数据开源为AI发展“推波助澜”

K 热点透视 rediantoushi

在6月9日—10日举行的2023北京智源大会上，“AI数据开源”引发广泛关注。AI数据为什么要开源？AI数据开源面临哪些挑战？它会是未来AI发展的重要趋势吗？科技日报记者带着这些问题采访了相关专家。

AI数据开源意义重大

有专家认为，AI数据开源对深度学习模型的发展意义重大。由于训练AI大模型需要大量资源，所以预计“赢家通吃”类AI系统的开发和管理将首先由少部分闭源实体所主导。

但遗憾的是，这种资源限制导致研究人员、非营利组织和初创公司等小规模实体因无法承担高昂的成本，几乎不可能从零开始训练自己的AI大模型。

以对话类模型为例，目前国内外众多已经开源的对话模型，其实都是基于语言基础大模型，再利用少量指令微调数据进行训练所得。

如果开源AI大模型的数据在质量上具有足够的竞争力，深度学习模型的规模化训练和运行成本将大幅降低。

北京智源人工智能研究院（以下简称智源）副院长兼总工程师林咏华对记者表示，大模型是AI未来发展的重要方向，其研究和应用将逐步成为AI发展的关键方向，并有望形成新一波AI推广浪潮，而AI数据开源将进一步促进大模型的发展。

深度学习需要大量的标注数据进行模型训练。在林咏华看来，过去10年，深度学习技术快速发展的重要原因，就是许多志愿者团体、国外科研团队一直在积极地收集、整理并开源用于深度学习的训练数据集。“当前AI大模型训练对数据量的需求，比之前的深度学习小模型对数据量的需求有了百倍，甚至千倍的提升。所以，尤其在过去一年，数据开源的问题日益受到广泛关注。”林咏华说。

背后挑战不容忽视

开源固然会为AI发展带来诸多好处，但其背后的挑战也不容忽视。其中之一，便是开源安全与合规挑战。林咏华认为，对传统的商业软件而言，开源中的安全、合规、许可证和代码质量风险等是使用开源组件必须面临的挑战。然而在AI大模型时代，更大的挑战则在开源数据集方面。

因此，AI数据开源应在协议许可的范围内进行。“用于AI大模型训练的开源数据必须是合法地从公开或可公开获得的资源中收集的数据。人们可以在开源协议允许的范围内，以AI大模型训练、AI算法开发为目的，对数据进行访问、修改和使用。部分数据可能要求使用过程中遵守更严格的协议。”林咏华表示。

此外，今天的基础AI大模型不只具备理解能力，还具有生成能力，它能够对外进行认知输出、价值观输出等，可能给社会带来巨大影响。“我们在训练基础大模型的时候，所使用的预训练数据会对AI生成内容质量起到很大程度的决定性作用。因此，开源数据的质量十分重要。”

林咏华指出，由于高质量的数据（如文章、图片、视频等）通常有版权，由于版权或商业因素导致的闭源以及数据孤岛等挑战会制约AI的发展，所以需要多方推动构建更多高质量的开源数据集，尤其是用于训练基础AI大模型的开源数据集。

LF AI & DATA基金会董事主席堵俊平对此也深有感触：“AI大模型就像一个贪吃的‘怪兽’，始终需要研究人员投喂更多的、质量更好的数据。”他说，当前数据几乎都是从“在网络上主动收集”“从第三方购买”“利用公开数据集”这三个渠道得来。在堵俊平看来，从第一个渠道得到的数据局限性较强，由于版权问题，很多公司只能从其私域获得数据；从第二个渠道获取的数据面临数据定价、数据质量等问题；而从第三个渠道获取的数据往往只能作为研究使用，在商用或者其他方面有很多限制。

开源渐成AI发展重要趋势

记者了解到，智源对2023年1月到5月底发布的、具有影响力的语言模型进行过统计。统计结果表明，国外发布的开源语言模型有39个，国内发布的开源语言模型有11个。

■ 视觉中国供图

“开源是推动AI技术进步的重要力量，AI开源开放生态及平台建设也日益受到重视。开源开放毫无疑问已经成为重要的AI发展趋势之一。”林咏华表示，“开源能够促进AI大模型科研创新，推动和降低AI大模型落地乃至整个AI产业落地的门槛。”

然而，通往开源的道路并非一帆风顺，在数据之外，算力也是开源路上的一只“拦路虎”。AI大模型训练依赖庞大的数据、算力。训练参数量级的增长使得算力需求也随之增长，算力集群正变得愈发庞大。

然而算力成本却是小型开发者的“不可承受之重”。拿到AI大模型开源数据后，往往需要对其进行微调和二次开发。但现实的情况是，对一些小型开发者来说，仅仅是做推理都很困难，就更别提对AI大模型做微调、二次开发。以ChatGPT为例，仅就算力而言，Open AI为了训练它，就构建了由近3万张英伟达V100显卡组成的庞大算力集群。有消息称，Open AI公司发布的新一代语言模型GPT-4甚至达到了100万亿的参数规模，其对应的算力需求同比大幅增加。

