

当前,人工智能技术创新呈现群体性突破,深刻改变人类生产生活方式,重塑全球产业格局。从2018年首次举办世界人工智能大会至今,上海跑出了令人瞩目的“AI加速度”,产业规模不断增加。算力筑基、模型引领、应用赋能……通过构建全链条产业生态,上海人工智能产业上下游协同、软硬件并进,创新成果加速涌现。

上海壁仞科技股份有限公司的展厅里,几台黑色算力柜相排而立,状态灯、链路灯不断闪烁。在算力“燃料”的支持下,AI正从单纯的内容生成向自主执行任务、深度参与科研、赋能工业生产转变。

当前,人工智能技术加速迭代演进,进入前所未有的活跃期。国家数据局发布的《全国数据资源调查报告(2025年)》显示,截至2025年底,全国智能算力规模达159万P(FP16,即每秒千万亿次浮点运算),居全球第二位;全年用于AI训练和推理的数据总量达199.48艾字节;全年词元调用量约21100万亿。

从算力基础设施到大模型创新,再到真实场景应用,上海人工智能产业链构建了完整闭环生态,成为观察我国人工智能产业发展的样本城市。截至目前,上海智能算力规模突破17万P;394家规上人工智能企业产业规模超6370亿元,同比增长39.5%。

7月17日至20日,2026世界人工智能大会暨人工智能全球治理高级别会议(2026WAIC)在上海举行,汇集1100余家企业,超300款产品实现全球首发,全景呈现AI产业从底层算力到应用生态的未来走向。作为前沿技术的展示窗口、加强对话交流的平台,世界人工智能大会助力上海加快建设全球人工智能创新高地。

先手布局

时间拨回到2019年,彼时,人工智能产业经历一轮投资热后进入冷静期,行业普遍关注AI究竟能否走出实验室。

这一年,曾任商汤集团股份有限公司总裁的张文在上海创立壁仞科技,主攻高性能通用图形处理器(GPU)赛道。

一年后,多位拥有多年高性能GPU研发经验的业界“老将”,在上海成立沐曦集成电路(上海)股份有限公司。“当时算力需求还没有爆发,但我们判断,伴随AI进一步发展,通用、稳定、可靠的算力将变得更加重要。”沐曦股份首席产品官、高级副总裁孙国梁说。

为什么选择上海?孙国梁坦言,科技创业的难点在于需求与技术之间经常存在“时差”;等需求真正爆发后再补技术,往往已经来不及;但在需求尚不清晰时提前投入,企业要穿越漫长的不确定期。上海的优势,恰恰在于通过构建资本、人才和产业链的生态闭环,有效缩短了技术从实验室走向市场的距离。

2021年前后,沐曦股份处于高端GPU芯片研发攻坚阶段,上海科创中心基金参与其A轮融资,支持企业持续研发。

对企业而言,除了资金,产业链协同优势同样重要。在上海,从芯片设计、制造、封测到设备材料,企业在较小半径内就能找到合作伙伴,实现全链条配套,这种协同有助于研发速度提升。2022年底,沐曦股份曦云C500交付流片;2023年6月芯片回片后,仅用5小时便完成芯片功能测试。此后,公司继续向软件生态和集群交付延伸。截至2025年底,沐曦股份GPU产品累计销量超5.5万颗,部署于10余个智算集群。

产业成长离不开城市的战略布局,上海还有诸多类似的“先手棋”。壁仞科技、上海天数智芯半导体股份有限公司、燧原科技等算力企业的成长过程中,都能看到上海国资等资本的身影。2026年以来,上海GPU芯片企业密集登陆资本市场,掀起了AI芯片抢占资本赛道的热潮。

如何围绕AI基础设施,构建更完整的产业链?在上海徐汇西岸,全国首个大模型创新生态社区“模速空间”坐落于此。其首批入驻企业上海无问芯穹智能科技有限公司是上海在算力产业链上落下的一枚“枢纽棋”。

