深度求索大模型："花小钱办大事"

AI世界

◎本报记者 崔 爽

一个来自中国的开源模型，在开年之际聚焦了人工智能（AI）行业的目光。

日前，杭州深度求索人工智能基础技术研究有限公司（以下简称"深度求索"）上线并同步开源DeepSeek-V3模型，同时公布长达53页的技术报告，介绍关键技术和训练细节。

和很多语焉不详的报告相比，这份报告真正做到了开源。其中最抓人眼球的部分是，V3模型能力大幅升级，但训练仅仅花费557.6万美元，仅用2048块H800显卡，耗时不到两个月。

美国人工智能初创公司Anthropic首席执行官达里奥·阿莫迪曾透露，GPT-4o的模型训练成本约为1亿美元。美国开放人工智能研究中心（OpenAI）创始成员之一安德烈·卡帕西点评，DeepSeek-V3让在有限算力预算内进行模型预训练这件事变得容易。

深度求索如何实现"花小钱办大事"？它是否走出了大模型发展的一条新路？

降低模型推理成本

深度求索一直是国内AI版图上位置相对独特的一家——它是唯一没有做2C（面向个人消费者）应用的公司，选择开源路线，至今没有融过资。

去年5月，深度求索发布DeepSeek-V2，以其创新的模型架构和史无前例的性价比爆火。模型推理成本被降至每百万Tokens（大模型用来表示自然语言文本的单位）仅1元钱，约等于开源大模型Llama3 70B的七分之一，GPT-4 Turbo的七十分之一，引发字节、阿里、百度等企业的模型降价潮。

个中关节在于，DeepSeek提出的MLA（多头潜在注意力机制）架构和DeepSeekMoESparse（采用稀疏结构的混合专家模型）结构，大幅降低了模型的计算量和显存占用，实现了高效推理和经济高效的训练。

简单来说，模型压缩、专家并行训练、FP8混合精度训练、数据蒸馏与算法优化等一系列创新技术大幅降低了V3模型成本。作为新兴的低精度训练方法，FP8技术通过减少数据表示所需的位数，显著降低了内存占用和计算需求。据报道，目前，谷歌等已将这项技术引入模型训练与推理中。

深度科技研究院院长张孝荣在接受媒体采访时说，DeepSeek的"出圈"是对其在大模型技术上的突破和创新的认可，其通过优化算法和工程实践，实现高性能与低成本的平衡。DeepSeek为整个行业的发展注入活力，也对大模型的技术路径和工程实践产生积极影响，推动高效训练、模型轻量化和工程优化。

有业内人士分析，V3在架构创新、训练效率和推理性能方面展现巨大潜力，尤其在成本和性能的平衡方面作出重要贡献。不过，与此同时，也仍有许多挑战需要解决，如需进一步扩展上下文长度、优化多模态数据处理等。未来的研究方向包括提升模型的推理速度、完善更高效的硬件架构设计，以及增强多模态学习和生成能力。

不堆算力创新算法

大参数、大算力、大投入，这条已经被验证行之有效的ChatGPT路径，实则是绝大部分创业公司难以承受之重。

据报道，仍处于研发过程中的GPT-5，已进行过至少两轮训练，每轮训练耗时数月，一轮计算成本接近5亿美元。一年半过去，GPT-5仍未问世。这意味着，新一代通用大模型的训练成本已达到十多亿美元甚至更高。未来这一数字可能持续攀升。

规模定律（Scaling law）是指在训练大模型时，数据量、参数量和计算资源越多，训练出的模型能力和效果越好。然而，一段时间以来，行业对规模定律可持续性的疑问不绝于耳。

V3的出现提供了新的解法。"Scaling Law不只停留在预训练阶段，而是往后训练，尤其是注重推理领域的后训练集、强化学习等领域扩展。"智源研究院副院长兼总工程师林咏华接受科技日报记者采访时说，这一点在国外以OpenAI o1发布为标志，国内则有DeepSeek使用强化学习训练发布DeepSeek R1这个具有很强挖掘和激活能力的模型。

视觉中国供图

在林咏华看来，V3的发布，也印证了利用R1可以很好进行能力提升。

行业相关探索还有很多，如Kimi将强化学习用到更多搜索场景，发布以逻辑思考和深度思考为核心功能的数学模型K0-math；蚂蚁技术研究院建立强化学习实验室，围绕如何在后训练及强化学习上进行更多模型能力的探索。林咏华期待，未来不仅是靠堆砌更多算力、参数和数据，而是靠真正的算法创新，持续在后训练阶段帮助模型提升基础能力。

