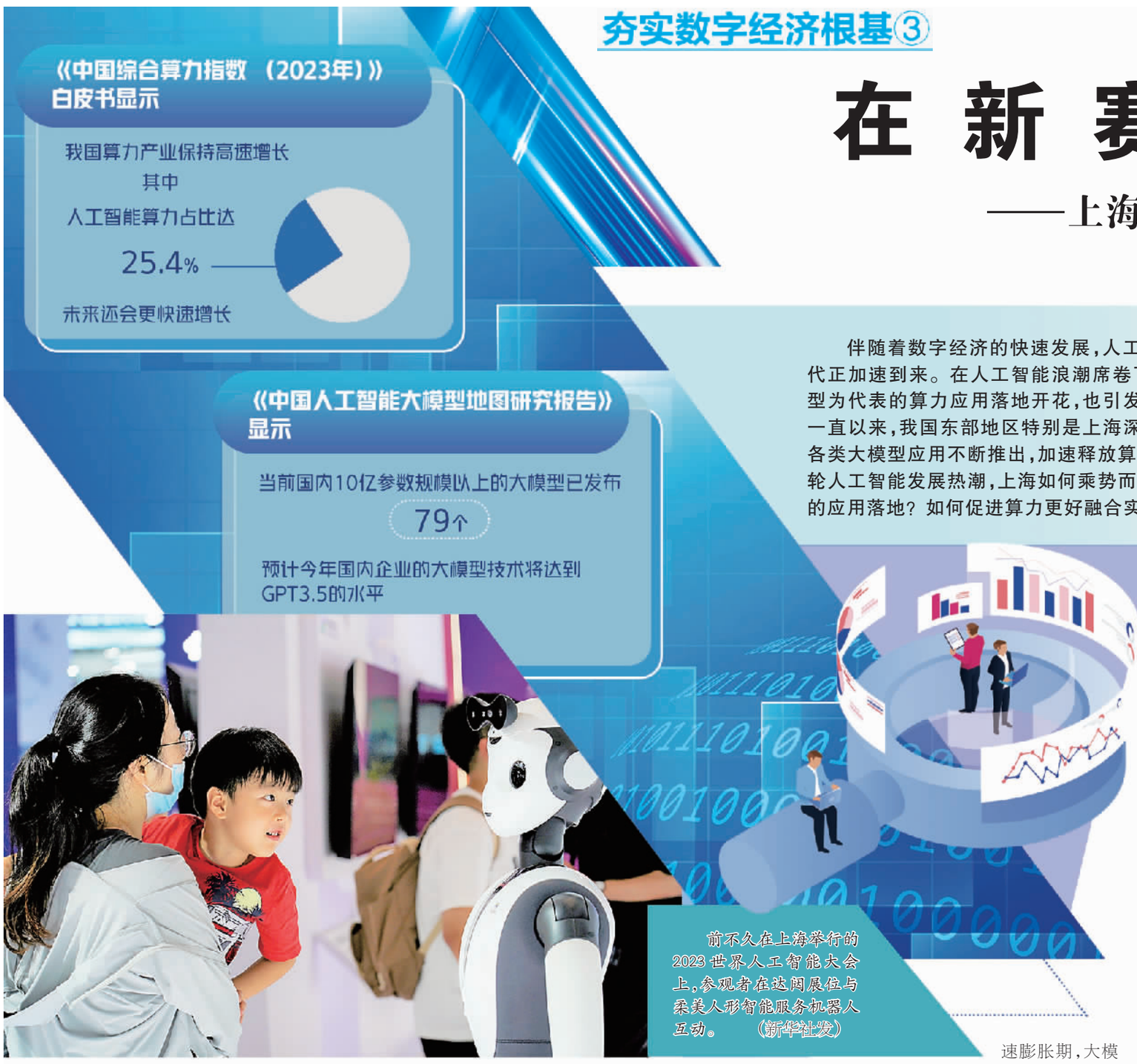
夯实数字经济根基③

在新赛道上奔跑

——上海人工智能算力应用调查

本报记者 李治国 李 景

伴随着数字经济的快速发展，人工智能技术持续突破，智能时代正加速到来。在人工智能浪潮席卷下，以生成式人工智能、大模型为代表的算力应用落地开花，也引发了算力需求的进一步增长。一直以来，我国东部地区特别是上海深入挖掘算力融合应用场景，各类大模型应用不断推出，加速释放算力资源服务潜能。面对新一轮人工智能发展热潮，上海如何乘势而上加速推进算力在更多场景的应用落地？如何促进算力更好融合实体经济？



