大模型百花齐放 数据质量决胜负

■中国城市报记者 邢 灿

金山办公大模型WPS AI官网上线，京东推出言犀大模型，佳都科技发布知行交通大模型、华为发布大模型时代AI存储新品……连日来，关于大模型的消息让市场应接不暇。

用“百模大战”来形容当下大模型的火热程度一点也不夸张。科技部新一代人工智能发展研究中心发布的报告显示，今年上半年我国10亿参数规模以上的大模型已发布79个。有机构预测，下半年国内还将进入大模型发布的密集期。

当前，我国大模型处于何种阶段？有何发展趋势？未来发展关键是什么？需谨防哪些风险？近日，中国城市报记者进行了相关采访。

多模态大模型是大势所趋

什么是大模型？浙江大学国际联合商学院数字经济与金融创新研究中心联席主任、研究员盘和林在接受中国城市报记者采访时介绍，模型就是算法，大模型就是多个算法，通过海量数据训练和调参来实现想要达到的效果，其形成基础依然是算力、算法、数据。

“其中，数据要求高质量数据；算力需要通过算力芯片的并行计算来堆运算能力；而算法则是大模型本身，融合了非监督学习、监督学习和强化学习等机器学习算法。”盘和林说。

北京师范大学政府管理研究院副院长、产业经济研究中心主任宋向清认为，我国人工智能大模型产业在技术创新、产业生态、融合应用等方面，相较美国等发达国家，起步虽然晚，但进步较快。

在宋向清看来，由于我国前期在人工智能领域作出多项精准预判和未雨绸缪的战略部署，不仅为大模型发展奠定了坚实基础，而且推动并建立了涵盖理论方法和软硬件技术的体系化研发能力，形成了紧跟世界前沿的大模型技术群。在此前提下，我国大模型技术和产业目前均已高居全球第一梯队，进入高速发展阶段。

谈及该阶段基本特征，宋向清认为，一是已形成完整的大模型产业体系，成为新的经济增长引擎；二是技术适应性明显增强，经过系列数据训练和微调后即可广泛适应于下游任务；三是我国自主研发的AI大模型在语言、视觉、推理、人机交互等领域不断推陈出新，出现若干世界领先的新算力、新动力和新效力，已经成为国际大模型技术市场和大模型应用市场上的重要力量。

“大模型真正实现自己的价值，一定是在产业应用中。”京东集团CEO许冉表示，大模型的价值=算法×算力×数据×产业厚度的平方。“前三个指标固然重要，但技术在产业场景落地应用，创造实际价值才是关键。当产业效率和产业的边界拓展得到质的提升以后，大模型才有了更重要的实际价值和意义，这将不亚于又一次工业革命。”

“当前大模型的热点逐渐趋于冷静。”金山办公技术总监熊龙飞在接受中国城市报记者采访时表示，因为大家发现如果大模型只是拿来聊天并不会带来明显实用价值，所以如今在大模型技术上寻求的方向会越来越务实，会更多考虑将大模型能力进行应用落地。

熊龙飞认为，现在语言大模型和视觉大模型是产业界应用较广的模型，但是并未做紧密的融合，而未来多模态的大模型是一个重要的趋势。人类不仅在靠视觉感知这个世界，也在用声音、触觉、文字等媒介感知世界，更多维度的信息会让大家对世界的感知和认知更全面，因此更多模态的关联感知会提高人工智能技术的能力上限。

以大模型WPS AI为例，熊龙飞介绍，目前WPS AI接入了金山办公旗下常见的办公组件，比如WPS文字、表格、演示PPT、PDF、拍照扫描等等，并且加速产品智能化发展。在此基础上，金山办公又推出了WPS智能文档、智能表格、智能表单，有了AI的赋能，办公软件操作起来会更加便捷，原先许多复杂的功能的使用门槛也将明显降低。