值得注意的是，"省钱模式开启"并不意味着算力式微。

V3发布后，360集团创始人周鸿祎发文称赞"DeepSeek的进步对推动中国AI产业发展是极大利好"，但他也认为，这并非说中国AI发展不需要高端算力芯片。囤显卡建算力集群依旧必要，因为目前预训练算力需求或许没那么大，但像慢思考这类复杂推理模型对推理算力需求大，文生图、文生视频的应用也需消耗大量算力资源。巨头们提供AI云服务，构建庞大算力基础必不可少，这与DeepSeek 降低训练算力需求是两回事，两者并不矛盾。

一位行业专家在接受记者采访时认为，2025年，大模型行业会进一步收敛，这种收敛既包括技术层面，也包括厂商层面。进入"百模大战"后期，要进一步提高模型计算效率，降低推理成本，对计算的架构分布、利用效率等都提出更为精细化的要求。

"烧钱"不是唯一逻辑

深度求索创始人梁文锋在金融行业征战已久。他成立的幻方量化早在2019年就开始大手笔投入深度学习训练平台。2023年7月，梁文锋创立深度求索，专注AI大模型的研究和开发。

据报道，包括梁文锋在内，深度求索仅有139名工程师和研究人员。在外界看来，这是一支"神秘的东方力量"。

但在一次采访中，梁文锋曾透露，深度求索并没有什么高深莫测的奇才，团队都是国内顶尖高校的应届毕业生，没毕业的博四、博五实习生，还有一些毕业才几年的年轻人。他特别提及，"V2模型没有海外回来的人，都是本土的"。

他也曾在访谈中说，过去30多年的IT浪潮，中国基本上扮演的是追随者角色，"随着经济的发展，中国也应该逐步成为技术创新的主要贡献者"。如今，V3的横空出世贡献了一个更高效率、更低成本的大模型发展样本，也让AI行业看到一种可能：虽然训练大模型依然需要大规模显卡集群，但"烧钱"不是行业唯一的逻辑，也并不是谁烧钱多，谁就注定赢得一切。

对此，周鸿祎评论道，V3用2000块卡做到了万卡集群才能做到的事。用这种极致训练方法训练专业大模型，算力成本会进一步降低，促使中国AI在专业、垂直、场景、行业大模型上更快普及。

5G智能系统让塔吊司机告别高空作业

◎本报记者 赵向南

近日，在山西省太原市阳曲县人民医院新院区建设现场，两台塔吊同步进行作业。与其他塔吊作业现场不同的是，这里的塔吊司机并未置身于百米高的驾驶舱，而是坐在地面操作室里，通过屏幕和手柄远程操作。科技日报记者在现场看到，吊装作业过程平稳且高效。

山西建设投资集团有限公司山西六建集团公司（以下简称"山西建投六建集团"）技术研发中心经理宋红旗介绍，塔吊配备了5G智能塔式起重机远程控制系统，将塔吊高空操作转变为地面远程吊装作业，施工安全稳定高效。"这套远程控制系统是国内首创，率先在山西建筑业项目进行了全过程应用。"宋红旗说。

该控制系统由山西建投六建集团主导，联合太原科技大学、广联达科技股份有限公司研发。系统由12个子系统构成，融合了5G、人工智能识别、智能控制等先进技术。配备该控制系统的塔吊，具备远端控制、自动辅助驾驶、群塔防碰撞、智能吊钩、集中控制等功能，实现了数字化、智能化和高效化作业。

"在自动辅助驾驶系统的帮助下，塔吊可自动规划用时最短、最安全的吊装路径，并提供塔机状态信息数据，实现精准吊装。"宋红旗介绍，群塔防碰撞系统可通过智能识别塔机健康状况和工作环境，实时监控塔机碰撞信息，提升主动安全性；智能吊钩系统可利用高清视频和人工智能边缘计算技术，实现吊钩可视化精准定位和风险点智能识别预警。

由于该控制系统集成于集中控制平台，操作人员在地面便可远程操控塔吊，且一位操作人员能分时操控多台塔机。宋红旗说，该系统可将塔机利用率提升15%以上，还能预防70%至80%的人为安全事故，降低高空作业风险。

近年来，随着5G、人工智能等技术的应用与推广，建筑行业正在进行转型升级，为城市建设和社会进步贡献更多智慧和力量。山西建设投资集团有限公司董事长倪华光表示，作为山西规模最大的综合性国有投资建设集团，公司将继续加大在智能建造领域的投入和研发力度，推动更多智能化、数字化技术的创新与应用。

5000米高原露天矿用上无人驾驶技术

科技日报讯（记者**崔爽**）1月7日，记者从华为获悉，由该公司与西部矿业集团（以下简称"西部矿业"）、中铁十九局集团有限公司三方共同建设的国内首个5000米高原露天矿无人驾驶项目成果日前在西藏昌都玉龙铜矿发布。

