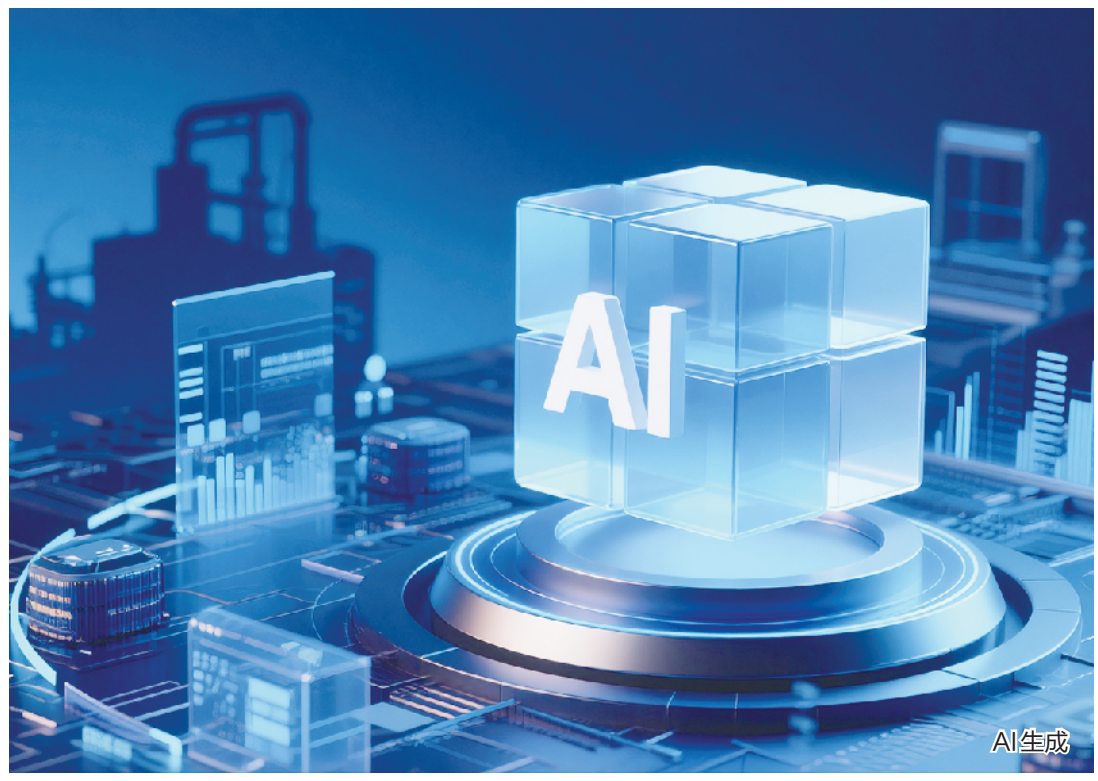
大模型发展“下一站”:全场景生态构建

随着人工智能技术迅猛迭代,大模型已成为驱动经济社会发展的关键引擎。在单一模型技术不断突破、参数纪录屡屡刷新的同时,各家大模型厂商不约而同将目光投向大模型服务链建设,着重构建全场景生态。

国务院不久前印发的《关于深入实施“人工智能+”行动的意见》提出,发展“模型即服务”“智能体即服务”等,打造人工智能应用服务链。面对千行百业智能化转型需求,全链条覆盖的服务能力、优秀的全场景生态构建能力,已成为大模型产业下一阶段发展的重点。



AI生成

加强软硬件协同

对如今的大模型行业来说,“周周有发布,天天有更新”已成为常态。

9月9日,百度正式发布文心大模型X1.1。它采用迭代式混合强化学习训练框架,通过混合强化学习,协同提升通用任务和智能体任务的效果;同时,通过自蒸馏数据的迭代式生产及训练,不断提升模型整体效果。尤其在更复杂的长程任务场景中,模型表现突出。