“就像软件行业常说‘二八定律’，80%的人只用了20%的功能，而办公软件因为其复杂性，用户往往需要一定的学习才能熟练使用，这是办公行业不能避免的问题和痛点。但有了AI，就能彻底改变软件的二八定律，软件将能理解用户的需求和意图，用户不需要再去学习复杂的软件使用方法，每个用户将用到更多的功能，也会让移动APP有更大的可操作性。”熊龙飞举例说。

数据质量或是决定关键

“我国大模型面临的门槛和困境很多。”盘和林认为，其一是我国高质量数据不多，各大互联网企业相互封闭自身生态，当前数据开放度不够，且由于各个平台对数据内容存在选择性，导致很多有价值数据无法形成，数据瓶颈是我国发展生成式AI最大的瓶颈。

河南省商业经济学会副秘书长胡钰表达了相似的观点：当前我国大模型面临多个难点。其中，数据生态存在先天不足，如在互联网内容资料中，中文数据不足2%，而且质量参差不齐。

宋向清认为，大模型发展首要突破的是大数据、大算力技术瓶颈，以及为大数据存储和大算力提升提供基础支撑的高科技智能化装备。“中文数据生态严重滞后，中文数据占比太少且质量参差不齐，这不利于中国大模型技术和中华文化在全球市场上的同步推广。”宋向清说。

胡钰在接受中国城市报记者采访时建议，提升互联网公开资料中文数据生态，数量和质量共同提升、为大模型发展提供优质数据。

值得一提的是，日前，大模型WPS AI在回复中国城市报记者关于“当前大模型面临哪些门槛或困境”这一问题时也提到“数据质量”问题。回复中提及，大模型的训练需要大量的数据，但是数据质量参差不齐，存在噪声、偏见等问题，这会影响大模型的性能。

熊龙飞介绍，目前大模型的训练门槛和成本还是非常高的，需要投入海量的GPU训练资源和大量人力做数据清洗和模型调优。而大模型本身的能力覆盖也非常有限，单靠大模型本身无法完成复杂的任务，尤其是在办公领域，很多任务不是简单的写几百字的文本，而是要实质性的解决用户办公场景的需求。

大模型不能成“脱缰野马”

新技术总会伴随新风险，大模型亦不例外。宋向清认为，大模型广泛应用的风险主要还是法律法规不能快速跟进调整而形成的应用漏洞。

“比如个人和单位数据隐私泄露风险，不法分子通过大模型进行网络钓鱼、欺诈等行为导致的恶意攻击风险；因基于不同大数据而形成的大模型算法偏见可能形成的公平性和公正性风险；因数据质量、算法设计等因素而形成的可能的不合理、不可靠或不准确误导性风险等。”宋向清说。

宋向清建议，强化立法进程、加快立法进度，通过立法加强数据保护和隐私保护，确保用户的数据不被滥用或泄露；同时，加大利用大模型违法犯罪的打击力度，加强安全防御和监测，及时发现和严厉打击来自大模型的各种攻击和威胁；从企业自身角度看，还必须强化算法设计和评估，确保算法的合理性、公正性和可靠性。

盘和林认为，大模型风险包括隐私和数据安全风险、内容欺诈风险、伦理风险。其中，伦理风险最小，不宜对AI替代人的作用过分夸大，AI并不具备这么高的智慧；而隐私和数据风险和内容欺诈风险则需要给予关注，当前最理想的状况是要求AI生成内容发布的时候给予来源标注。

记者注意到，将人工智能作为三大先导产业的上海，已在考虑规范发展问题。上海市经信委副主任阮力表示，上海市将人工智能作为三大先导产业之一，不仅积极打造具有国际影响力的世界级产业集群，同时坚持科技向善理念，持续推进人工智能健康规范发展。

阮力提到，上海将加强大模型风险管理，启动高风险人工智能产品和服务清单式管理相关研究，积极推动大模型相关治理研究，探索适应大模型特点的创新监管方式。

江西东乡：企业生产忙 夏日经济旺

7月12日，在位于江西东乡经济开发区的江西养乐星生物科技有限公司生产车间里，工人正在赶制果冻订单。

盛夏时节，该区各类夏日引用品企业开足马力忙生产，满足消费者订单需求，带火带旺夏日经济。

人民图片