目前，有一些研究机构希望用技术的革新抵消巨大的算力成本。最直接的手段是通过训练技术的革新加快AI大模型推理速度、降低算力成本、减少能耗，以此来提高AI大模型的易用性，让开源数据更好地发挥价值，但这只能从工程上对算力资源的约束起到缓解作用，并非终极方案。

有业内专家表示，解决算力问题最终还是要回到AI大模型自身寻找突破点，一个十分被看好的方向便是稀疏大模型。稀疏大模型的特点是容量很大，但只有用于给定任务、样本或标记时，模型的部分功能才会被激活。也就是说，这种稀疏大模型的动态结构能够让AI大模型在参数量上再跃升几个层级，同时又不必付出巨大的算力代价，一举两得。

此外，开源社区的作用同样不容忽视。开源社区是推动开源发展的重要基石，开源的最初发源点，就是来自于社区开发者的贡献。“Linux系统的成功很大程度上得益于开源社区。30多年来，Linux系统发展成为拥有海量全球用户的操作系统，其成功以及长久不衰的秘诀就是开源，尤其是内核社区成千上万开发者的贡献。”林咏华举例说。

“开源开放可以使得我们站在前人的肩膀上前行。”林咏华总结道，“这些年AI领域取得的成果大多受益于开源，如果没有开源，AI不会发展到今天。”

裴宸纬

K 创新杂谈 chuangxinzatan

让科研人员大胆探索

■ 北湖

前不久在烽火通信科技股份有限公司采访时了解到，为应对数字时代网络传输容量大、速率高等挑战，该公司创造性地提出了“智慧光网”的新理念。智慧光网前人没做过，技术开发周期较长、投入大，有一定的失败风险。该公司以研究贡献考核科研人员，不看研发过程一时成败，只要输出成果有价值，就会得到认可。在这一机制鼓励下，科研人员大胆尝试，矢志攻关，最终让智慧光网由理念变成了现实，如今正加速落地，实现产业化。

这不禁让人联想到“手撕钢”的攻关故事。“手撕钢”薄如蝉翼，工艺控制难度大、产品质量要求高，中国宝武山西太钢不锈钢精密带钢有限公司在突破相关关键技术时，困难接连不断。比如，“手撕钢”经过260米长的带钢通道时，最容易出现抽带断带。有时，抽带断带一周出现十几次，造成不小损失。一次次失败让科研人员很受挫。紧要关头，该公司明确考核“新规”，不仅宽容失败，而且只要攻关过程中取得进步就给予及时激励。这让科研人员重拾信心，最终在历经700多次失败后，实现了“手撕钢”量产。

创新从来都充满荆棘，原创性研究的难度更大，试想如果“只能成功，不许失败”，可能会让许多科研人员望而却步。智慧光网和“手撕钢”的创新实践说明，让科研人员卸下包袱、轻松上阵、迈开步子，能够充分释放创新潜能、实现“从0到1”的突破。着力培育鼓励创新、宽容失败的科研生态，能够促使科研人员大胆探索，挑战未知，敢碰“九死一生”的真问题，并形成不惧失败、追求成功的创新风尚。

实际上，科研本身是个不断尝试、长期积累的过程，“失败”并非没有意义。包信和院士团队潜心研究20多年提出了“纳米限域催化”新概念，成功的背后有许多意想不到的失败；赵忠贤院士团队找到令同行振奋的铁基高温超导材料，其过程同样是“山重水复疑无路，柳暗花明又一村”。如同钱学森先生所说的：“正确的结果，是从大量错误中得出来的，没有大量错误作台阶，也就登不上最后正确结果的高座。”

当前，我国重大创新成果竞相涌现，一些前沿领域开始进入并跑、领跑阶段，科技实力正在从量的积累迈向质的飞跃，从点的突破迈向系统能力提升。从模仿式的追随转向开拓性的引领，需要广大科技工作者以更为广阔的视野、更加自觉的使命担当，勇闯“无人区”，努力实现更多“从0到1”的突破。重大原创性突破和关键核心技术攻关难度大、周期长，把鼓励创新、宽容失败的好政策好机制落实落细，让全社会理解科学、支持创新，将会收获更多高质量的研发成果。

“东方故乡——中华大地百万年人类史”展览开展


图为观众观看古生物和古人类化石。■ 洪星摄

科学导报讯 6月22日，中国科学院古脊椎动物与古人类研究所（以下简称中科院古脊椎所）和中国国家博物馆携手打造的“东方故乡——中华大地百万年人类史”展览在北京开展。展览依托中科院古脊椎所与国家博物馆收藏的220余件（组）文物，结合各类场景还原和多媒体技术手段，展现了近百年来中国在古人类学、旧石器考古学和古DNA研究等方面的最新成果。