同样在9月,阿里巴巴旗下通义千问也推出了其最新模型,在中英文理解、复杂指令遵循、工具调用等维度实现性能增强,同时显著减少了知识幻觉。

在追求性能指标不断提升的同时,大模型产业也正走向以配套工具开发、应用场景适配、产业落地为核心的全场景生态能力建设阶段。

大模型生态是一个涵盖底层技术研发、模型训练与优化,以及应用开发、产业落地、人才培养等多个环节的复杂系统。完善的大模型生态能促进技术快速迭代创新,加速人工智能在各个领域的普及应用。

中国工程院院士郑纬民此前曾谈到国产AI芯片的生态问题。他直言,当下国产AI芯片的适配生态不够好,如果生态足够好,哪怕只有60%的性能也会有更多用户使用。

同样的问题也出现在大模型领域。随着人工智能技术快速迭代,仅靠算力堆叠已无法实现模型性能的同步线性提升。模型算法与底层硬件、软件、开发框架间的协同效果显著影响其训推效率与性能表现,人工智能软硬件生态协同的重要性更加凸显。

中国信息通信研究院人工智能研究所软硬件与创新生态部主任李论认为,大模型的升级迭代需要在庞大的软硬件系统上进行实验,模型的原始创新和应用迭代落地已非常依赖先进的软硬件协同技术生态体系,框架、芯片、集群、网络等与算法间的协同愈发紧密。

从全球范围看,一些国内外头部企业正在纷纷加快推进更大规模集群的软硬件协同优化。例如,百度自2010年起布局人工智能,先后发布我国第一个开源深度学习框架飞桨、人工智能芯片昆仑芯,以及文心大模型等一系列人工智能软硬件技术产品,构建起覆盖芯片、框架、模型、应用的完整技术生态。

“我们预计,在下个阶段,软硬件协同和生态体系的构建会是全球大模型创新和算力设施建设的竞争焦点。”李论说。

完善配套工具研发

大模型生态并非“为建而建”,有效降低技术门槛、推动大模型应用广泛落地是关键目标。

一个成熟的大模型产品落地,从训练调优到部署应用,背后是一系列工程环节的紧密协作。当前,阻碍AI应用开发落地的一大瓶颈是大模型训练、部署成本高。这很大程度是模型配套的上下游工具链不健全导致的。

百度AI技术生态总经理马艳军举例说,在大模型的大规模分布式训练中,静默数据损坏是一类通常由硬件设备引发的极具隐蔽性和破坏性的故障。它一旦出现,容易造成训练效果严重劣化甚至失效,难以追溯原因,往往需要停机进行长时间压测才能识别,检测成本极高。针对这类问题,团队自主研发的飞桨核心框架v3.2利用流水线并行训练中固有的“空泡期”,在硬件设备空闲间隙,在线、无损插入运行精度检测代码,无需停机即可精准定位故障硬件,从而在不影响训练的前提下,极大提升超大规模训练的长期稳定性与可靠性。

在部署端,大模型的高性能推理效率和成本是业界长期关注的问题。马艳军介绍,团队基于飞桨研发的大模型高效部署套件FastDeploy 2.2版本,可以提供大模型高效部署及高性能推理全栈能力。这一套件不仅可以服务于文心大模型,还可以兼容多种协议、格式,高效运行文心系列及其他主流开源大模型。

不仅是百度,华为、阿里巴巴等企业也在不断完善大模型配套工具研发,降低大模型开发应用门槛。例如,华为面向昇腾AI开发者提供的全流程开发工具链MindStudio,可为开发者提供端到端的昇腾AI应用开发解决方案,使开发者高效完成训练开发、推理开发和算子开发。

扩大生态“朋友圈”

好用的工具带来的是应用百花齐放,而更多应用方的加入也会让大模型生态更加丰富多元。

中国中车集团有限公司科技质量与信息化部数字化创新处副处长陈鉴分享了人工智能技术在中国高铁气动外形设计上的应用案例。

传统的气动评估方法采用精细建模和高精度模型,计算周期长、使用门槛高、资源消耗大。中国中车集团有限公司以既有的仿真和实验数据为基础,构建高速动车组的气动载荷标准数据库,并基于文心大模型和飞桨的科学计算能力,构建起空气动力学仿真大模型。“一个外形设计想法,过去可能需要做大量实验,几个月才能有答案。现在借助人工智能仿真计算,最快几分钟就可以得到结果,并且准确率很高。”陈鉴说,这一技术的应用大大加速了气动外形设计迭代,显著提高研发效率。

