并非越大越好，模型选择要结合需求

AI世界

◎本报记者 都芃

8月底，微软和英伟达相继发布小型语言模型，引发业界热议。两家公司均称，新发布的小模型能在算力资源消耗和功能表现之间取得平衡，甚至可以在某些方面媲美大模型。不仅如此，苹果、美国开放人工智能研究中心（OpenAI）等也发布了参数规模更小、性能更强的小型语言模型。

小模型通常指参数少、占用算力资源少、反应速度快、可以本地化运行的小型语言模型。在大模型竞争日趋激烈的今天，多家人工智能企业及研发机构为何另辟蹊径，加码小模型赛道？

大模型训练成本高

大模型赛道太“卷”了——这是部分业界人士对当下大模型产业发展的评价。随着各家人工智能厂商在大模型领域不断加大投入，如今百亿级甚至千亿级参数的大模型已不再稀缺，大模型产品同质化趋势也愈发明显。

但模型真的越大越好吗？模型越大，意味着消耗的资源越多，成本越高。今年4月，OpenAI首席执行官萨姆·奥尔特曼在麻省理工学院演讲时提到，“我认为我们正处于巨型模型时代的结尾”。在他看来，未来人工智能新的进步并非来自于越来越大的模型。

且不论运行所需费用，仅在训练阶段，大模型就要花费巨额成本。OpenAI前研究副总裁、人工智能初创公司Anthropic首席执行官达里奥·阿莫迪曾提到，目前像GPT-4o这样的模型训练成本约为1亿美元，而当下正在开发的人工智能大模型训练成本则可能高达10亿美元。他预计，未来3年内，人工智能大模型的训练成本将上升至100亿美元甚至1000亿美元。

对于国内大模型产品而言，成本同样居高不下。百川智能创始人兼首席执行官王小川曾提到，大模型每1亿参数对应训练成本在1.5万到3万元人民币之间。一个千亿级参数的大模型，单次训练成本在3000万至5000万元人民币之间。

高端算力短缺等因素也是困扰国内大模型发展的难题。2023年，科大讯飞与华为联合发布首个全国产算力平台“飞星一号”，以此为基础训练出的讯飞星火大模型已实现自主可控。但整体来看，相比国际先进水平，国内大模型产品仍有较大提升空间。

此外，在应用端，端侧部署是目前人工智能大模型发展的热门方向，但由于所需算力资源过于庞大，大模型几乎无法在手机、人形机器人等小型终端上本地部署，限制了大模型的应用场景。例如，在目前发布且有实际演示的模型中，10亿参数量模型尚可在手机上运行，一旦模型参数级别跃升至百亿级，在手机端运行就变得非常吃力，几乎无法正常使用。在许多场景下，模型规模越大并不一定能给用户带来更好的使用体验，这也给小模型留下了更多发展空间。

小模型有多重优势

大多数小模型参数量在几百万至数千万，结构也更简单。参数量缩小带来的明显改变是对功耗以及算力需求的降低。

目前主流旗舰手机的芯片算力可以达到40—50TOPS（1TOPS代表处理器每秒钟可进行1万亿次操作）。若再叠加专门开发的功耗控制策略，许多智能手机都能轻松“驾驭”小模型。

模型虽小，但在部分专门领域，其功能表现并不输大模型。例如OpenAI推出的轻量化模型GPT-4o mini在常见的多轮对话等功能上，与GPT-4o表现不相上下。

针对当下大模型存在的“幻觉”问题，即机器可能输出“无中生有”的内容，小模型通过专注于学习某个细分领域的精华数据，可降低不相关、意外或不一致的输出风险，显著降低“幻觉”现象出现概率。

此外，相比部署在云上的大模型，小模型具有个性化程度高、响应速度快等特点，这使其更贴近用户端需求。同时，小模型的本地化部署也能更大程度保障用户的数据控制权和隐私权。

