科学大模型：“上线”之路还有多远

深瞳工作室出品
采写：本报记者 孙明源 华 凌 徐庆群
策划：赵英淑 滕继濮

只需输入一段文字，电脑便会将其转化成栩栩如生的画面；只需问一句“哪里的饭好吃”，导航软件就能带你吃遍当地风味……大模型通常具有高度的通用性和广泛的适用性，已经在自然语言处理、图像识别和语音识别等众多领域大放异彩。

然而，这仅是人工智能大模型应用的“冰山一角”。尤其是在科研领域，其无限潜能还有待深入挖掘。

2024年12月7日，地球科学领域垂直大模型——“元古大模型”在中国地质大学（武汉）发布，可对古生物化石进行复原。去年9月，在北京发布的全球首个多模态地理科学大模型“坤元”因具备处理地理科学相关问题的专业能力，被称为“智能地理学家”。

“科学大模型或许可以引发科研范式和方法上的革命。”北京智源人工智能研究院（以下简称“智源研究院”）院长王仲远告诉记者，科学大模型作为一种新兴工具，目前尚未在高校、科研院所以及企业进行大范围应用，除了技术层面的原因，其推广还面临诸多挑战。

赋能科学研究

大模型参与科研活动的基本原理是什么？用一个词来回答，就是“模拟”。正如语言大模型可以模拟语言文本信息一样，科学大模型旨在模拟复杂的科学现象。

中国空气动力学研究与发展中心研究员钱炜祺介绍，广义的大模型，是指具有大量参数和复杂结构的深度学习模型。参考目前业界主流观点，可将大模型分为大语言模型、视觉大模型和科学大模型。

其中，科学大模型主要处理和分析数值、科学领域数据，对其理解物理规律和知识生成的能力。“相比语言和视觉信息，科学数据通常具有超高维度、非线性、强空间差异性等特点，为此大模型需要理解的物理现象极其复杂。相较于大语言模型和计算机视觉大模型，科学大模型发展的成熟度相对较低。”钱炜祺说。

尽管研发难度高、挑战大，科学大模型目前已经取得了许多突破，并且已用于科研实践，在药物研发、材料科学、分子模拟、天气预报、流场预测等领域发挥作用。

2024年12月1日，福布斯中国与全球商业研究院联合发布“2024中国新时代颠覆力创始人评选”名单。北京分子之心科技有限公司（以下简称“分子之心”）创始人兼首席科学家许锦波入围。

“现在，人工智能已经改变了分子生物学的研究范式。过去研究者要基于氨基酸序列来研究蛋白质功能，现在可以直接基于人工智能预测出的结构进行功能研究。”许锦波说，“我们还在运用蛋白质生成大模型进行精准的蛋白质优化与设计，这颠覆了过去生物医药、生物制造等产业领域的蛋白质发现与改造方式。”

用模型解开万物之奥秘，正是大模型辅助科研的最大优势。在其近年来的应用方向当中，生物计算领域成果颇丰。

智源研究院于2024年6月推出的“全家桶”中，就包括生物计算大模型。该模型搭建了全球首个数字孪生心脏电功能超实时仿真系统，包含了19种细胞生理状态变量和70多个公式，能够实现复杂的心脏电生理与病理的仿真。

除了生物计算，科学大模型也在其他领域得到应用。百度深度学习技术平台部架构师胡晓光告诉记者，当前大模型与智能体已经在科学计算领域得到应用。例如，中国科学院自动化研究所依托百度的“飞桨”和“文心”大模型，研制出材料学科科研智能体。一些前沿实验室正在采用大语言模型，进行材料属性预测和结构生成。

华为轮值董事长胡厚崑认为，人工智能将数学计算和科学模型的方法结合，可以高效处理海量数据，解决原来传统科学研究范式无法解决的问题，帮助科研工作者突破科研瓶颈。

“科学大模型拥有非常大的潜力。”王仲远说，“目前人工智能大模型作为新兴工具整体上还处于起步阶段，但一些大模型已经在包括科研在内的许多领域发挥了作用。”