整个展览共分为物竞天择、矗立东方、智慧灵长、现代之路4个单元。“物竞天择”部分以古脊椎动物演变历程展现生物多样性及与环境的依存关系；“矗立东方”部分以元谋人、蓝田人、泥河湾人、北京人等重要发现展示中国直立人的演变形态；“智慧灵长”部分展现具有承上启下意义的智人的发展历程；“现代之路”借助基因组数据成果，展示现代人的直系祖先，即早期现代人的演变格局。

中科院古脊椎所所长邓涛介绍，近年来，该所在国内率先将高清晰度CT、同步辐射、三维激光扫描、数字图像分析等新技术手段应用于古人类研究，革新了本领域的研究范式，大大推进了该领域研究进展。尤其是古核基因组捕获技术的突破与应用，使大规模研究古DNA成为可能，在世界范围内推动了人类学、演化遗传学等相关学科发展。

孙明源

脚踏实地结硕果 知识产权管理成果获评国家优秀案例

——访中国知名知识产权管理专家于丽萍

作为一名专业的知识产权经理人，荣获“2020年度中国杰出知识产权经理人”称号无疑成为于丽萍职业生涯中一个值得铭记的时刻。但在她内心深处，却更看重另一项荣誉，那就是凭借其知识产权管理方面的成绩，她领导下的彩虹鱼海洋科技公司入选了“2021年度知识产权信息公共服务优秀案例”。

业绩成果入选国家优秀案例

年度知识产权信息公共服务优秀案例是中国国家知识产权局为总结和推广知识产权公共服务先进做法和典型经验、更好地服务创新创业主体而发起的一项遴选活动。这项覆盖全国、权威性极高的活动，有着非常严苛的筛选、复评标准。最终入选的典型案例都有着非常好的创新和示范引领作用，用以直接推动知识产权管理制度建设，帮助完善国家知识产权公共服务体系，提升知识产权管理效能，助力创新成果惠及更广大人群。这项荣誉可以说是所有知识产权管理人和团队梦寐以求的重磅荣誉。

2021年11月25日，由国家知识产权局主办的“全国技术与创新支持中心（TISC）及高校国家知识产权信息服务中心交流研讨活动”在北京成功举办。国家知识产权局副局长何志敏、教育部科技发展中心副主任张拥军出席活动并讲话，世界知识产权组织（WIPO）中国办事处主任刘华通过视频致辞。会议上公布了2021年度知识产权信息公共服务优秀案例，彩虹鱼公司以“凝视深渊，挑战极限——上专助力彩虹鱼科创之旅”为主题报送的知识产权成果案例，最终从全国98家机构报送的153个案例中脱颖而出成功入选。这不仅是当年上海市唯一获此殊荣的案例，也是当年全国唯一入选的中小型科创企业知识产权创造、保护和运用的优秀案例。

这项荣誉的获得代表彩虹鱼公司的知识产权管理工作具备了一定的专业性和系统性，进一步提升了公司的知识产权管理水平和核心竞争力，为企业正在筹备的科创板上市提供了强有力的支撑，对全国拟上市科创板企业也有着很好的参考价值。

荣誉不是终点，不断迈向新高度

获得这项荣誉，是对于丽萍及其团队专业能力的直接认可，也对未来的发展提出了更高的要求。于丽萍说：“获得这个荣誉我们非常高兴。说明我们的工作得到了认可。另外，我想说的是，这份荣誉的获得，既要感谢团队的努力，也要特别感谢我们的合作方上海专利商标事务有限公司。最终得到这个认可，说明我们此前的工作方法和方向是正确的。这个荣誉给了我们很大的鼓舞，也是彩虹鱼公司知识产权管理团队一个全新的起点。未来我和我们团队会一如既往地做好专业工作，我们要用自己的点滴努力，激励并发掘创新的闪光点，用经过我们的挖掘和引导产生的创新，争取为公司、为社会做出更多扎扎实实的贡献。”

过去几年彩虹鱼公司的知识产权工作在于丽萍的带领下，聚焦公司核心技术，在知识产权规范化挖掘、管理、运用及保护等方面表现出显著的成效，覆盖到产品提供、销售、采购、研发、管理等各个环节，形成了一套专业化的体系。而且于丽萍始终坚持对专利质量矢志不渝地追求，坚定地追求创新，才得以令她和她的团队成为行业的标杆。

