

“AI革命”的进度条,走到哪儿了?

□半月谈记者 张漫子

ChatGPT 刚诞生的2022年底,就算我们把AI当回事,AI却没我们什么事——大厂之间的角力、大国之间的博弈,离大众还很远。但时间来到2024年,日常生活已被AI占据:AI电脑、AI手机、AI软件、AI游戏、AI陪伴,我们真真实实触摸到了AI之躯,它的应用比它的名字还要普及。

一个更为强烈的感受是,AI变了。它的“智”和“能”不断进化、迭代,不断引我们到未曾想象的新世界。回望人工智能风起云涌的一年,“AI革命”的进度条,走到哪儿了?

瘦身

大模型,“瘦身”了。今年的大模型“武林”依然精彩,只是简单粗暴的追参数、比大小已成明日黄花。“小了,但要强”,成为国产大模型厂商的追求。

有研究者测算,如果大模型保持现在的发展势头,到2028年左右,现有的数据储量就将全部用完。还有3年,就要撞上南墙了吗?大模型向何处去?也许,我们需要新的参照系。

面壁智能首席科学家、清华大学副教授刘知远拿出一把新标尺:能力密度。“大模型并不是参数越大越好。同样的模型能力可以放入一个更小的参数

规模,这表明模型的能力密度不断增强。达到同等性能模型所需参数数量越小,意味着模型性价比越高,效率越高,能力越强。”刘知远说。

新的认知,发散出新的思路:炼大模型,不如炼优模型。不是大模型不给力,而是小模型更具性价比。这一年,越来越多的模型厂商将目光聚焦在小模型身上,更注重AI算法的调优。一边降参数、一边提性能,以最小的资源释放最大的能量,大模型、小模型不再因训练数据量的差异而出现性能代差。

AI处理的重心也不再依赖云端算力,开始向端侧转移。这催生了设备单元大模型的新需求。借助一系列优化措施,单元大模型在特定任务上的表现,甚至优于预训练大模型。它可以广泛适配不同类型终端,并通过合理连接,激发群体智能。

“我们在手机和PC端体验到的端侧AI,大部分是通过适配器、分类器挑选的微调小模型,以及经过自定义开发的智能体,操作中无需调用全部模型参数,无疑更具性价比、更具效率。”面壁智能联合创始人、CEO李大海说。

破壁

第二个强烈感受,大模型变“勤快”了。除了越来越瘦、越来越轻盈之外,它越来越实干,不再是那个只说不干的“口嗨派”。

2024年11月底,智谱AI开放日上最新亮相的AI智能体已经化“chat”为“act”,身体力行替人类点外卖。更惊喜的是,这款国产AI智能体已初步实现跨App、多步骤解决真实任务。

安装AutoGLM、点击“绿色手机”图标,对着智能体说出“帮我查询一顿火锅需要的食材,然后下单送到家里”的指令,AI就会自动打开小红书App页面、搜索火锅食谱,找到需要的食材,然后关闭小红书App,打开美团App进入

超市频道,勾选食材下单——在无需人工干预的条件下,一口气完成54个步骤。

尽管目前刚刚满月的它并不比人类的操作要敏捷多少,但这至少意味着AI终于能跳出“对话框”,帮人类做点什么事了。“今年,AI实现从技术驱动到需求驱动的转变。从业者深入思考如何让大模型进入真实的物理世界,真正行动起来。”智谱CEO张鹏说。

规模效应时代的AI,值得关注的就3件事:算力、算法、数据。而今,在“致用”大方向上寻找价值的AI亟须打通3件事:感知、思维、行为。

物理空间与认知空间的交互,激发

了人类的想象力与创造力,是抽象思维的起点,对于进化中的AI而言,也就分外重要。

当然,现在的AI“感知”到的可能仅仅是事物不完整的、甚至碎片化的特征,如何在这样的基础上让AI“灵光乍现”,有所领悟,甚至产生思维的萌芽?正如2024世界人工智能大会上诸多专家所言,“AI智能体才是未来的真正挑战”。由“知”到“行”,跨越非易。

“智能体改变的是人与机器的互动方式,基于理解需求、规划与决策、执行行动和自我反思,它将带来符合直觉的人机交互——从人适应机器,到让机器适应人。”张鹏说。

生长

进入“青春期”的大模型,还在蓬勃“发育”之中。它将长成什么样子?我们不妨回到人工智能的本质:模仿人类智能,完成复杂任务。这意味着它不仅仅是一名吸收人类已有知识库的学生、一个可以与人交谈的智能设备,更不是被动回答人类问题的反馈机器。人工智能的未来,应该是“有灵且美”的类生命体。AI的感知能力、交互能力,必将不断丰富和拓展。

拓展,要从哪几方面展开?

