大模型“混战”

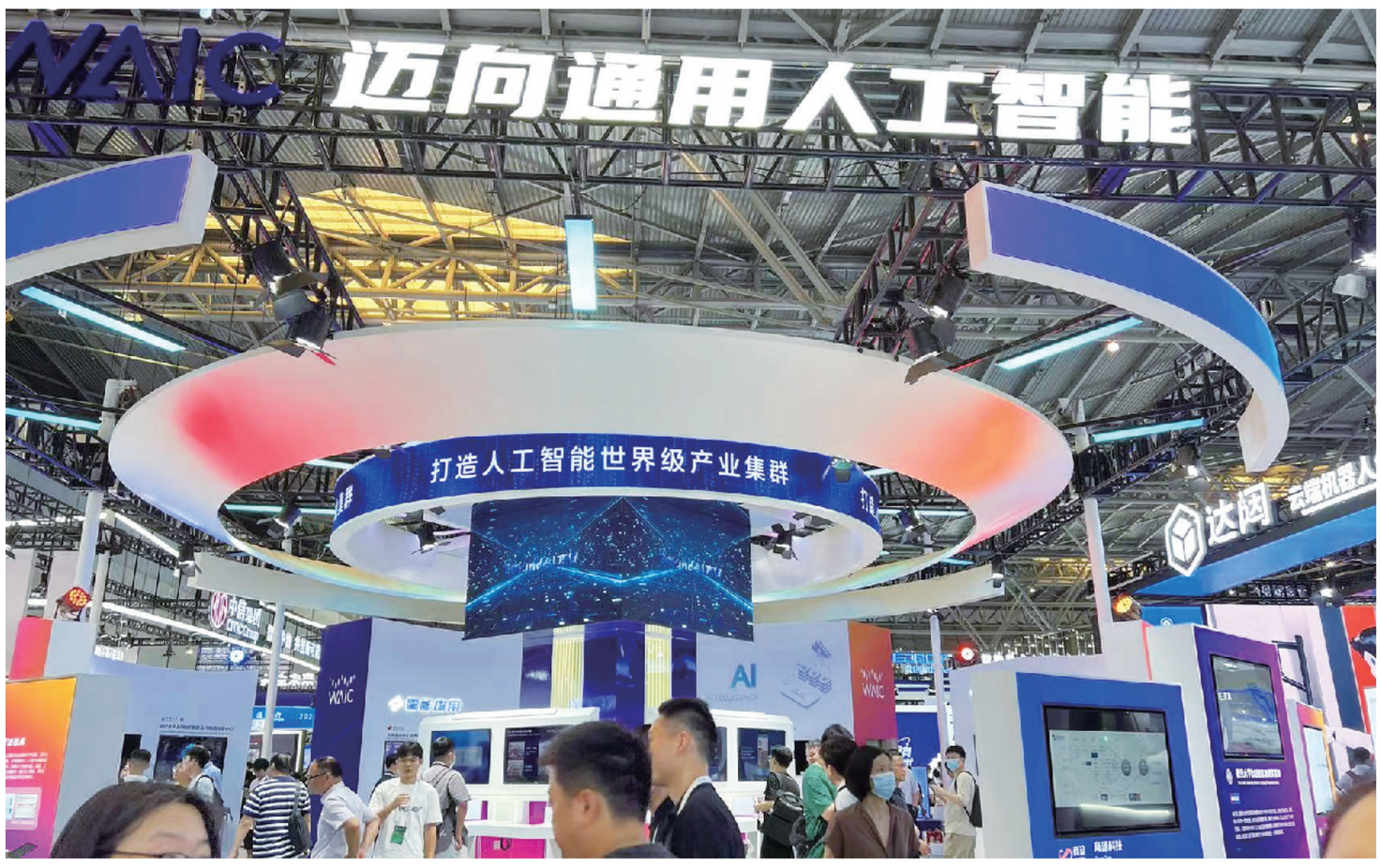
◎ 记者 马云飞 实习生 钱晨珠

7月6日上午10点，连续多天高温的上海依旧“热浪滚滚”。以浦东新区博成路为中心，四周沿建筑物延伸出的广告牌下、通往世博展览馆的路上、展厅前的空地上，早已人声鼎沸、人头攒动。以“智联世界、生成未来”为主题的2023世界人工智能大会（WAIC）在此拉开帷幕。