据介绍，玉龙铜矿无人驾驶技术的核心在于"云网车"协同。基于华为云的人工智能算力底座，无人驾驶训练及迭代、智能运营监管、高精地图服务和路径规划功能得以实现。同时，项目组联合西藏移动架设多座5G基站，实现5G信号全覆盖，使车载传感器能实时上传信号，让地图数据以分钟级更新，从而优化效率并保障作业安全。

针对高寒高海拔的特殊环境，矿车采用多传感器融合感知技术，确保全天候稳定作业，在线运行率始终保持在99%以上，实时感知系统也为生产作业构建了立体化安全防护体系。

值得一提的是，无人驾驶技术的智能化调度系统优化了作业流程，减少了道路维护次数和车辆维修频率，降低了燃料和轮胎的消耗。相比传统模式，两个编组10辆无人驾驶矿卡每年可节省成本约600万元。

无人驾驶技术是推动传统矿山行业转型升级的关键一步。西部矿业党委书记、董事长张永利表示，项目的交付将全面提升高原露天矿山运营效率、作业安全性，公司未来将以玉龙铜矿为示范基地，进一步拓展无人驾驶技术的应用场景，打造更加安全、高效、绿色的现代化智能矿山。

上海聚力构建世界级人工智能产业生态

◎本报记者 王 春

从"魔都"到"模都"，上海正以开放的姿态拥抱全球合作伙伴，赋能大模型创新企业集群。上海计划通过一系列措施，到2025年底建成世界级人工智能产业生态。

2024年底，上海发布《关于人工智能"模塑申城"的实施方案》，提出打造超大规模自主智算集群、构建多层次语料供给体系等基础设施的目标。在前不久举行的上海市"人工智能+"行动推进大会暨中国—金砖国家人工智能发展与合作中心运营基地启用仪式上，"模塑申城"五大公共服务平台发布。

这五大公共服务平台包括上海市智能算力公共服务平台运营系统、模塑申城语料普惠计划、大模型评测与验证平台、"百人百项"青年科学家计划和上海国投—徐汇融资服务中心。五大平台将着力降低"人工智能+"应用落地成本，为各类创新主体提供优质、便捷、普惠的公共服务。

此次大会上，上海市首批"模塑申城"行业应用示范基地发布。这将推动"人工智能+"金融、制造、教育、医疗、文旅、城市治理等重点行业的应用落地，带动上下游协同创新，共促产业生态发展。

同时，中国—金砖国家人工智能发展与合作中心运营基地在上海徐汇区正式启用。该基地旨在广泛链接金砖国家乃至全球人工智能创新和产业资源，推动全球人工智能生态系统互联互通，为我国面向其他金砖国家乃至全球开展人工智能国际合作提供平台。

作为上海首个人工智能发展集聚区，徐汇区跑出人工智能产业发展"加速度"。

2023年，全国首个大模型创新生态社区"模速空间"在徐汇区成立。抓住通用人工智能的变革趋势和产业风口，徐汇区落地上海创智学院等一批国家战略平台，并引进中国电信人工智能研究院等一批大模型企业机构，成功创建上海市生成式人工智能创新生态先导区。"模速空间"揭牌一周年以来，已吸引近百家大模型初创企业入驻，带动该区200余家大模型企业加速集聚。目前，徐汇区已有34个大模型通过国家网信办备案，占上海市的近60%。

作为"模速空间"的入驻企业，上海羚一人工智能科技有限公司将"AI+工业"确定为企业发展方向。该公司总经理陈启明介绍："与我们合作的许多工业企业在金砖国家有港口基建、装备制造等项目。通过我们提供的从数据到应用的一体化智能体解决方案，可满足制造企业在设计、采购、制造、管理等领域的AI+业务协同需求。"陈启明期待，中国人工智能技术能走出国门，为更多国家的企业提供服务。

徐汇区还在此次大会上发布了推进"人工智能+"十大行动，涵盖模型能力跃升、创新策源突破、场景应用牵引、孵育范式升级、产业集群倍增等方面。十大行动旨在推动徐汇区加快建成全国人工智能高地。其中，"创新策源突破行动"提出，加强新型研发机构间深度合作，推动科学智能创新研究，加快具身智能和类脑科学等前沿技术攻关。

青海首个电缆隧道智能巡检机器人"上岗"

科技日报讯（记者**张蕴** 通讯员**丁有鹏**）1月9日，记者从国网西宁供电公司获悉，由该公司研发的智能巡检机器人日前在沿110千伏杨乐线隧道架空轨道完成自主巡视任务，标志着青海首个电缆隧道智能巡检机器人正式"上岗"。