一是行动。过去24个月,大模型主要增强的是人工智能“智”的维度,也就

是作为一位参谋,读懂问题、搜索答案、形成回复。对于人工智能“能”的维度,也就是在拆解目标、做出决策、执行任务的层面,它的增益并不显著,远没有赶上“智”的进化。

二是自主。也就是变被动响应为主动思考,根据感知到的具体情境做出灵活的有针对性的决定和行动,主动规划、主动推理,主动在现实世界采取行动。

三是预测。收到“把水杯放在桌上”的指令后,为什么人类能把装着水的水杯正面向上,平放在桌面而不是有起伏的书堆之上,是因为人类能够预测当水

杯放在“斜坡”上会失去平衡而滑落到地上。机器“拟”人,就需要掌握预测能力。这就要求AI不仅可以理解物理世界的规律,还能对这些规律产生记忆,在合适的时候调出这些记忆。

可以预见,当主动行动、快速预测的AI智能体“炼”成后,我们的手机、电脑、软件、机器人,就可以自己“跑”了。而如何能让人工智能的“能”追上“智”的发育,如何能获得行动、自主、预测这三种能力,尚无现成方案可依,不仅需要新的范式,也需要开拓新的路径。

来源:新华社

在美国消费电子展上看AI智能体发展风向

□新华社记者 谭晶晶 黄恒 高山



这是美国拉斯维加斯消费电子展Old World Labs展区的机器人展示(1月8日摄)。

拉斯维加斯消费电子展(CES)7日至10日举行。人工智能(AI)技术驱动的新应用、新解决方案是本届展会最大热点,AI智能体成为参展企业热议词汇。

□新华社发(曾慧摄)

美国拉斯维加斯消费电子展(CES)7日至10日举行。人工智能(AI)技术驱动的新应用、新解决方案是本届展会最大热点,AI智能体成为参展企业热议词汇。

科技界对智能体这一概念的理解是多方位的。总体来说,智能体是一种能够执行复杂任务并与人类进行自然交流的AI系统。智能体从多种渠道获取大量数据,能够通过复杂推理、规划来自主解决多步骤问题,确定工作策略并执行。智能体是新的数字劳动力,可以协助甚至代替人类完成不同任务或重复劳动。

美国高德纳咨询公司 will 将智能体列为2025年十大战略性技术趋势之一。业界专家认为,智能体是AI技术衍生的一种新业态,代表了生成式AI发展的下一波浪潮,有助于提高各行业生产力和运营效率。

本届CES上,智能体成为许多参展企业展板和介绍册上的热词,同时也登上CES官网及各大媒体热搜,相关标识在各展厅随处可见。

美国英伟达公司首席执行官黄仁勋在CES主旨演讲中阐述了智能体发展蓝图:开发者可以借助视觉语言模型、大语言模型等工具,构建、测试和运行能够分析大量视频和图像内容的智能体,并将其训练为不同专业领域的任务专家。

黄仁勋表示,智能体有望成为下一个机器人产业,很可能是一个市场体量巨大的机会。英伟达还推出一款名为Cosmos的世界基础模型,该模型能够生成逼真的虚拟现实场景,应用于机器人、自动驾驶汽车等结合物理实体的智能体的开发。

展会上,功能各异的智能体闪亮登场:研究智能体能够轻松对大量繁杂文件进行分析、总结,比如企业财报、论文等,并生成交互式视频,方便使用者学习;视频分析智能体能够分析各类视频数据,比如交通流量、工厂监控、桥梁质量监测等数据,从中摘取核心信息,生成交互式报告,提供改进建议等;软件安全智能体能够帮助开发者持续扫描软件漏洞,提示采取应对措施;虚拟实验室智能体能够加速化合物设计和筛选,加快寻找潜在药物候选者的步伐;陪伴智能体拥有“长期记忆”,能够与使用者贴心聊天,理解和回应其情感需求。

各大展厅里,众多类型的消费电子产品也都有智能体加持:机器人、电动汽车和电动自行车、智慧厨房、智能家居、医疗诊断……各类产品搭载的智能“小助手”在功能性和交互性等方面不断进化。