大小模型协同发展

当然，对于实现通用人工智能这一终极目标而言，小模型远远不够。小模型在当下的快速发展，更多是企业出自平衡成本与用户需求后的理性选择。

科大讯飞副总裁、研究院院长刘聪认为，不能泛泛谈大模型与小模型孰优孰劣，必须结合具体使用场景来评估。他举例说，如果只是让人工智能写一个具体行业的文案，或是只对具体行业的文字进行翻译、润色等处理，一个中小规模的模型就完全够用。但如果是在开放信息环境中，对不特定的内容进行提取、识别、分析等操作，大模型的表现毫无疑问将更好。

在刘聪看来，大、小模型相结合或将成为未来人工智能发展的重要方向，确定某一任务是使用大模型还是小模型更好，取决于其泛化性和效率要求。“归根结底要围绕具体需求展开，这两者不是非此即彼的关系。”他说。

具体在研发层面，大、小模型之间的关系更多是协作而非竞争。当下，许多科技巨头的做法是先训练出通用能力足够强的大模型，再借助大模型对数据进行初步筛选处理。站在大模型“肩膀”上的小模型，可以用质量更高、数量更少的数据完成训练，以更低成本实现不输大模型的效果。“大模型的目标是找到性能的天花板。以此为基础再优化小模型，和从零起步做一个小模型相比，效果完全不同。”刘聪说。

我国首个区块链专用计算硬件开放架构发布

科技日报讯（记者**崔爽**）记者8月30日从国家区块链技术创新中心获悉，我国首个区块链专用计算硬件开放架构BUDA（Blockchain Unified Device Architecture）日前正式发布，并被命名为“菩提”。该架构为区块链与隐私计算的底层软件提供了统一的专用硬件功能、实现规范和调用接口，可大幅提升区块链网络中数据要素安全可信流通效率，为扩大区块链应用生态、全面加速国家级区块链网络建设、实现我国数据要素互联互通提供更高效能。

近年来，区块链在数字经济场景中的应用愈发广泛。2021年1月，国家区块链技术创新中心团队推出我国首个自主可控的区块链软硬件技术体系——长安链。凭借全自主、高性能、强隐私保护等优势，长安链打破国外区块链底层技术垄断，为区块链大规模组网应用奠定基础。中国信通院数据显示，长安链已连续两年位居国内区块链底链市场占有率第一。

“我们一方面持续优化硬件技术，另一方面也一直在努力打造更加普适、易用的硬件开放架构，让整个国内区块链生态都能享受专用计算硬件加速带来的性能红利。”长安链硬件研发中心负责人说。

据介绍，BUDA“菩提”包括系统架构、功能实现规范、接口规范等。国内任意厂家均可参考开放的系统架构和功能实现规范来设计区块链与隐私计算专用硬件，并可参考开放的接口规范，让不同区块链软件平台调用相关功能，实现区块链与隐私计算整体系统性能提升。该架构还可以支撑不同区块链之间的连接与协作，助力建成链间“朋友圈”，降低不同应用链上主体数据交互难度，促进区块链与隐私计算专用硬件功能兼容，实现可互换性、互操作性和一致性。

国家区块链技术创新中心负责人表示，扩大区块链软硬件技术开源开放，有助于解决区块链专用硬件设计能力发展缓慢、部分区块链应用场景受性能限制、不同区块链底层平台间互联互通效率低等问题。同时，这也将进一步激发行业创新活力，让更多行业伙伴具备区块链与隐私计算专用硬件生产能力，从而共同推动区块链技术普及应用，加速国家级区块链网络建设，促进区块链网络内数据要素更加高效地流通共享。

大模型“域见医言”上线

医生诊断病情有了专业AI助手

◎洪恒飞 林捷 本报记者 江耘

分析报告、诊断病情……这一切将变得更加高效，因为医生如今有了专业人工智能助手。日前，在浙江杭州举行的中华医学会第十八次检验医学学术会议上，医检大模型——“域见医言”正式发布。基于该大模型的智能应用“小域医”同步上线，它具备智慧报告解读、疾病诊疗助手等功能，可大幅缩短医生决策的时间，减轻医生负担。