不仅如此,该公司还与百度飞桨联手打造出国内首个虚拟传感器模型。该模型可在不增加既有传感器的基础上,根据车辆已有的电流电压等现成数据,通过一系列计算,推算出与车辆安全运行相关的其他数据。这仿佛给车辆安装上多个虚拟传感器,可以更早发现故障隐患,对车辆进行更为精细的健康管理,使故障检测准确率在现有传感器检测的基础上再提升10%。

人才是支撑大模型生态建设的重要基础。目前,百度已联合湖北省12所高校共同成立百度飞桨(湖北)人工智能教育创新中心,湖北省22所高校基于飞桨与文心开设学分课程。同时,百度飞桨作为副理事长单位参与湖北人工智能学院建设。

多位专家认为,随着底层技术不断优化、配套工具日益丰富、应用场景加速落地,大模型产业将加快从技术迭代走向生态构建。

据《科技日报》 作者:都芃

人工智能个人电脑走出“新手村”

近日,联想、雷神、宏碁、积核等PC(个人电脑)企业在柏林国际电子消费品展览会上展出了AI PC产品。虽都是新品,但侧重点各有不同:积核展出的GEEKOM A9 Mega更强调本地AI算力;雷神强调其三引擎嵌入式神经网络处理器(NPU)架构、游戏内AI对话、实时多语言翻译等电竞功能;联想则在电脑形态上下了功夫,不仅展示了显示屏的横纵切换,还展示了可实现面部追踪、语言控制和满足人体工程学健康功能的AI电脑支架。

AI PC究竟该长成什么样,能发展到什么程度,距离“终极形态”还有多远?

离“终极形态”还有差距

AI PC应该是什么样的?各PC品牌供应商对AI PC的形态和功能有着相似的美好期许。

在产品形态上,小米公司软件产品负责人刘靖超表示:“AI PC更像是伙伴、助手,具有了灵魂,有主动性。无论传统PC还是AI PC,虽然对用户的终极价值都是提升效率,但是二者的上限不同,传统PC的上限是用户,而AI PC可以帮助用户突破自我。”上海英众信息科技有限公司副总经理邵世佳认为:“PC不再只是‘应用的容器’,而是能随时理解语音、图像、文字,并进行本地推理的智能体。”

在硬件表现上,荣耀终端股份有限公司PC产品总经理朱臣才期待通过AI调度与优化,让PC在性能、续航、通信、静音、音效、显示等方面实现全方位提升,成为兼具轻、薄、强、静、优、久特性的“多边形战士”,为用户带来均衡且卓越的使用体验。邵世佳则希望,采用“CPU(中央处理器)+GPU(图形处理器)+NPU”的异构协同,NPU负责常驻AI推理,保证高能效。

在交互体验方面,朱臣才表示,荣耀将大语言模型与多模态模型应用于PC的人机交互优化,提升办公与学习效率。荣耀PC的智能体可实现“一句话自动驾驶”,用户只需口述意图,智能体便可自动完成文档总结并将结果发送至手机,实现高效协同。基于AI意图识别,荣耀PC能够与手机、平板电脑、打印机等多设备实现以人为中心的无缝协同,让服务与数据在多终端间自然流转,进一步提升跨端生态的智能化体验。邵世佳期待:“在系统层面,操作系统直接提供模型调用接口和多模态交互能力,AI能力嵌入到快捷键、任务栏、侧边栏等系统级入口;在功能层面,本地模型仓可在无网络下完成会议纪要、实时翻译、屏幕理解等任务,既降低延迟,也保护隐私。”

