大模型发展亟需高质量“教材”相伴

AI世界

◎本报记者 罗云鹏

1月5日，美国人工智能公司OpenAI表示，正在与数十家出版商洽谈达成文章授权协议，以获取内容来训练其人工智能模型。2023年12月27日，《纽约时报》起诉OpenAI和微软公司，指控这两家公司未经许可使用其数百万篇文章训练人工智能模型。而早在2023年3月，就有消息显示谷歌Bard模型的部分训练数据来源于ChatGPT。

这些事件剑指同一个问题——大模型高质量语料短缺。“对于从头开始训练的模型，语料短缺会在非常大的程度上限制大模型发展。”近日，哈尔滨工业大学（深圳）计算机科学与技术学院教授邵睿在接受科技日报记者采访时说：“增加语料对于提升大模型能力的边际效益正在减弱，高质量语料的缺乏正日益成为限制大模型发展的瓶颈。”

大模型发展如火如荼，但高质量语料的缺乏正日益成为大模型发展的瓶颈。 视觉中国供图

大模型训练语料短缺问题严重

科技部新一代人工智能发展研究中心2023年发布的《中国人工智能大模型地图研究报告》显示，从全球已发布的大模型数量来看，中国和美国大幅领先，占全球总数的80%以上。

虽然大模型发展如火如荼，但大模型高质量语料短缺已成为全球共性问题。公开资料显示，大模型对数据供给要求极高。比如，训练GPT-4和Gemini Ultra大概需要4万亿至8万亿个单词。麻省理工学院等高校研究人员预测，到2026年之前，机器学习数据集可能会耗尽所有可用的高质量语料数据。研究机构EpochAI亦公开表示，最早在2024年，人类就可能会陷入训练数据荒，届时全世界的高质量训练数据都将面临枯竭。OpenAI也公开表达过对数据告急的担忧。

值得注意的是，当前大模型数据集主要为英文。中文语料面临的短缺问题更加严峻。

中国工程院院士、鹏城实验室主任高文曾公开表示，全球通用的50亿大模型数据训练集里，中文语料占比仅为1.3%。

上海数据交易所市场发展部副总经理章健此前公开表示，当前大模型行业存在语料供应不足的问题，特别是在垂直细分领域，一些共享、免费下载的语料数量虽然大，质量却不高。“我们在追求语料数量增长的同时，也要重视质量。”章健说。

高质量语料应具备七大特征

那么，何为高质量语料？记者采访时，包括腾讯、商汤科技、哈尔滨工业大学（深圳）等企业和高校专业人士均给出一致答案：高质量语料应具备多样性、大规模、合法性、真实性、连贯性、无偏见和无害等七大特征。

邵睿表示，高质量语料应具有多样性高、句式流畅的特点。腾讯机器学习平台算法负责人康战辉认为，语料的多样性是保证语料质量的基础，要通过不同的途径采集新闻、小说、诗歌、科技文章等不同类型的语料。这有助于大模型学习到更丰富的语言表达。

同时，高质量语料要具有较大规模，因为大模型需要大量语料来学习语言规律并提高泛化能力。只有拥有充足的语料，大模型才能更好地捕捉细微的语言特征。

此外，高质量语料应是合法且无害的。不合法或有害的语料可能导致模型产生不恰当的回答或建议，或无意中泄露他人隐私。

“高质量语料还应该具有真实性和连贯性，以便让大模型更好地理解语境并生成符合逻辑的回答。”康战辉说，语料库应该充分反映语料的多样性并避免偏见，这样大模型在不同场景下回答不同用户的问题时才能做到尽可能科学客观。

完善相关机制提高语料质量

记者在采访中了解到，目前训练大模型的语料有一部分是从数据公司购买的，也有一部分是从网络公开语料或者公开数据集中获取并整理的。“从数据公司购买的语料质量较高，并且大多是垂域数据，但其数据量较少且价格较贵。”邵睿说，“网络公开语料通用性较好，数据量大，但数据质量无法保证，数据格式难以统一。”

“人类产生的有效信息，包括大量高价值信息可能不一定是互联网数据，而是沉散在各行各业里的数据。”商汤科技发言人说，“怎样更多汇聚数据，设计更多、更好的网络结构，用更多的计算资源去支撑更大容量的高质量语料，产生更强的智能，是一个至关重要的问题。”这位发言人认为，要解决语料问题，不仅要靠增加语料总量，还需要提高语料质量，甚至要考虑完善数据交换机制，推动人工智能数据基础设施化。

正如这位发言人所说，目前业界正在采取一些措施，推动数据交换机制的建设。记者梳理发现，2023年7月，深圳数据交易所联合近50家单位成立开放算料联盟。该联盟将围绕高质量中文训练数据和多模态训练数据，协调数据要素、数据治理、训练数据、数据标注、合成数据等相关标准制定，协助数据交易所增加与大模型相关的新品类和新专区。