前不久在上海举行的2023世界人工智能大会上，参观者在达闼展位与柔美人形智能服务机器人互动。（新华社发）

上海市日前印发的《立足数字经济新赛道推动数据要素产业创新发展行动方案（2023—2025年）》明确提出，建设高效协同的算力体系，建设“E级”超算载体、人工智能公共算力平台，因地制宜部署边缘计算资源池，对接“东数西算”国家战略，建设枢纽型算力调度平台，到2025年，算力总规模较“十三五”时期末翻两番。

上海迅速布局算力这一新基建，正是基于当前以大模型为代表的算力应用的落地开花。最近1年，人工智能成为全球科技产业的热门焦点领域。爆款产品ChatGPT（自然语言处理大模型）以其强大的对话输出功能，让人工智能有效辅助生产生活，人工智能产业随之向前迈进一大步，迅速成为资本青睐的“香饽饽”。

随着全球科技巨头纷纷入局，人工智能应用在大模型领域打开新局面。普遍观点认为，上游算力基础设施的持续建设、算力规模的不断扩大、数据处理能力的迅速提升，造就了下游算力应用端ChatGPT等大模型的成功崛起，让人工智能发展迎来“拐点”。

在这场全球参与的科创竞技中，我国紧跟趋势走在前沿。其中，以上海为代表的生成式人工智能（AIGC）探索将人工智能带到新高度，国家也出台了《生成式人工智能服务管理暂行办法》，及时规范AIGC的开发及应用。作为算力落地场景的最大突破口，国内AIGC的发展态势如何？应用效果如何？面临哪些挑战？记者走访了诸多业内企业与专家，探寻在算力支撑下，人工智能到底怎样“为我所用”。

大模型火爆

今年的中考、高考结束后，ChatGPT、文心一言、复旦MOSS、讯飞星火等大模型紧跟热点，纷纷下场写起作文。看了MOSS写的上海中考语文作文《会心之乐》后，上海市市南中学语文高级教师陶璐说：“我觉得MOSS总体上写得不错，但大模型没有自我意识，它怎么能真正理解‘会心之乐’呢？”上海复旦五浦汇实验学校校长、语文特级教师黄玉峰更是笑言，“如果要我打分，它肯定不及格”。

尽管人们对生成式人工智能抱有不同态度，但它已经进入我们的生产生活。

“大模型是指具有大量参数的机器学习模型，可以在训练过程中处理大规模的数据集。ChatGPT就是具有超大规模参数的大模型。要实现大模型的训练，必须有强有力的算力支撑。”天风全球前瞻产业研究院联席院长孔蓉表示，按照国内的发展速度，预计今年国内大模型可以达到GPT3.5的技术水平；得益于海外开源大模型的技术，会有一批国内企业达到这一水平。

国内大模型的涌现速度，也印证了孔蓉的判断。3月16日，百度推出搭载文心大模型的文心一言；4月8日，华为更新盘古大模型；4月10日，商汤科技推出商量SenseChat；4月11日，阿里巴巴推出通义千问；此后，360、字节跳动、科大讯飞、京东、腾讯等公司也纷纷推出了自己的大模型。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，“百模大战”并非戏言。7月份在上海举办的第六届世界人工智能大会上，国内通用型大模型顶尖产品悉数到位，集中展示国内外总计30多款大模型。

国内大模型发展火爆，离不开国内算力资源作为基础。

燧原科技创始人、董事长兼CEO赵立东表示，在大模型的技术萌芽期，训练一个GPT3参数规模的大模型成本极为高昂。根据谷歌的经验，用大模型来代替谷歌每秒32万个查询，将新增约360亿美元推理成本。此后，发展到应用加速膨胀期，大模型应用赋能千行百业，所需算力又会是一波乘数效应。

上海市集成电路行业协会会长张素心表示，“为解决算力需求问题，国产芯片应汇聚合力，扩大开发者群体，形成生态闭环，继而加速产业发展乃至国际化之路”。

东浩兰生会展集团董事长陈小宏告诉记者，在第六届世界人工智能大会上，围绕大模型的训练需求，沐曦曦思N100、瀚博SG100、昆仑芯2代AI芯片等大模型应用芯片集中亮相，夯实了国内算力资源的底气。可以说，随着大模型的火爆，算力领域既感到压力，也充满动力。

应用突围

无论是实现大模型落地应用，还是提升算力供应，都需要努力构建自主创新架构、满足市场多元需求。当前无疑是一个极佳的时间窗口。

孔蓉在美国硅谷调研发现，相比国内企业争相布局大模型，美国科技企业的研发已经以AI应用为主。在美国企业中，AI应用已相当普遍，写文章、写邮件、数据分析、发布招聘广告等都离不开AI。“国内大模型井喷之后，当务之急是应用落地。毕竟大模型研发出来就是要为生活和工作服务的。”孔蓉表示。