110千伏杨乐线电缆隧道位于地下4米深处，全长650米、宽2米。其供电范围覆盖整个西宁市城东区，关系着千家万户的用电。为实时掌握高压电缆设备运行状态，提升巡视工作效率，2024年1月，国网西宁供电公司自主实施电缆隧道智能巡检机器人研制项目，成功研制出青海省首个电缆隧道智能巡检机器人。

这款智能巡检机器人由车体、驱动电机、控制箱、红外激光雷达、360° 全角度摄像头、温度和湿度探头等部件构成，可依托通道顶部工字形轨道自动行驶，具有视频图像识别、红外热成像测温、环境及气体检测、音频采集分析、双向语音对讲等功能。运维人员可利用智能巡检机器人实时监控电缆隧道内设备的运行状态，在发现缺陷隐患时尽早处理，确保电网安全稳定运行。

"西宁市电缆总长149.48公里，人工巡检需两个半月才能完成，而智能巡检机器人只需一周。它不仅使工作效率大大提升，也让检测数据更精准。"国网西宁供电公司电缆运检班班长张明帅说。

据悉，下一步，国网西宁供电公司将深化智能设备应用，通过"智能巡检+人工巡检"，构建"立体巡检+集中监控"运维模式，确保人民群众温暖过冬、亮堂过年。

图为电缆隧道智能巡检机器人。 **沈仪**摄

兼顾安全保护和流通效率

中国电子云可信数据空间解决方案发布

科技日报讯（记者**操秀英**）1月9日，记者从中电云计算技术有限公司（以下简称"中国电子云"）获悉，该公司日前发布云数一体可信数据空间解决方案，并携手国家对地观测科学数据中心无人机遥感数据资源分中心、中盾安信、中科加能、朗新科技、百度智能云等行业合作伙伴，共同发布高质量共建可信数据空间联合倡议。

可信数据空间是基于共识规则，联接多方主体，实现数据资源共享共用的一种新型数据流通利用基础设施，是数据要素价值共创的应用生态，是支撑构建全国一体化数据市场的重要载体。近期，《关于促进数据产业高质量发展的指导意见》《关于促进企业数据资源开发利用的意见》《可信数据空间发展行动计划（2024—2028年）》等文件印发，积极推动可信数据空间建设迈出新步伐。

中国电子云数据运营部总经理宋超认为，可信数据空间建设主要面临三个难点。第一，可信数据空间需要汇聚多路数据资源，承接各类主体，这些数据资源的类型以及各主体提供的数据方案模式都不同，如何进行有效整合是个技术难点。第二，对数据的统一管控面临难题。第三，可信数据空间是国家数据基础设施建设的一个重要内容，与其他数据基础设施的连接，有待进一步研究。

"中国电子云采用云数深度融合的方案，基于自主创新的分布式云原生操作系统和数据流通控制技术，提供一种广泛的数据主体安全可信接入能力，帮助客户统一开展数据资源管理、数据开发运营和数据资产保值增值等多类业务。"宋超说，方案为解决可信数据空间建设的难点问题提供了一个可行路径。

宋超进一步介绍，该方案构建"云边协同、多元共创、动态管控、全栈安全"的分布式数据价值可信流转网络，实现多源数据融合价值共创。"方案能够完整解决一个城市、一个行业、一个企业的数据价值流通链条中的全栈流通体系建构问题。"宋超说。

中国电子云副总裁冯进认为，该解决方案的独特之处在于提供了一个标准化产品，兼顾安全保护和流通效率，可用于大模型训练推理等创新场景，让可信空间能切实发挥国家数据基础设施的作用。

此次发布的高质量共建可信数据空间联合倡议提出，各方应积极响应国家数据要素政策，推动数据要素高效流通及价值转化，有效解决数据"不敢享""不愿供""不流通""难利用"等关键问题，携手落地可信数据空间，推进城市、行业、企业数据基础设施落地，畅通数据流通大动脉。

据介绍，中国电子云可信数据空间已面向城市、行业和企业等领域进行探索实践，目前已在部分城市、央企，以及高端装备制造、育种等行业落地。未来，中国电子云将携手合作伙伴，共同推动可信数据空间应用生态发展，实现多源数据融合价值共创，赋能产业数字化转型和业务创新，共同助力建设全国一体化数据市场。

据悉，中国电子云作为数据创新领域先行者，主动开展数据基础设施建设探索和实践，参与到北京、深圳、大连等地的数据基础设施规划设计中。在数据"治得专"上，中国电子云支撑了地方政府、央国企、关键行业等150余个数据平台建设，为数据运营打下良好基础。在数据"供得出"上，中国电子云数据港对接全国17个城市的50多类高价值社会数据资源，自营12个数据产品，实现数据运营闭环。在数据"流得动"上，中国电子云支撑深圳数据交易所全国性数据交易服务平台的建设与迭代升级。