当前市面上能够见到的AI PC产品已经或多或少地实现了上述愿景,但距离理想中够智能、懂用户、能自我进化的产品形态还远远不够。甚至从AI技术与产品结合的密切程度来看,PC类产品的表现还不如手机、平板电脑等移动设备。PC产品的本地算力、模型应用能力以及可支撑的模型规模还不能满足用户的期待。

“现阶段AI PC是指具备一定的AI功能的PC终端。”华为AI终端首席产业代表史浩说,“距离真正意义上的AI PC还有不小的差距。”

史浩举了个例子,就像是汽车自动驾驶级别分类一样,如果根据PC产品的智能化程度进行分级,那么现在市面上大部分AI PC产品,可能仍在最基础的智能阶段上下徘徊,与AI PC理想中的“终极形态”至少还有3至5年的迭代时间。

应用落地是关键

2024年被认为是AI PC元年。市场分析机构称,预计从2025年至2034年,AI PC市场的年复合增长率将达到42.8%。

而支撑市场增长的有三大因素——产品体验、应用落地、产业推动。

在这三者之中,产业推动是基础。上下游厂商在NPU硬件、大模型生态、开发工具上共同投入,降低了应用落地的门槛;应用落地是关键——更多多模态、个性化的AI应用上线,使用户直观感受到价值;产品体验是根本——端侧大模型使PC在办公、学习、创作等场景下更智能,形成差异化卖点,才能吸引到更多消费者购买。总而言之,如何利用硬件基础,整合应用资源,给消费者带来更好的体验,才是决胜AI PC应用市场的关键所在。

这一逻辑对于电脑供应商而言同样适用。日前,联想公布的最新季度财报称,2025年4月至6月,联想AI PC产品的出货量占到联想个人电脑总出货量的30%以上;天禧个人超级智能体的用户活跃度也保持了增长势头,平均周活跃率达到40%。

“应用”同样是联想的发力重点。联想相关负责人称,在联想的天禧生态中,智能体层面已经聚合了超过2000家企业入驻天禧AI空间,涵盖教育、视频、办公、创作等应用,以更加丰富的场景推动智能体落地与创新;模型方面则覆盖全部主流大模型,包括文心一言、通义千问、百川等,客户可以自行选择和配置,实现多模型无缝切换;算力和系统方面实现了国际和国内全兼容。

产品迭代逻辑变革

为什么当前仍有很多消费者不愿意为AI PC买单?其根本原因在于,PC的AI性能还达不到消费者期待。一方面,缺少杀手级应用场景支撑用户换机;另一方面,应AI需求采用的高性能硬件推高了整机价格。

在这场由AI带动的终端变革中,有一项重大的产业发展逻辑变化:终端形态和功能的变化,越来越难以靠一家或者个别几家企业主导。要想真正实现AI PC的应用体验跃迁,将需要整个生态链的协同发力。

举个简单的例子,相较于云端智能的解决方案,“隐私性”是AI PC这种端侧设备吸引消费者购买新机的杀手锏。但要想使大模型真正能够实现数据处理的私密性,却很难做到。在本地学习用户数据、构建用户专属个人知识库、帮助用户打造更具个性化的AI体验,这对于本地算力、模型规模和运算效率乃至整机功耗和整机成本来说都是巨大的考验。

“隐私性”简简单单三个字,仅这一项性能的提升,就需要多方的协力配合:首先,芯片供应商应提供具有足够算力的处理器;其次,模型开发商要通过量化和蒸馏,让小模型在端侧就能完成大部分需求。而电脑、系统开发商则要对用户的各项任务进行调度编排,使不同类型的处理器能够高效、低耗地处理各项任务,尽量让更多的需求在本地解决,少量复杂任务交给云端。只有实现性能、体验和隐私三者的平衡,AI PC才有机会实现在提升用户AI体验的同时保证隐私性。

在这一过程中,终端厂商的身份悄然发生了变化。联想相关负责人表示:“终端厂商将从产品提供者进阶为生态组织者,以场景需求为基础面向用户整合产业资源,提供软硬件一体的交付体验。”

