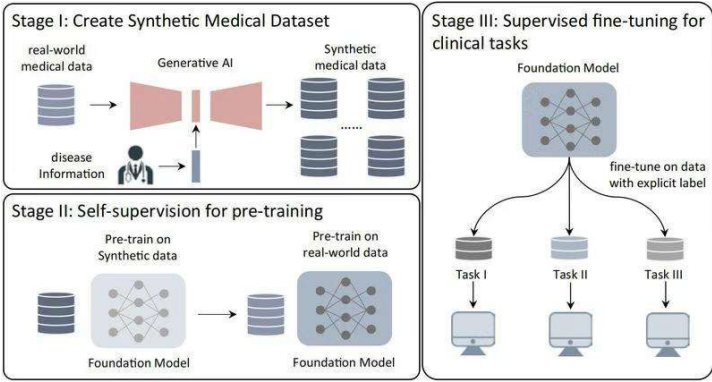
相关成果日前在《自然 生物医学工程》(*Nature Biomedical Engineering*)上,以题为《数据高效的高性能医学基础模型训练策略》(A data-efficient strategy for building high-performing medical foundation models)的论文发表。该策略利用生成式AI生成大规模合成数据集,从而扩展有限的真实预训练数据集,为解决数据稀缺难题提供了新思路。

在数据有限的情况下开发基础模型

基础模型是一种基于深度神经网络和自监督学习(SSL)技术,在大规模、广泛来源数据集上训练的AI模型。相较于只能完成特定任务的专用AI模型,基础模型的独特之处在于其强大的泛化能力。通过大规模数据预训练提取通用特征,垂直领域的基础模型可以完成各种不同下游任务。

然而,大规模高质量数据的获取面临挑战。以医疗为例,收集大规模真实数据集不仅成本高昂、耗时漫长,还具有隐私泄露风险。在恶性肿瘤诊断、工业焊缝瑕疵检测等数据稀缺场景,传统基础模型训练方法效果受限,难以推广。

在数据有限的情况下,如何开发基础模型?在AIGC领域深耕10余年,颜波团队将目光瞄准了AI合成数据,采用大量合成数据让模型学习,弥补现有真实数据的不足,实现让模型“训得



■ 数据高效的医学基础模型训练框架

好”。团队在少量公开的真实医学数据上微调可控生成式AI,整合特定疾病知识,并以疾病概念作为条件生成大规模合成医学数据集。医学基础模型先后在合成数据和有限真实数据上使用SSL技术预训练,以初始化模型参数并学习精确医学表征。最后,团队通过带明确标签的监督微调基础模型,使其适配特定任务。

让基础模型“见多识广”

如何确保合成数据在模型训练中的有效性?一方面,合成数据均基于真实数据标签生成,与真实数据特征非常接近;另一方面,团队在训练过程中引入条件混合增强,使数据特征更加多样。颜波用“做模拟题”比喻这一过程:

“真题是有限的,但我们可以根据真题编写更加灵活、多变的模拟题。让学生先做模拟题了解题型、再做真题,学生考试就能考得更好。脱胎于真实数据、具有丰富特征的合成数据可以增加数据集的多样性,基础模型也能‘见多识广’,从而提升性能。”

团队遵循该方法构建了首个基于合成数据的视网膜基础模型RETFound-DE,使用16.7%的真实数据即在多项疾病诊断任务和《自然》杂志报道的RETFound模型性能相当。使用20%真实数据构建的胸片X光基础模型CXRFM-DE同样展现了较好的性能和泛化能力,进一步验证了

该训练框架的有效性。

该创新可在多领域应用

除了智慧医疗之外,该创新基础模型训练方法为所有因数据限制而难以构建基础模型的领域提供了新的解决方案,目前已经在工业界瑕疵检测、监控异常分析、新材料设计、和船舶发动机智能制造等产业实际痛点中得到初步应用。高质量数据缺乏问题是所有AI领域均面临的重要问题。颜波介绍,“在算法层面尝试解决小样本问题效果有限,在数据层面进行突破为这一问题提供了新思路。未来,这一理念将为智能制造、智慧医疗等各行各业提供支持。”

该研究还对推动合成数据的应用提供了启发。团队验证了合成数据在泛化能力、标注效率和训练效率上的价值,并详细探索了合成数据在均衡样本分布、合成数据量控制、克服数据偏差等方面的使用方式。尽管合成数据在AI基础模型训练中展现出巨大潜力,但需注意过度依赖合成数据可能影响泛化能力,并引入数据偏差,确保合成数据在AI模型训练中的正确使用,需要有效的验证和监管。

该工作由计算机科学技术学院数字媒体实验室完成。

原文链接:

<https://www.nature.com/articles/s41551-025-01365-0>

实习记者 曾译萱

我校牵头发布长三角气温指数

由我校大气科学研究院赵艳霞研究员牵头编制的长三角气温指数(系列)3月22日发布。这一创新性金融气象指数的落地,标志着我国气象数据要素与金融市场的深度融合进入新阶段。

当前,全球气候变化正引发极端天气频发,进一步加剧了金融和相关行业的物理风险和转

型风险。在金融领域,从农业生产季节性波动,到能源消费的峰谷变化,再到资本市场定价模型的修正需求,气象因子已深度嵌入经济运行脉络。气象数据正在成为金融市场定价的重要变量,通过构建科学量化的气温指数,可以有效对冲企业经营中的“气候不确定性”。