美国高通公司中国区董事长孟樸在接受新华社记者采访时表示,智能体的核心是建立了一个新的AI入口,无论使用文字还是语音等输入方法,只要将相关要求输入这个端口,具体执行什么任务都可以重新改写。“智能体的应用场景越来越丰富,将大大提高人们生活和工作便利性和效率。”

智能体的发展也为企业开拓了更多创新空间,有助于形成更完善的AI产业链。孟樸认为,与往届相比,今年CES展会一大变化是AI在参展产品中“全面开花”,AI应用从概念到真正实现落地。“AI不再只是吸引观展者的概念,而是能看到更多的AI产品和应用,看到它们具体能实现什么功能。”

算力支撑让人工智能在中国行稳致远

仅需一秒钟,上海的海量数据可抵达甘肃庆阳,由算力基础设施完成分析运算;广东的创意设计团队,也可借助这个“超强大脑”完成渲染超精细的3D模型。

记者在国家数据中心集群(甘肃·庆阳)“东数西算”产业园区看到:智慧屏不断闪烁,数据实时更新。算力服务正在从黄土高原直通长三角和京津冀等地。

中国科学院计算机网络信息中心研究员陆忠华表示,全社会对算力的需求日益增长,人工智能应用的算力需求更加突出。

算力发展的政策部署紧密推开——

2021年5月,我国启动全国一体化算力网络国家枢纽节点建设,加快推进“东数西算”工程;2022年初,我国8个国家算力枢纽节点全部获批,并规划设立10个国家数据中心集群,“东数西算”工程正式启动;2023年,工业和信息化部等6部门印发《算力基础设施高质量发展行动计划》,提出到2025年,算力规模超过300EFLOPS(EFLOPS是指每秒百亿亿次浮点运算次数),东西部算力平衡协调发展。

算力网络“高速路”加速构建——

针对数据中心无法联动带来算力资源的低效配置,制约“东数西算”成效发挥等问题,国家正在通过连通已有的、不同体系架构的算力中心,合理配置、共享、调度、释放更多算力,并降低应用门槛。

贵州省已建成全球首条400G算力通道,与全国38个城市实现网络直联,与全国主要城市平均网络时延在20毫秒以内,实现由“存储中心”向“存算一体、智算优先”的转变;天津市河北区投资12.7亿元建设立足该区、面向全市的算力中心,目前已帮助本地石化、医药和新材料等9条重点产业链解决一系列技术难题,提高了生产效率和创新能力。

中国信息通信研究院云计算与大数据研究所副所长李洁认为,当前,计算产业和网络产业相对独立,存在标准缺失、数据共享不够、资源接口不统一等壁垒,未来仍需大量的标准化工作和技术研究工作。

算力生长与绿色能源“双向奔赴”——

数据中心作为耗电大户,电力成本占其运营成本的60%左右。加快推进数据中心节能降碳势在必行。

“控制数据中心的能耗不仅关乎计算服务的成本,而且关乎国家碳排放的核心发展要求。”中国工程院院士邬贺铨说。

来自2024中国算力大会的数据显示,全国在用算力中心机架总规模超过830万标准机架,算力总规模达246EFLOPS,位居世界前列。算力中心平均电能利用效率(PUE)降至1.47,创建国家绿色数据中心246个。

一批前沿技术正加快应用,助力数据中心节能降耗。如阿里云首创基础设施和IT设备融合一体化浸没液冷数据中心,将所有IT设备浸入专利的冷却液中,PUE达到极低的1.07。一些地区加快引入可再生能源,助力数据中心绿色转型。如宁夏预计到2025年,实现中卫绿色数据中心集群PUE平均值不高于1.2,可再生能源利用率达到65%。

专家表示,越来越多的绿色电力将为算力保驾护航,而越来越聪明的气候大模型将更精确、长周期地预测天气,提高可再生能源发电效率。

世界知识产权组织今年7月发布的报告显示,2014年至2023年,中国生成式人工智能专利申请量超3.8万件,居世界第一。

这些日新月异的大模型,更加需要坚实有力的算力“底座”。必须“装好行囊再上路”,让AI在中国行稳致远。

来源:新华社



工作人员在位于贵州的中国联通贵安数据中心微模块机房进行巡检。

□新华社记者 陶 亮摄