当算力需求日益旺盛,如何把不同芯片、模型和场景高效匹配起来?如何做好算力调度与服务?“上海既能‘聚’资源,又能‘转’价值。”无问芯穹总经理曾书霖告诉记者,公司从7人团队、不足50平方米的办公空间起步,如今已成长为300多人的创新团队。2025年,无问芯穹与模速空间、上海仪电、沪芯沪模等产业上下游企业联手打造国内首个算力生态超市,让无问芯穹大模型服务平台在实践中不断优化,让每一度电、每一颗芯片生产更多词元。

“依托这种生态自组织能力,无问芯穹Agentic MaaS大模型服务平台日均词元调用量相较于去年底增长超40倍。”曾书霖说,在2026世界人工智能大会上,无问芯穹展示了“前店后厂一中心”自进化Agentic Infra(通用人工智能基础设施)业务架构,该架构由算力集散中心+词元工厂+AI生产力商店构成,实现“用智能进化智能,用生产力加速生产力”的完整闭环。

从建设算力到调度算力,上海的先手布局逐步覆盖芯片、软件、平台和服务,一条算力基础设施链条开始成形。



技术迭代越快,想要依靠某项技术的单点突破来带动产业整体跃升就越困难。如今,产业创新越来越像一场系统作战。在智算领域,一颗GPU芯片能否充分发挥作用,不仅取决于芯片本身,还取决于存储、互联、服务器、散热、供电、软件和应用等环节能否协同。

在这方面,上海独具优势。壁仞科技AI框架架构副总裁丁云帆向记者讲述了一次由4家上海企业共同完成的前沿技术攻关。

故事从迅速增长的算力需求说起。当前,大模型正向万亿级参数、多模态混合专家(MoE)架构和超长上下文演进,智能体将AI从简单问答推向代码开发、方案设计等复杂长程任务,对训练特别是推理算力提出了更高要求,但这依靠“堆卡”并不能解决问题。“这类类似于企业部门协作,简单任务可以由每个人独立完成,复杂任务则需要跨部门协作。如果资源调度不合理,沟通速度不够快,即便拥有上万名员工,也很难发挥上万人的效率。”丁云帆说,模型规模越大,GPU之间的数据交换越频繁,通信就越容易成为系统瓶颈。

超节点技术由此走向前台。它通过高速互联把多张GPU、多台服务器连接成一个整体。然而,新的难题随之而来:卡与卡、服务器与服务器之间,如何实现更大带宽、更低时延和内存语义的高效通信?“传统电互连依靠电信号传输数据,随着传输距离拉长,信号衰减、干扰和稳定性问题会逐渐显现。”丁云帆介绍,前沿的光互连技术或许能提供解决方案,但对于将光互连真正装进国产GPU超节点,当时国内还缺少成熟的工程化验证,没有人敢“第一个吃螃蟹”。

巧合的是,在上海另一支团队已在光方向探索多年。

2016年,沈亦晨在国外知名高校完成博士研究。他判断,摩尔定律已逼近极限,而光在一些方面能突破晶体管本身的限制。“AI集群规模不断扩大,传统电互连的带宽、延迟和功耗会成为越来越突出的瓶颈;而光互连、光交换在科学机理上存在突破的可能。”沈亦晨不想空等机会,于是选择创业,成立了上海曦智科技股份有限公司。

在上海,产业链上下游企业距离近,合作反馈周期快,曦智科技就是受益者。2018年,沈亦晨将曦智科技研发中心转移至上海浦东,因为这里有高密度的产业集群。

很快,曦智科技与壁仞科技“擦出了火花”。2025世界人工智能大会上,由曦智科技、壁仞科技、中兴通讯联合打造的“分布式OCS全光互连芯片及超节点应用创新方案”备受关注。随后,上海仪电联合曦智科技、壁仞科技、中兴通讯发布国内首个光互连光交换GPU超节点“光跃LightSphere X”;2026年3月,国产光互连光交换超节点“光跃128卡商用版”正式落地2048卡集群。该方案构建起高带宽、低延迟、灵活可扩展的自主可控智算集群新范式,端到端性能相比传统技术提升30%以上。

在这项合作中,4家企业紧密协同、各司其职;壁仞科技为接入光互连,重新设计GPU板卡及全新载板互连;曦智科技负责光电光转换、光信号传输与交换;中兴通讯将GPU板卡、光互连部件与服务器系统集成;上海仪电提供智算平台和应用场景。如果缺少其中任何一方,整套系统都难以大规模落地运行。