“历经4年技术深耕和研发迭代，百度现已升级到文心大模型3.5。”百度首席技术官、深度学习技术及应用国家工程研究中心主任王海峰表示，文心大模型3.5在效果、功能、性能等方面有了明显提升。“凡是与语言文字或程序代码打交道的应用场景，都可能有文心一言的用武之地。”王海峰表示，不少行业如能源、金融、教育等，已经成为文心一言的应用场景。

要把通用大模型应用到不同行业中，仍存在不少突破口。对此，垂直的行业大模型应运而生。这类行业模型、专属模型脱胎于通用大模型，经过有针对性的专业数据精调后，就可适用于垂直领域，为某些特定行业服务。

不久前，星环信息科技（上海）股份有限公司发布了为金融领域量身定做的大模型“无涯”。作为业界首款面向金融智能量化投研的领域大模型，它将在金融投研、量化投资和智能推理等领域有力辅助分析师、研究员和投资经理的日常工作，对股票、债券、基金、商品等各类市场事件进行复盘、传播和推演。同时，基于大模型的事件驱动与深度图引擎，其可实现对事件语义刻画、定价因子挖掘、时序编码、异构关系图卷积传播等功能，帮助企业更好应对复杂的市场环境和业务需求。

不过，目前来看，大模型应用仍存在一些问题。星环科技创始人孙元浩表示，大模型有时会一本正经“胡说八道”，原因之一在于训练大模型至少用时半年，这导致实时新闻资讯、市场行情等快速变化的信息难以内置到模型中，因此需要各类大模型发布商不断更新语料与数据，这是一个不小的挑战。

北京中关村科金技术有限公司专注于垂直行业和细分领域的大模型应用，瞄准新一代得助对话引擎，他们推出了全新的AIGC应用——“超级员工”，如智能客服、外呼机器人、智能陪练、智能质检、坐席助手等。该公司副总裁张杰介绍，“超级员工”以助手的形式在金融、零售、政务等多个行业头部企业试用，原先需要10分钟完成的营销文案，现在10秒即可完成；外呼客服话术师助手让原来30个话术师的工作量如今由2人即可完成，且语义理解准确度从85%提升至94%。“大模型具备的超强语言理解能力，让‘最后一公里’的销售过程实现数字化转型成为可能。这既能帮助企业通过智能对话服务实现降本增效，也能有效提升用户体验、拓展服务外延。”张杰说。

众多通用大模型和垂直大模型同台竞技，结果必然是优胜劣汰。张杰表示，在成本约束以及充分的市场竞争环境下，众多大模型产品必然面临“洗牌”。

面临“洗牌”的不仅是大模型，算力领域同样如此。“百模大战”可喜亦可忧，必须直面算力“烧不烧得起”的课题。大模型训练和推理最终要回归商业逻辑，实现经济效益而非“不计代价”，这就涉及算力能否像水电煤一样“普惠”。

燧原科技创始人兼COO张亚林表示，“随着大模型间的竞争逐渐展开，大模型在数量上会收敛，生态也会相应浓缩和集约，这对于建立AI芯片新生态来讲，是非常有利的机会”。

上海天数智芯半导体有限公司不久前宣布，天垓100加速卡的算力集群，基于北京智源人工智能研究院70亿参数的Aquila语言基础模型，使用代码数据进行训练，已稳定运行19天，且模型收敛效果符合预期，测试证明天数智芯已经具备支持百亿级参数大模型训练的能力。天垓100加速卡的算力集群率先完成百亿级参数大模型训练，迈出了自主通用GPU大模型应用的重要一步。这一成果证明天垓产品可以支持大模型训练，打通了国内大模型创新发展的关键“堵点”，对于我国大模型自主生态建设、产业链安全保障具有重要意义。

孔蓉说，“像ChatGPT这样级别的大模型需要上万张芯片和加速卡支持，就目前国内一些商业化应用来看，不需要比拼最高算力，而是应比拼实际效率，因此性价比会是重要的影响决策因素”。赵立东也认为，“芯片是‘用进废退’，越用才能越好用。在渐进式过程中培育算力生态、迭代算力产品，这个过程是我们必须经历的”。

提升竞争力

不久前印发的《上海市推进算力资源统一调度指导意见》提出，开展上海全市算力基础设施及算力资源输出能力摸排，形成算力清单。基于算力资源底数，推动头部企业接入上海市人工智能公共算力服务平台，构建一体化算力调度服务体系和平台基础框架，实现算力资源统一编排。

“某种程度上，算力决定了市场竞争力。”商汤科技董事长兼CEO徐立说，在AI大模型时代，模型参数量将以指数级速率提升，数据量随着多模态的引入将大规模增长，由此带来算力需求剧增。