医学检测是医学诊疗过程中的关键一环。研究显示，70%的临床决策信息来自医学实验室报告。然而，检测技术增多、操作流程烦琐、临床场景复杂多变等因素，大幅增加了临床医生、检验医师分析报告的工作量。此外，由于培养周期长、难度大，富有临床经验的复合型检验医师人才紧缺。

为破解这些难题，广州金域医学检验集团股份有限公司耗时近两年开发训练出“域见医言”。该公司在通用语料基础上注入超20亿Token（数据单元）医检语料，目前已有超2万名企业专业技术人员、临床专家、检验医师参与“域见医言”测试。

记者体验后发现，类似使用ChatGPT、通义千问等大模型应用，使用者可通过对话框与“小域医”互动，问询诊疗相关内容。

以智慧报告解读功能为例，对涉及生化、免疫、质谱、基因组等多学科报告的复杂病情，使用者可上传报告图片或PDF，由应用进行解读。单份报告通常10秒内即可完成解读。疾病诊疗助手功能可提供全面的临床诊疗辅助，包括疾病查询、病历整理、鉴别诊断、医嘱辅助和体检评估等服务，帮助患者接受更及时、精准的治疗。

广州金域医学检验集团股份有限公司副总裁兼数字化管理中心总经理李映华介绍，和其他行业大模型不同，“域见医言”不依赖特定大模型底座，能适应各类通用多模态大模型。它支持与形态学、病理、基因等专业领域的大模型以及医检特定场景的专用模型、工具进行衔接。通过技术融合，它可以有效提升在整体判读与识别方面的精准度。

浙江省肿瘤医院检验科主任张毅敏认为，精准医疗时代下，医学检验人要有临床思维，也要有数据思维。医检大模型可以将临床、检验和数据更紧密地结合起来，有利于医生出具更精准的检查报告，更好服务患者。

“未来机器人什么样？年轻人说了算！”

——第七届中国高校智能机器人创意大赛侧记

◎洪恒飞 胡格格 本报记者 江耘

近日，第七届中国高校智能机器人创意大赛在浙江省余姚市如约而至。智能茶道机器人、穴位按摩机器人、家庭备灾应急机器人……来自27个省区市382所高校的1366支参赛队、4192名参赛选手携上千款机器人，在中塑国际会展中心同台竞技。

“请看我们的衣物晾序智控护理搭配一体机。它具有检测衣物掉色程度、建立App衣库、晾衣收衣、搭配建议、雾化香薰等功能。”温州理工学院产品设计专业大四学生郑胜丹向现场观众介绍一款她和同学发明的“智悉衣管家”机器人。

第七届中国高校智能机器人创意大赛现场。 大赛组委会供图

大赛现场，600多个家用智能机器人展位上，一批着眼未来生活的服务型机器人尽显巧思，涉及家务劳动、情感交流、个人卫生、家庭管理、安全防护等领域。

“上啊！”“稳住！”“胜利了！”杭州电子科技大学信息工程学院选手陈星培手捧自家“宝贝”轮式格斗机器人，在“战场”上击败对手，满面笑容走下赛场。“这就是我们参加大赛的意义所在，以赛代练，不断进步，下次还来。”陈星培说，前期参加大赛的省赛后，团队第一时间对机器人进行了改造，使其战斗实力大增。

大赛现场，一轮胜过一轮的比拼，一浪高过一浪的喝彩，将氛围一次次推至顶点。火热的大赛现场，折射出余姚机器人产业的蓬勃发展。

《“十四五”机器人产业发展规划》提出，到2025年，我国成为全球机器人技术创新策源地、高端制造集聚地和集成应用新高地；到2035年，我国机器人产业综合实力达到国际领先水平，机器人成为经济发展、人民生活、社会治理的重要组成。

余姚机器人产业发展起步较早，2013年已开启机器换人及智能化改造进程。近年来，余姚以机器人智谷小镇为核心，稳步构建机器人产业生态圈。产业强则人才聚，人才聚则产业兴。余姚不断加大对青年人才的招引力度。