根据官方披露的数据，此届世界人工智能大会，无论参展企业数量还是展览面积均创历届之最，5万平方米的世博主展览馆吸引了超过400家参展企业，首发首展新品达到30余款。

“今年的大会与往届相比，有一个最大的不同——在一个新的背景之下，即去年年底ChatGPT的出现，将人工智能推到了一个新的风口上，可以说人工智能、尤其是通用人工智能，在当前这个时期，已经成为了人类社会最热门的话题，没有之一。”在2023世界人工智能大会开幕式上，华为轮值董事长胡厚崑表示，他在展览区转了一圈，只关注到了两件事：一方面是大模型的研究，另一方面是大模型在不同行业的应用。

正如胡厚崑所言，今年的人工智能大会展区可以称得上是“大模型”的竞技场，百度的文心一言、阿里巴巴的通义千问、华为云盘古大模型、讯飞星火、商汤日日新等AI大模型产品令人目不暇接。在加速大模型方面，底层AI芯片亦是本届大会的亮点之一，瀚博半导体、燧原科技、算丰、海飞科等算力芯片企业亦在现场争相“秀肌肉”。



马云飞 摄

1 通用大模型：遍地开花

随着ChatGPT的横空出世，基于大模型的人工智能技术发展进入新阶段，科技部人工智能发展研究中心5月底发布的《中国人工智能大模型地图研究报告》显示，当前国内10亿参数规模以上的大模型已发布79个。

在此基础上，“大模型”相关话题也自然成为本届世界人工智能大会最热的焦点，国内通用型大模型（指能够处理多种任务和领域的模型）顶尖产品几乎悉数到场。

今年3月，作为国内AI领域的代表企业之一，百度发布大语言模型“文心一言”，成为我国首个类ChatGPT产品。在此次世界人工智能大会上，“文心一言”大模型登台亮相。记者留意到，在百度展台上，“文心一言”的官方介绍这样写道：在中国最新涌起的互联网革命浪潮中，“文心一言”可以称之为国内最有希望在短期内赶超国际水准的AIGC产品。文心一言的综合评分已与ChatGPT所差无几，在国内大语言模型中位列第一。

根据“文心一言”技术人员在展台现场的解说，“文心”的设计初衷就是打开一个AI与人类真正能够沟通交流的窗口，克服了以往看似智能、实则“傻瓜式回应”的缺陷，当“文心”技术应用到未来的各个平台，例如电商平台客服，作为集成了大模型理解能力、推理能力和学习认知能力的智能客服，拥有可以自主地解决退货或者重新发货问题的能力，这些便是它的进步所在之处。它代表的是一种通用的能力，而不仅仅是写一首诗或者画一幅画这样简单的操作。

阿里云则带来了今年4月推出的“通义千问”大语言模型。根据展台工作人员介绍，通义千问具备多轮对话、文案创作、逻辑推理、多模态理解、多语言支持等功能。具体到应用上，通义千问可以跟人类进行多轮的交互，也融入了多模态的知识理解，有着非常强的文案创作能力，能够续写小说、编写邮件等。

此外，在展会上，华为的盘古大模型亦引人驻足。作为一个完全面向行业的大模型，盘古大模型包括NLP大模型、多模态大模型、视觉大模型、预测大模型和科学计算大模型，可在各个垂直场景应用，已融入金融、制造、政务、电力、煤矿、医疗、铁路等10多个行业，支撑400多个业务场景的AI应用落地。而在此次展会上，药物分析大模型、矿山大模型、气象大模型、海浪大模型等均纷纷亮相。