彩虹鱼公司作为深渊科技领域当之无愧的龙头企业，肩负着推动行业科技发展甚至人类福祉、人类文明发展的使命和能量。于丽萍心中所想的，就是让公司的知识产权管理团队能够匹配彩虹鱼公司的行业地位，争取为彩虹鱼公司寻求更大的发展空间。于丽萍始终保持着对知识的敬畏之心，用她自己的话说：“在点滴的知识进步中，世界才能逐渐变得更好。我希望能在自己的岗位上真正地有所贡献，哪怕在一个小小的位置，我也会心怀使命、心怀梦想，通过自己的努力和创造，尽可能地发挥出自己的生而为人的禀赋和能量。”于丽萍是这么说的，也是这么做的。

资料显示，彩虹鱼知识产权管理团队是于丽萍一手创立的，多年来，她一边着手建立彩虹鱼知识产权管理团队和运营体系，一边脚踏实地地深入一线实操知识产权挖掘、整理、申请、保护等工作。在繁芜丛杂的工作之中，于丽萍始终保持着勤奋务实、乐于分享、乐于助人的精神，用她的思考、勇气、努力、谦逊和真诚，不仅赢得了同事的信赖和尊重，也很快便打造出一支专业、高效的知识产权管理团队。

在于丽萍卓越的个人魅力和团队管理能力之下，彩虹鱼公司的知识产权管理团队一路乘风破浪，砥砺前行。在知识产权管理体系建设方面，于丽萍充分发挥了彩虹鱼公司在深海科技尤其是深渊科技领域的技术优势和科研资源，为公司制定了一套行之有效的知识产权创造、运用、保护全链条管理体系。

截至目前，围绕公司的核心技术或主营业务，彩虹鱼公司已经实现了海洋物联网系统集成、数据采集、监测检测、智能硬件设备等八大类核心技术100多件的核心专利布局，获得80余项授权专利，其中发明专利占比50%以上，形成了初具规模的知识产权矩阵和壁垒。这种高质量的产权管理成果为彩虹鱼公司在企业项目申报、合同招投标、高新技术企业申请等诸多方面提供许多助力。其中的多项专利已经产生了数以百万计的经济价值，为企业的良性发展提供了坚实的基础和创新活力，具有深远的品牌效益和社会效益。 **王丽**

于永生：用科技铸就“中国重载第一路”辉煌

（上接A1版）机车、车辆、可控列尾既是3个独立的装备，分别都有自己的技术参数和设计标准，又共同组成了一个系统。于永生说：“每个独立装备的个体最优并不能达到系统的最优。这就需要我们站在系统的角度提高三者的适配性，调整3个独立装备的技术标准、参数。但要调整哪个装备、调整到什么程度，这个装备的工艺、成本是否支持，需要在太原局集团公司重载铁路技术研究中心这个以需求为牵引的合作平台上不断协调才能完成。”

于永生介绍，这个旨在破解制约重载铁路发展难题、推进科技成果转化和应用的平台，通过与多家高等院校、科研院所、装备生产制造企业深度合作，聚合了强大的人才和研发优势。

作为项目牵头人，于永生从系统最优角度提出要求和目标后，中南大学、大连交通大学、中车齐齐哈尔车辆有限公司、眉山中车制动科技股份有限公司、克诺尔公司等参研合作单位会共同跟进，不断优化设计标准、生产工艺、技术参数，让项目向着整体最优的目标趋近。

“科研既无捷径，也无坦途，发挥协同作用能够大幅提高科研效率和成果转化率。”体验过数学模型误差过大不达预期的失落，也曾陷入试验失败找不出症结的僵局，更无数次在深夜愧疚无法陪伴家人左右，但于永生还是执着坚守年少的梦想，共参与了“重载铁路技术升级深化研究”“重载组合列车自动驾驶技术研究”等10余项重载课题、“重载铁路技术升级深化研究”等4项国铁集团重大课题、“基于MPC的重载组合列车差异化控制研究技术研究”等8项太原局集团公司重载专项课题。

如今，于永生依旧执着于重载铁路技术研究与应用，全身心扎在大秦铁路上，同时也把科研的种子播撒在铁道线上，让科技报国的理想在守护重载铁路安全的实践中落地生根、开花结果。

“外星人”科技：数智赋能助山西焦企降本增效

（上接A1版）为解决这一瓶颈，公司从全国各地引进了涉及工业自动化、数字孪生、大数据开发挖掘、软硬件集成、人机交互等专业技术人才，团队也从开始的十几人发展到现在的一百余人，吸引了大量优质人才留晋、归晋。

人才的聚集、平台的落成，使山西省煤焦化工企业的盈利能力和抗市场风险能力进一步增强。公司助力的项目已成为全国焦化行业领先的数字工厂代表，同时实现着产品和生产持续优化的高效循环。

“未来，企业将进一步加大技术研发力度，深耕全链路数据采集及数据治理，也必将助力煤焦化企业提质、降本、增效，推进山西煤焦化行业数字转型升级和智慧化运营更上一层楼。”石永刚信心满满地说。