“当时技术很新,产业化存在不确定性。最终大家愿意共同承担风险,不仅因为产品本身具



上海振华重工集团位于南通生产基地的智能板单元车间里,多套智能制造设备构成全链条智能制造体系,有效提升钢结构产品的制造效率与品质。(资料图片)

有创新和落地价值,也因为上海大力支持AI前沿技术发展。”丁云帆说,这是上海的前瞻性布局,它不仅解决眼前的算力需求,更为下一代超大规模算力系统储备新的互连技术。“从前沿技术到真实系统,光互连需要在实践中持续历练。”沈亦晨说。

面对Agent应用和巨大的算力需求,壁仞科技即将推出新一代产品,实现架构、内存、带宽、互连和软件的全面升级,打造“芯片+系统+生态”全栈交付能力。壁仞科技还将与多家生态伙伴共同致力升级NPO光互连技术,打造全光互连、分布式解耦架构的超节点,提供高性价比的Token解决方案。

2026世界人工智能大会暨人工智能全球治理高级别会议期间,类似的跨界合作还有很多。本届大会围绕世界模型、开源智能体、Token经济、OPC等议题,深入探讨行业新趋势。展览首发首秀多款AI新产品,包括MiniMax M3多模态大模型、阶跃Agent操作系统、近存计算3D芯片,以及多款人形机器人、AI灵巧手等。

落地成势

算力底座是否有效,最终要看能否支撑模型进化,解决实际问题。

成立于2022年初的上海稀宇科技有限公司通过自主研发多模态通用大模型,已服务超过200个国家和地区的逾3亿名个人用户,并服务上百万企业客户和开发者。

如何在短时间内取得如此亮眼的成果?公司副总裁严奕骏说,公司坚持技术驱动,但光靠模型本身是不够的。“从上游半导体、芯片制造到软件、云平台,再到人工智能本身,产业链协同能力决定了公司的发展速度。上海构建的AI产业链生态,让技术积累和产业落地同时推进成为现实。”

调查手记

构建正循环生态

李治国

人工智能产业的竞争,是创新生态与制度效率的竞争。上海在人工智能领域构建产业正循环,算力基建、产业链协同与应用场景等供需互促,为我国人工智能产业发展提供了可参考的样本。

上海模式的深层逻辑,在于以耐心资本、产业密度、场景开放的组合策略,缩短了从技术研发到商业验证的链路和周期,由区域创新系统分散经营主体难以独自消化的系统风险。

推进人工智能产业向更高层级发展,要构建分层的算力战略,避免低效同质竞争。在全国算力网络“东数西算”框架下,统筹长三角、京津冀、粤港澳大湾区等区域产业资源,建立“国家调度+市场交易”的双层运营机制,将地方法算力生态超市等探索,整合为全国算力互联互通的标准协议与清算结算体系,让分散于各地的产业资源从物理连接走向价值连接。

建设跨主体协同制度。在人工智能产业链条中,芯片、互联、软件、算法具备高度耦合特性,单点突破难以转化为系统竞争力。对此,可以借鉴“链长制”经验,设立AI基础设施系统创新专项,以“揭榜挂帅”方式组织芯片设计、封装测试、云平台、场景方组成创新联合体,并建立风险共担基金与首台套保险补偿机制,让联合攻关从偶发性的企业自发行为上升为制度化的行业组织行为。

建立场景驱动创新的对接平台。上海以场景开放牵引技术落地的实践表明,需求侧的“用”是带动供给侧的“研”的关键杠杆。各地可以根据自身条件,发布“AI+行业”场景开放清单,建立场景需求与服务商能力的标准化对接流程,形成“需求定义—技术响应—迭代优化—标准推广”的正向循环。

从宏观层面来看,制度创新是产业发展关键变量。AI产业的“正循环”不能止步于市场的自发演进,需要以制度之力疏通堵点、降低交易成本、稳定市场预期,推动我国人工智能产业在全球竞争中形成持续的发展优势。

本版编辑 闫伟奇 美 编 夏 祎