科大讯飞则携带星火认知大模型及行业应用亮相今年世界人工智能大会。记者在现场注意到，科大讯飞的认知大模型应用展示细分成了八大板块，分别为工业、金融、汽车、数字员工、办公、教育、医疗和SparkDesk，旨在以多元能力融合垂直场景。

其中医疗板块的咨询热度极高，现场的讲解员表示，该系统适用于医院或医疗机构，医生可以为出院后的患者提供全方位诊后管理，并通过识别病人出院时的医嘱和病历信息，形成一个病人的基本画像，基于《医疗指南》自动生成个性化的康复计划，涵盖用药指导、康复运动指导、健康饮食建议、复诊提醒等全部场景。“康复计划”在被医生审核通过之后，病人可以通过小程序，实时接收到干预的内容。

围绕降低大模型开发门槛，商汤科技展示了“日日新sensenova”大模型体系，以及在该体系下的一系列大模型产品更新和落地成果。

“大模型的突破掀起了人工智能的新一轮技术革命，随之而来的是产业需求呈现爆炸式增长，全新的应用场景和应用模式正迅速涌现。商汤希望通过‘大模型+大装置’持续推动AI基础设施能力的跃进提升，不仅打造通用能力更加强大的基础模型，也进一步高效融合不同垂直领域的专业知识，构建更懂行业、更具专长的专业大模型，从根本上降低大模型的下游应用成本和门槛。”7日7日上午，在“大爱无疆·日日新”人工智能论坛上，商汤科技董事长兼CEO徐立如是表示。

马云飞 摄

2 垂直大模型：百舸争流

当下，以AIGC技术为代表的新一轮人工智能浪潮正在席卷。除了通用大模型外，在此次世界人工智能大会上，多家企业展出了如在政务、公共安全、金融等领域陆续落地的，以及更具针对性的、符合垂类场景需求的行业大模型。

作为WAIC的“老朋友”，大数据基础软件供应商星环科技已六次参展。在此次大会上，该公司重点展示了两个行业大模型：针对大数据行业全生命周期各场景的大数据分析大模型Transwarp SoLar求索，以及面向金融量化领域、超大规模参数量的Transwarp Infinity星环无涯金融大模型。

其中，星环无涯金融大模型是业界首款面向金融智能量化投研的领域大模型，支持股票、债券、基金、商品等市场事件的全面复盘、总结及演绎推理，以及政策研报的深度分析；而求索是基于通用大语言模型，通过对大数据分析领域语料的重新训练微调而产生，相较于通用大语言模型，可以更好地理解大数据分析领域的专业术语、缩写、常见词汇和语法，更适合用于大数据分析领域的自然语言处理任务。

“未来在金融、政府、能源、交通等每一个行业与领域，都会诞生领域或者行业的大模型，这些大模型具有专家的能力，可以在上面构造复杂的应用。”星环科技创始人、CEO孙元浩向包括《国际金融报》在内的记者表示，在特定领域，行业大模型将会成为发展的主流，比如金融行业会出现金融量化大模型，为基金经理投资提供决策辅助支持；在传统行业比如冶金领域，基于大模型驱动的控制技术应用将得到快速发展。

“由于大模型反馈的结果是基于对训练语料的学习而产生的答案，因此行业大模型在具体的落地过程中，需要学习大量行业的专精语料和经验知识，才能确保返回结果的精准性和专业度。”孙元浩进一步指出。

语言智能科技企业蜜度亦重点展示了三款行业领域AI大模型，应用于出版、媒体、政务、教育等行业。具体包括用于辅助写作、新闻稿件辅助生成的蜜巢知识问答与内容生成大语言模型，这是首个软硬件一体国产化知识问答与内容生成大语言模型；国内首个智能校对领域大语言模型“蜜度文修”；蜜巢智能舆情分析大语言模型。