同样是2023年7月，在2023世界人工智能大会现场，中国大模型语料数据联盟成立。同年8月，上海人工智能实验室宣布，联合中国大模型语料数据联盟成员单位共同开源发布“书生·万卷”1.0多模态预训练语料。这次开源的数据总量超过2TB，包含超5亿个文本、2200万个图文交错文档、1000个影像视频。

除了建设更为完善的体制机制，数据清洗等技术手段也能在一定程度上解决高质量语料短缺难题。但要看到，这些技术手段有较高门槛。商汤科技发言人透露，该公司在数据清洗的过程中投入了上千块GPU的算力。OpenAI在无数场合介绍过GPT-4训练的经验，但从未公开过数据清洗的经验。

评论

牢牢把握人工智能时代中国内容“生成权”

◎赵 运 饶高琦

2023年被誉为“生成式人工智能之年”，以ChatGPT为代表的大模型在当今社会中的作用愈发凸显。模型的效能和准确性在很大程度上取决于其训练所依赖的语料。根据公开信息，ChatGPT的大部分语料来自政府公告、新闻报道、科技论文、经典文学作品、历史档案、艺术作品等公开领域内容。这些语料在很大程度上反映了美国社会和文化领域的主流观点，也决定了大模型输出的内容。

若大模型的训练语料充斥着具有偏见性的内容，生成的内容也可能带有同样的偏见。

“傅满洲”原是西方小说中的虚构角色，后被美国电影界采用，被描绘为典型的东方反派。其外貌枯瘦、留有八字胡和长指甲，阴险狡猾。当这种具有文化偏见的内容被大量纳入模型训练语料库时，生成式人工智能就极可能输出这些有偏见的观点。如在外国一些AI作画程序中，输入“华裔”或“亚裔”作为引导词，就可能会生成类似“傅满洲”这样的形象。

在可以预见的未来，当人们不可避免地需要依赖人工智能生成的内容时，生成内容的可信度和价值观将深刻影响使用者乃至整个社会的思维导向。偏见和误解不仅会在人工智能生成的内容中被复制和放大，还可能加深人们对特定文化群体的刻板印象，使人们陷入“语种信息茧房”。

因此，在人工智能时代，应当更加重视中国知识、中国文化、中国故事在公开领域的中文表达。这不仅可确保大模型训练语料的真实性、准确性和多样性，也是守护我国文化表达权和文化阐释权的关键，更是牢牢把握住人工智能时代中国内容“生成权”的关键。

【作者赵运系中图科信数智技术（北京）有限公司战略规划部经理，作者饶高琦系北京语言大学国际中文教育研究院副研究员】

前海深港人工智能算力中心启动

科技日报讯（记者**罗云鹏**）记者1月12日获悉，前海深港人工智能算力中心已于日前启动。该算力中心是大湾区最大规模、算力最先进的智算中心，由前海管理局、商汤科技、香港科技园公司三方共同推进，前海科技创新集团与商汤科技联合投资建设。

据悉，该算力中心一期算力规模达500Petaflops（每秒500千万亿次浮点运算），AI算力规模每秒50亿亿次，相当于一小时可完成16亿张图像处理、190万小时语音翻译、0.7万公里自动驾驶AI数据处理。在对外服务方面，该算力中心采取模型即服务模式，可为不同客户提供定制化的模型服务，实现产业精准赋能。

近年来，人工智能产业飞速发展，智能算力需求不断攀升，智算中心作为人工智能产业发展的底座和基石备受关注。

据悉，前海深港人工智能算力中心不仅将提供高效、安全、可靠的算力资源，还将针对产业发展、社会应用提供算法模型以及数据资源。

“人工智能在2023年实现了跨越式发展，进入以大模型为基础的AI 2.0时代。先进智能算力作为当前最具活力的新型生产力，已经成为重要的战略资源。”商汤科技董事长兼CEO徐立表示，公司将充分发挥在人工智能行业的优势，在深港合作、算法大模型、人工智能产业与投资等领域加强产业带动与应用示范。

据了解，算力中心基于SenseCore商汤大装置建设而成。该装置可降低人工智能应用成本，提高算力效率，为高校、科研机构、企业提供算力服务和合作平台。

新模型可实现零样本动物社交身份识别

科技日报讯（记者**罗云鹏** 通讯员**刁雯蕙**）1月12日，记者从中国科学院深圳先进技术研究院获悉，该院脑认知与脑疾病研究所研究员蔚鹏飞及其团队将AI技术运用到动物身份识别和神经科学研究中，提出一种研究社交行为的小样本学习计算框架模型。该模型可解决精确检测动物社交行为中的多个难点，有望创新社交行为神经环路机制的研究范式。相关研究成果发表在《自然·机器智能》上。

近年来，AI技术在传统行为学研究领域的应用日益广泛，DeepLabCut、SLEAP、MoSeq等AI动物行为追踪技术正成为神经科学家重要的研究工具。然而，上述技术在分析多个动物目标、动物自由社交行为时仍无法实现海量数据标注。此外，使用这些技术也会出现连续追踪的动物目标身份识别精度低等问题。