上海市经信委副主任汤文侃表示，“十四五”期间，上海将加强全市算力资源统筹、调度和共享，提升算力资源利用率，加速数据要素流通，全面释放数据价值。

上海临港新片区6月份发布的《临港新片区加快算力产业集聚发展三年行动方案》提出，到2025年，临港新片区将形成以智算算力为主、基础算力和超算算力协同的多元算力供给体系，总算力超过5EFLOPS（FP32），AI算力占比达到80%，算力产业总体规模突破100亿元，集聚相关企业及机构超过100家，打造具有全国影响力的算力产业集聚区，建设一批算力示范应用标杆场景。

一系列算力布局，是为了夯实大模型应用的底座。随着AIGC深度应用的展开，不仅对算力、数据、算法提出了更高要求，也对安全、隐私、伦理提出更多挑战。只有在确保数据安全和隐私保护、健全人工智能伦理与安全的前提下，才能让AI技术真正释放出应用价值。

针对算力问题，华为轮值董事长胡厚崑表示，华为已在内蒙古乌兰察布市建设数据中心，初期阶段部署了数千卡规模的人工智能集群，在同等算力下，计算效率提升10%以上。

不久前，UCloud优刻得AIGC算力底座正式亮相。优刻得董事长兼CEO季昕华表示，优刻得推出涵盖数据中心、计算平台、管理平台、网络服务、应用服务、生态接口的一系列产品和解决方案，可为用户提供完全物理隔离的独享机柜、服务器、网络、存储资源，结合完整的安全方案和专家服务，确保用户的大模型平稳运行。

“历史的机遇、技术的变革，将数据智能推向了前所未有的高潮，也带来了更加严峻的数据安全挑战，数据流通迈向密态化是未来趋势。数据密态要求下，隐私计算的方法体系、平台框架、技术标准都面临全新变革。”蚂蚁集团副总裁兼首席技术安全官韦韬呼吁更多同行参与到开源和生态建设工作中。“开源隐私计算核心产品一直是我们对行业的态度，未来将进一步加大隐私计算的开放力度和广度，与行业一道构筑AI智能时代数据安全护城河。”韦韬说。

尽管挑战不少，但不可否认，以大模型为核心的人工智能时代正加速到来。“未来10年，新一轮科技周期将启动。”孔蓉认为，在AI推动下，XR、机器人、自动驾驶、影视内容等行业将进入爆发式变革时期。

腾讯研究院、同济大学、腾讯云共同发布的《人机共生——大模型时代的AI十大趋势报告》明确指出，通过建设可控、可用的安全生态，推动模型落地和应用，AI技术将为各行业带来更多机遇。大模型时代带来的创新和发展，将推动人工智能走向更广阔的未来。

正如工业和信息化部副部长徐晓兰所言，以深度学习为代表的新一代人工智能和以大模型为代表的通用人工智能不断取得技术突破，将成为智能产业的根技术和智能经济的基础设施。这意味着人工智能产业生态将酝酿一个又一个“爆点”，并等待着创业者去把握与挖掘。

调查手记

最硬的检验标准是应用

李治国

随着科技跨越式发展，人工智能领域研究不断取得突破性进展，大模型技术的研发尤为引人注目。然而，对于大模型来说，考验不在于模型的大小和复杂度，而在于其落地和服务能否经得起实践的考验。

从文心一言的应用到星环科技的测试，再到大模型支撑下的“超级员工”，这些应用案例都生动表明，大模型的生命力在于实际应用效果，只有真正解决用户需求的大模型，才拥有未来。

大模型发展需要竞争。目前，大模型研发已进入白热化阶段，国内外企业争相推出大模型产品，但只有优秀的产品才能在市场充分竞争中脱颖而出，实现优胜劣汰。大模型发展需要进化。大模型应用不应仅停留在实验室阶段，而要落地到各个产业领域，这就需要企业找准应用场景，推动大模型不断进化和发展。大模型开发需要合作。大模型研发需要大量人力、物力和财力，企业间的合作尤为重要。为避免资源浪费和低端化同质竞争，企业间要打破壁垒，携手营造合理有序的市场竞争环境，让大模型发展始终走在良性轨道上。

需要强调的是，大模型竞争也需要规则和监管。大模型应用需严格遵守相关法律法规和行业标准；监管部门应加强监管，保障模型使用中的公共利益与数据安全。

对于大模型这一新生事物来说，应用效果是最硬的检验标准。我们期待更多企业和研究机构推出更优质的大模型产品，并通过市场竞争和成本约束，助推大模型健康发展，推动AI提升生产效率，提高人们的生活品质，带来更多的现实可能和想象空间。

位于上海临港新片区的商汤人工智能计算中心。（资料图片）

本版编辑 郎 冰 闫伟奇 美 编 高 妍