“未来机器人什么样？年轻人说了算！”浙江大学机器人研究院常务副院长陆国栋说，作为大赛承办单位，希望通过大赛培养学生提出问题、解决问题、落地创作、后期孵化的能力。大赛从无到有、从有到兴，在全国高校受欢迎程度持续高涨。他希望通过这一窗口活动的举办，让参赛学生与余姚结缘，甚至未来到余姚就业创业，为推动余姚机器人技术发展、智能制造业转型升级提供青创力量。

专家支招“车路云一体化”发展

◎本报记者 刘垠

截至2023年底，我国共建设17个国家级测试示范区、7个车联网先导区、16个智慧城市与智能网联汽车协同发展试点城市，发放测试示范牌照超5200张，累计道路测试总里程8800万公里……日前，在第四届沈阳智能网联汽车大会上，工业和信息化部原党组副书记、副部长苏波一一列举我国智能网联汽车产业发展成绩。

“以‘车路云一体化’为主要方向的智能网联汽车，承担着培育新质生产力、加速塑造汽车产业高质量发展新动能、新优势的重大使命。”苏波说，“车路云一体化”的发展，标志着我国车联网产业已从技术验证过渡至规模落地的重要阶段，进入拓展应用的新进程。

2024年，我国启动智能网联汽车“车路云一体化”应用试点工作。中国工程院院士李克强说，“车路云一体化”应用试点取得显著进展，但在推进过程中仍面临一些挑战。比如，许多城市在投入资源后，未能形成有效商业闭环。当前，“车路云一体化”应用示范还处在初级阶段，在车辆运用、系统运行方面仍以单车智能为主，这导致获取数据很被动。

数据是“车路云一体化”可持续发展的重要驱动因素。中国电动汽车百人会副理事长兼秘书长张永伟认为，让数据得到分级、分类挖掘和运营，是“车路云一体化”实现商业回报的关键。政策和标准是其落地和应用的重要保障，既需要国家部委制定相关政策，更需要地方推进政策、标准和法规创新。

如何更好推动“车路云一体化”发展？苏波建议，要加强技术创新、完善产业生态、强化基础设施建设、拓展市场应用等；应持续投入研发资源，突破关键核心技术，推动自动驾驶及智驾芯片、车联网、云计算等领域的技术创新和突破；要加强汽车、通信、交通等产业协同合作，推动产业链、创新链、资金链和政策链深度融合，为“车路云一体化”发展提供支撑。

李克强认为，推动“车路云一体化”，要确保项目形成商业闭环，使基础设施建设形成持续的投入产出机制。“车路云一体化”还需进行系统设计，明确架构、方法论和流程规范，制定与国际接轨的标准规范，推动技术体系全球推广。

链接

政策红利加速释放 产业发展持续升温

随着越来越多利好政策出台，“车路云一体化”建设不断升温。

今年初，工业和信息化部等五部门印发《关于开展智能网联汽车“车路云一体化”应用试点工作的通知》，加快推动智能网联汽车技术突破和产业化发展。7月，工业和信息化部等五部门发布《关于公布智能网联汽车“车路云一体化”应用试点城市名单的通知》，确定北京、上海、深圳、广州、武汉、重庆等20个城市（联合体）作为首批“车路云一体化”应用试点城市，以加快形成全国可复制可推广的经验。

为促进“聪明的车”“智慧的路”“强大的云”融合协同发展，各地在“车路云一体化”赛道上加速奔跑。

5月，北京发布“车路云一体化”新型基础设施建设项目招标公告，项目投资额达99亿元，计划在全市选取约2324平方公里范围，对6050个道路路口进行智慧化改造；6月，武汉的“车路云一体化”重大示范项目获得武汉市发展改革委批准，备案金额达170亿元；当前，重庆正加紧推动试点工作，计划建设智能路口3000个、路侧单元4000个，覆盖范围5000平方公里、道路里程5000公里……

业内专家认为，“车路云一体化”大规模建设的序幕已经拉开，行业将迎来快速发展期。

图说智能

医生操作 机器“操刀”

随着技术的发展，远程手术从概念走进现实。手术机器人的出现，让外科医生得以远离手术台，操作机器进行手术。图为观众在2024世界机器人大会上体验一款手术机器人。 郭海鹏

本版图片除标注外由视觉中国提供