在这样的产业发展逻辑下,要想实现AI PC产品体验感革命性变化,需要产业链各环节各司其职,分头发力。

在硬件层,NPU算力需持续提升,使参数量在70亿至130亿的模型能够在本地运行;同时通过更高带宽的大缓存,缓解大模型的带宽压力。

在模型层,端侧大模型的效率和精度需持续提升,实现更好的量化、稀疏和长上下文支持,让小模型也能接近云端体验。在系统层,系统开发者要改进CPU、GPU、NPU的调度和内存带宽管理,使本地AI既能流畅运行,又不会拖累续航和散热。在生态层,建立统一接口和工具链,推动应用场景更深入地融入工作流,使AI应用能在PC上跑起来,而不是停留在“演示功能”层面;同时通过安全内存等技术,保障AI在本地运行既安全又可审计。

此外,在PC向着更智能、更省电、更懂用户的方向发展的进程中,整机企业也在持续推出优化方案。REDMI Book Pro 2025系列发布AI智能亮度功能,能够通过识别用户场景和用户使用偏好,智能调整屏幕亮度,为用户提供个人化的护眼体验;联想推出多项AI终端原生技术,包括多模态自然交互、终端推理加速引擎、可信计算、主动检索增强生成等。

面世一年有余的AI PC仍然是新鲜事物。就像几年前的新能源汽车一样,正处于百家争鸣之时,形态、功能各有侧重,产业生态和市场规模也在向前推进。说不定再过几年,消费者可以拥有一款真正意义上的、具有颠覆性创新的AI PC。

据《中国电子报》 作者:姬晓婷

AI算力驱动需求 液冷散热产业链渐成熟

伴随着全球AI算力的需求上涨和芯片功耗的攀升,液冷正取代传统风冷被视作AI时代下解决高算力芯片散热问题的主流方案,业界看好液冷市场规模的增长和技术突破。

近日,英伟达正推动上游供应商开发一类名为微通道水冷板的水冷散热组件,以应对AI图形处理器芯片随代际更替不断上升的发热,单价是现有散热方案的三至五倍。

中信建投在近期的一份报告中指出,2025年是英伟达AI芯片液冷渗透率大幅提升的一年,随着单芯片功耗的提升,后续液冷市场规模将明显增长。伴随液冷产业链成熟度的提升,液冷在国内市场的渗透率预计也将快速提升,进一步打开市场空间,建议重视液冷板块。

根据中国信通院测算,2024年我国智算中心液冷市场规模达到了184亿元,较2023年同比增长66.1%,预计到2029年我国智算中心液冷市场规模将达到约1300亿元。

不少国内企业正在加快相关部署。润禾材料近日在投资者互动中表示,截至目前,公司冷却液产品已量产并形成销售。未来将结合产品的客户需求与战略发展规划适时推进产能建设规划。申菱环境近期也在互动平台表示,液冷是公司重要战略业务,当前面临较好发展机会,市场端的需求随着AI发展而快速增加,公司新的液冷产线已经投产。预计未来相关业务持续快速发展。

不过,值得注意的是,国内液冷产业仍处于发展初期,多家液冷服务器概念上市公司曾表示,相关业务的收入仍有限。

今年8月,大元泵业发布公告称,2025年一季度,公司能够识别的直接用于数据中心液冷的产品销售收入约160万元,仅占公司营业总收入的0.43%。半年报电话会中,管理层表示,公司目前对数据中心液冷领域业务的战略定位较高,在产品端和销售端投入有倾斜资源,但相关行业仍处于较为初期的研发阶段,尚无法确定未来行业成熟后具体产品对应的价值量和毛利率情况。

淳中科技近日表示,关注到市场对液冷服务器概念板块关注度较高,公司业务不涉及液冷服务器的生产制造,仅参与液冷测试平台等测试环节,2025年上半年该业务未形成收入。伴随着AI算力时代的到来,“重押”液冷正在成为行业趋势,但产业爆发仍需要时间。

据《第一财经日报》 作者:陈杨园