基于此，研究团队提出了双向迁移学习计算框架模型。使用这种模型，科研人员无需提前标注动物身份数据，即可实现多动物社交身份识别。据了解，这种识别的准确率超过90%，可完全满足动物社交实验的精度需求。

“双向迁移学习计算机框架模型的设计思路受大脑工作机制的启发。在非社交场景中，区分每一只动物的身份非常简单。这些模型已经认识的动物身份信息，可以迁移到多动物社交的场景。”蔚鹏飞说，此模型解决了AI需要人工标注大量数据才能实现多动物身份识别的问题，实现了零样本多动物社交身份识别。

“多动物行为量化是解读动物社交行为的关键，在神经科学和生态学中有着广泛的应用意义。”《自然·机器智能》期刊高级编辑特伦顿·杰德对该研究评价道。“未来，AI赋能的神经科学研究将为实施更加精准、个体化的无创神经调控提供指导，有望帮助人类进一步理解复杂精神疾病。”蔚鹏飞说。

扫码即能生成预问诊病历

AI医生助理提高看病效率

科技日报讯（记者**江耘** 实习生**卢馨怡**）记者1月12日获悉，浙江大学医学院附属邵逸夫医院（以下简称浙大邵逸夫医院）大运河院区近日启动试运行时引入了AI医生助理。患者可以通过扫描院区内展板上的二维码，使用微信小程序“邵医智慧门诊”，根据页面提示完成病情录入，生成预问诊病历，方便医生提前了解病史。

该小程序是一款基于大语言模型的AI医生助理。浙大邵逸夫医院门诊部主任丁勇告诉记者，AI医生助理可通过模拟临床医生诊疗思维对患者进行提问和引导，进行多轮一问一答的语音对话，快速完成患者主诉症状、伴随症状、诊疗情况、家族史等医疗信息的采集，生成一份完善的预问诊病例。

记者了解到，AI医生助理的无微不至，得益于大模型底座的先进计算能力、医疗大语言模型持续调优、医学语音引擎智能识别与转化、主动剔除无关干扰因素、AIGC格式化病历文本生成和智能影像与图片识别算法六大技术支撑。

医患间的沟通交流在诊疗过程中非常重要，无效沟通和信息的不对称不仅会造成时间的浪费，导致医生工作效率降低，还容易引发患者的不满情绪。现场前来问诊的患者吕先生说：“这次我通过AI医生助理将以前的诊疗记录全部上传，节约了诊治时间。”

据悉，浙大邵逸夫医院大运河院区试运行期间，将进一步扩大AI医生助理的应用范围，为探索AI赋能医学领域提供更多经验。

图为浙大邵逸夫医院挂号区内，一块印有AI医生助理小程序二维码的展板。 卢馨怡摄

上海推进IPv6技术演进和应用创新

新华社讯（记者**陈爱平**）上海市通信管理局、中共上海市委网络安全和信息化委员会办公室、上海市发展和改革委员会日前联合印发《上海市推进IPv6技术演进“智网上海”行动计划（2024-2025）》，全面推进IPv6技术演进和应用创新发展。

据介绍，IPv6是国际标准化组织IETF（互联网工程任务组）制定的下一代互联网协议版本，是全球公认的下一代互联网商业应用解决方案。“IPv6+”是面向5G和云时代的IP网络创新体系。

根据这份行动计划，到2025年末，上海将高标准全面建成“IPv6+”创新之城，主要目标和任务包括网络基础能力持续增强、技术创新取得显著突破、技术产业生态基本建构、重点行业应用成效凸显、安全保障能力显著提升等方面。

上海将充分释放“IPv6+”等创新技术潜能和优势，为5G、光网、算力等新型网络基础设施打造智能底座。到2025年末，上海将打造200个面向行业、园区、企业的IPv6专网；每百个重点场所拥有的网络切片接入站点数量超过70个；在企业组网和上云等场景中，新增用户开通企业专线50%以上采用“IPv6+”创新技术。

上海将推动“IPv6+”标准制定、技术研发和产业链协同体系初步形成，建设“IPv6+”公共服务平台、创新平台以及验证中心。

这份行动计划还提出，促进“IPv6+”与上海经济社会各行业全面深度融合，政务、广电、制造、金融、医疗、交通、教育、能源、互联网等重点行业“IPv6+”融合应用水平大幅提升。到2025年末，上海每个重点行业形成10个以上标杆应用，形成一批高质量示范项目；IPv6安全防护能力显著增强，基于IPv6的下一代互联网产业生态基本构建，为城市数字化转型筑牢“新基座”。

图说智能

机器人加油员上岗

近日，湖南省长沙市一加油站推出的加油机器人服务吸引了不少客人。在应用程序预约下单后，驾驶员只需要将车辆停至指定区域，加油机器人就会通过视觉识别技术定位车辆油箱盖，并自主完成开盖、加油、关盖等一系列操作，实现无人全自动加油。图为加油机器人给汽车加油。 中新社记者 杨华峰/视觉中国