“通用大模型的研究，是为了让不具备模型构建能力的企业能够享受到大模型带来的便利，更适合实力更强的头部企业去做，行业龙头公司则更适合去聚焦行业，去做垂直行业大模型。”蜜度首席技术官刘益东在接受采访时称，目前，蜜巢系列行业大模型已逐步尝试部署在政务、媒体等内容生产强需求场景当中。与众多平台型公司专注于通用大模型不同，蜜巢是为解决企业端政务端用户需求而诞生的垂直行业大模型，这也是公司对于大模型赛道的态度，即以解决实际场景问题为目的来做大模型。

在此次大会上，达观数据也正式发布了名为“曹植”的大模型。据该公司相关工作人员介绍，作为一款垂直专用的国产大语言模型，“曹植”具有长文本、垂直化和多语言的特点，将为金融、政务、制造等垂直领域提供智能协作、语义分析等各种AIGC应用。

“大模型未来的发展关键要与垂直行业相结合。未来大模型真正得以运用，还是需要和每一个垂直行业的深度结合，去解决每一个行业里面的真正痛点才行。”根据达观数据董事长陈运文的说法，“大模型未来在企业的落地形态一定是大模型和多个企业垂直小规模的组合，真正的机会在垂直行业市场落地。”

3 AI芯企：竞秀“肌肉”

“大模型应用目前处于从0到1的阶段，算力‘普惠’存在挑战，相比训练阶段，大模型的推理部署需要更大的算力。”在此次展会期间，燧原科技董事长兼CEO赵立东称，大模型目前处于技术萌芽期，按照GPT-3的训练成本推算，GPT-4的训练成本高达数十亿美元；待到技术膨胀期与应用萌芽期，推理AI业务需要增加360亿美元的成本。

不可否认的是，大模型蓬勃兴起，算力成为这一波产业浪潮的最核心因素。在此次展会上，瀚博半导体、天数智芯、燧原科技、登临科技、算丰、海飞科等国内AI芯企均展出了用自家产品运行大语言模型、AI绘画、文生PPT等AIGC交互演示。

作为国内人工智能算力提供商，燧原科技已连续第四年参会，此次带来了多款产品，包括人工智能训练芯片邃思2.0、人工智能训练加速卡云燧T20、人工智能训练OAM模组云燧T21、人工智能推理芯片邃思2.5、人工智能推理加速卡云燧i20等，并展现出全新文生图MaaS平台服务产品“燧原曜图™”（LumiCanvas™）。

在燧原科技展台上，记者留意到，以燧原曜图为代表的数个可交互设施组成了引人入胜的“互动体验站”，此外该公司亦展现了多款搭载云燧i20、云燧T20、邃思2.0等产品的解决方案。据工作人员介绍，目前燧原科技的众多产品已在智慧城市、智慧金融等领域投入使用。

昆仑芯主要展示的新产品包括第二代AI芯片和基于第二代AI芯片的加速卡产品R200系列。记者在展台现场了解到，昆仑芯2代AI芯片是国内首款采用显存的通用AI芯片，对于推动国内AI芯片技术研发和商业落地都具有重要价值。相较第一代产品，这款芯片通过芯片架构、指令集等底层核心技术的优化，实现性能、能效、易用性的提升，通用计算核心算力提升了2至3倍。

高端GPU芯片公司瀚博半导体亦是WAIC的“老朋友”。此次展会上，瀚博半导体发布了包括SG100全功能GPU芯片、LLM大模型AI加速卡以及高性能生成式AI加速卡等6款新品，为AI大模型、图形渲染和高质量内容生产提供完整解决方案。

据瀚博半导体展台工作人员向记者介绍，瀚博SG100芯片采用7nm先进制程，具备业界领先的渲染性能，同时兼具低延时高吞吐的AI算力和强大的视频处理能力。此外，该工作人员亦提及，随着大模型技术的不断发展和应用，GPU行业也将迎来更多的机遇和挑战。具有并行计算能力的GPU芯片作为大模型计算的“大脑”，将为大模型生成学习提供源源不断的算力支撑。而瀚博半导体本次新品发布将持续助力更多AIGC应用，进一步拉低文字、图像、视频等优质内容创作的门槛。

（实习生黄琬婷对此文亦有贡献）