来源:大气与海洋科学系

为锂电池开发与优化提供关键见解

高分子科学系汪莹课题组展示高效处理多目标优化的创新方法,为锂电池材料的开发与优化提供了关键见解。该成果以“Multi-Objective Optimization of Ionic Polymer Electrolytes for High-Voltage Fast-Charging and Versatile Lithium Batteries”为题近期在线发表于(*Advanced Materials*)。

随着便携设备、电动汽车和大规模电网对高能量密度、快速充电电池的需求增长,传统液态电解质因易燃性和高压下的副反应(如界面副反应、气体析出、过渡金属溶解等)受到限制。离子聚合物电解质(IPEs)兼具安全性和界面适应性,但其实现多目标优化(如离子电导率、电化学窗口、循环稳定

性)需要在高维化学空间中探索,传统试错法效率低、成本高。

在锂电池材料探索中,高效识别最优离子聚合物电解质(IPEs)配方是一个关键挑战。团队采用结合核函数高斯过程(GP)与多种采集函数的方法,在仅探索2.8%的化学空间后,快速锁定了最优的TFSI基IPEs配方,这种多目标优化方法显著降低了实验成本。这一配方在高压和快速充电条件下表现出优异性能,为开发高性能锂电池提供了创新性路线。

论文链接:

<https://advanced.onlinelibrary.wiley.com/doi/10.1002/adma.202500941>

来源:高分子科学系

社科基金优秀项目

调查编目日本所藏中国古旧地图

由中华古籍保护研究院副研究员陈照任项目负责人的《日本公藏中国古旧地图之现状与特色》获评2025国家社科基金鉴定优秀项目。

日本公共图书馆、大学图书馆、公文书馆和博物馆等公共机构收藏了大量中国古旧地图,对于研究中国历史尤其是近代中国具有重要的参考价值。本项目系统地调查和收集了日本2264家公共机构所藏、成图于1949年以前的中国地

图信息,经过整理和校勘,共获得18022幅/册地图信息,将版本相同以及同一系列地图进行归并后,最终汇总为包含4795幅/册地图信息的联合目录。对日本所藏中国古旧地图这一重要资源进行调查和编目,有助于国内外学界和社会各界了解日本公藏中国古旧地图情况,为地图史学等相关学术研究提供参考工具。同时,也有助于推进海外地图类中国古代文献的回归和保护工作。

为全球化跨文化对话提供历史镜鉴

由外国语学院语言文学学院教授包慧怡任项目负责人的《英国中世纪世界地图中的“东方叙事”图文研究》获2025年度国家社科基金冷门绝学专项结项鉴定优秀。

地图从来不只是测绘学的产物,而是地图制作者或委任者历史地理观、政治宗教观、地域边疆观、族裔文化观的载体。本项目立足原文(拉丁文、古英语、中古英语、古法语等),通过对大量中世纪地图和同时期文学手稿的首次翻译和解

读,图文互动地探究这些异域想象背后的语言思想观、史地风俗观、宗教伦理观等。首先提出“叙事地图”作为分析框架,主张将中世纪地图视为叙事主体,而非辅助性地理工具,试图在传说与事实的缝隙间梳理线索,为不同类型地图上的异域叙事勾勒轮廓并研究其成因。通过揭示英国中世纪文化中的偏见与权力关系,为中世纪研究和全球化视野下的跨文化对话提供了参考框架。

来源:文科科研处

报道光热气泡的3D周期性轨道运动模式

航空航天系邓道盛/胡曼课题组近日报道了光热气泡在正己烷中的三维(3D)周期性轨道运动现象,相关成果以“3D Periodic Orbiting of a Photothermal Bubble”为题发表于国际权威物理学期刊 *Physical Review Letters*。该研究采用的实验体系是近红外激光透过高导热的蓝宝石玻片加热正己烷液体,通过两个正交的高速相机观测到了正己烷中光热气泡的3D轨道运动模式,通过理论建模和数值模拟,揭示了正己烷中气泡的3D轨道运动机理是双向光热Marangoni流和驻点流(热浮力流碰

撞固壁面形成)协同作用的结果。

气泡在液体中的运动现象在日常生活和工业应用中非常常见,并且这种两相流通常会产生有趣的动力学现象。近年来,由于外部刺激操控的水下气泡在药物输送、光刻、热管理、微组装和微流体芯片等领域有广泛应用,揭示气泡的运动模式与液体的热力学和流体力学特性之间的内在关联,对于理解复杂流动机理和拓展相关气泡技术十分必要,成为研究热点。邓道盛/胡曼课题组前期利用近红外激光对液体介质的体相加热作

用,结合固壁面的高效热传导,分别实现了气泡在纯水介质中的一维(1D)垂直弹跳运动和乙醇中的二维(2D)振荡运动模式。

这项研究不仅揭示了光热气泡在不同液体中的复杂物理化学流体动力学,还为气泡介导技术的发展提供了新的视角,例如,在微流体和微操纵领域,可利用气泡的多维度多模式移动操控提高效率。

全文链接:

<https://link.aps.org/doi/10.1103/PhysRevLett.134.104004>

来源:航空航天系