面临诸多挑战

在王仲远看来，过去10余年间，人工智能技术的几次重大突破，并非单纯算法层面的研究突破，其本质是一个数据、算力、算法、评测等多团队高度协同的算法类系统性工程的落地。

在人工智能领域，特别是在科学大模型的研发上，要想取得突破性的创新，需要庞大且复杂的团队作战与协同，大量集中的资源投入以及技术路线的研究探索与试错，单凭一所高校或者一家企业很难做到。

“例如，作为一个研发周期长、成本高的行业，生物制药比较依赖已有的研发模式。如果引入新工具，需要慎重考虑其对成本、风险以及对收益分配的影响。”王仲远说，再如教育领域，特别是在与未成年人相关的应用场景，应用新技术需要社会各方的审慎思考，这涉及许多细微复杂的问题。

西安电子科技大学电子工程学院教授、情感机器（北京）科技有限公司首席科学家吴家骥注意到，在高校的科研环境中，科学大模型的应用也面临诸多挑战。

“那些简单的、可用公式表达的科学问题，基本都被解决了。目前科学问题公式的复杂程度，已经超越了人类理解能力的极限。”吴家骥表示，科学大模型的工作极具交叉性，从提出好公式到设计出好的训练系统，从传统科研实验流程到数据驱动的AI实验范式，从找到好答案到提出好问题，这些都对传统认知提出挑战。

高端人才匮乏

“高校和企业，各有各的难处。例如，高校受资源和机制所限，企业则背负着营收压力，导致它们在大模型应用和研发方面，有时会‘伸不开手脚’。”王仲远说，除了制度和资源，科学大模型在科研环境中落地最需要的基础条件是人才。

2022年，许锦波在北京创立分子之心，很快聚集了一批顶尖复合型人才。这些成员兼具AI蛋白质研究和产业实践的经验，其中核心研发团队博士占比90%以上。

“但从整个科学大模型领域来看，复合型人才非常稀缺。”许锦波告诉记者，以蛋白质生成大模型为例，除了必备的算法、算力、数据等基础条件外，应用此类大模型还需要具备两大专业能力。一是融合计算机、生物、物理等多学科，熟识人工智能、分子动力学、量子计算等多种方法，且能在实践中并行考虑序列与结构、主链与侧链、进化与组学的跨领域融合能力；二是走出实验室，下沉至真实产业环境，在需求、验证、落地上贴近产业需求的能力。

北京社会科学院副研究员王鹏此前接受采访时表示，人工智能技术发展日新月异，要求从业人员具备不断更新知识体系和技能储备，这对人才培养提出了更高要求。

人才短缺不仅限制了人工智能技术的创新和发展速度，也影响了相关企业在市场中的竞争力，但这也为有志于投身人工智能领域的人才提供了广阔的发展空间和良好的职业前景。

亟待多方发力

如钱炜祺所说，科学大模型在几类大模型当中研发门槛最高，如何持续提升科学大模型的质量，并推动其应用落地？

钱炜祺以空气动力学领域大模型为例，该领域目前已有成果大多借鉴了计算机领域通用技术，未来还需探索发展适用本领域的模型架构。

空气动力学数据具有样本少、规模大、模态多、获取成本高等特点。要想做好相关的大模型，就必须基于领域特点进行技术攻关。例如，可围绕空气动力学相关基础理论和人工智能领域发展迁移学习、小样本学习和多模态学习等模型算法，解决数据不足、学科贴合度不高的问题。

钱炜祺提醒，大语言模型、计算机视觉和科学大模型并不是像“烟囱”一样各自独立发展的。它们相互之间已实现关联、调用、融合，可以共同解决特定场景、特定领域问题。因此，科学大模型的研发和推广并不局限于自身，人们应该关注大模型技术的整体发展。

钱炜祺预测，随着数据不断丰富、算力提升、算法改进，空气动力学领域大模型将改变信息分发和获取模式，革新数据和知识生产模式，实现全自动交互完成任务目标，成为科研工作的“加速器”。

许锦波说，除了技术本身，大模型的发展也和人才和制度息息相关。

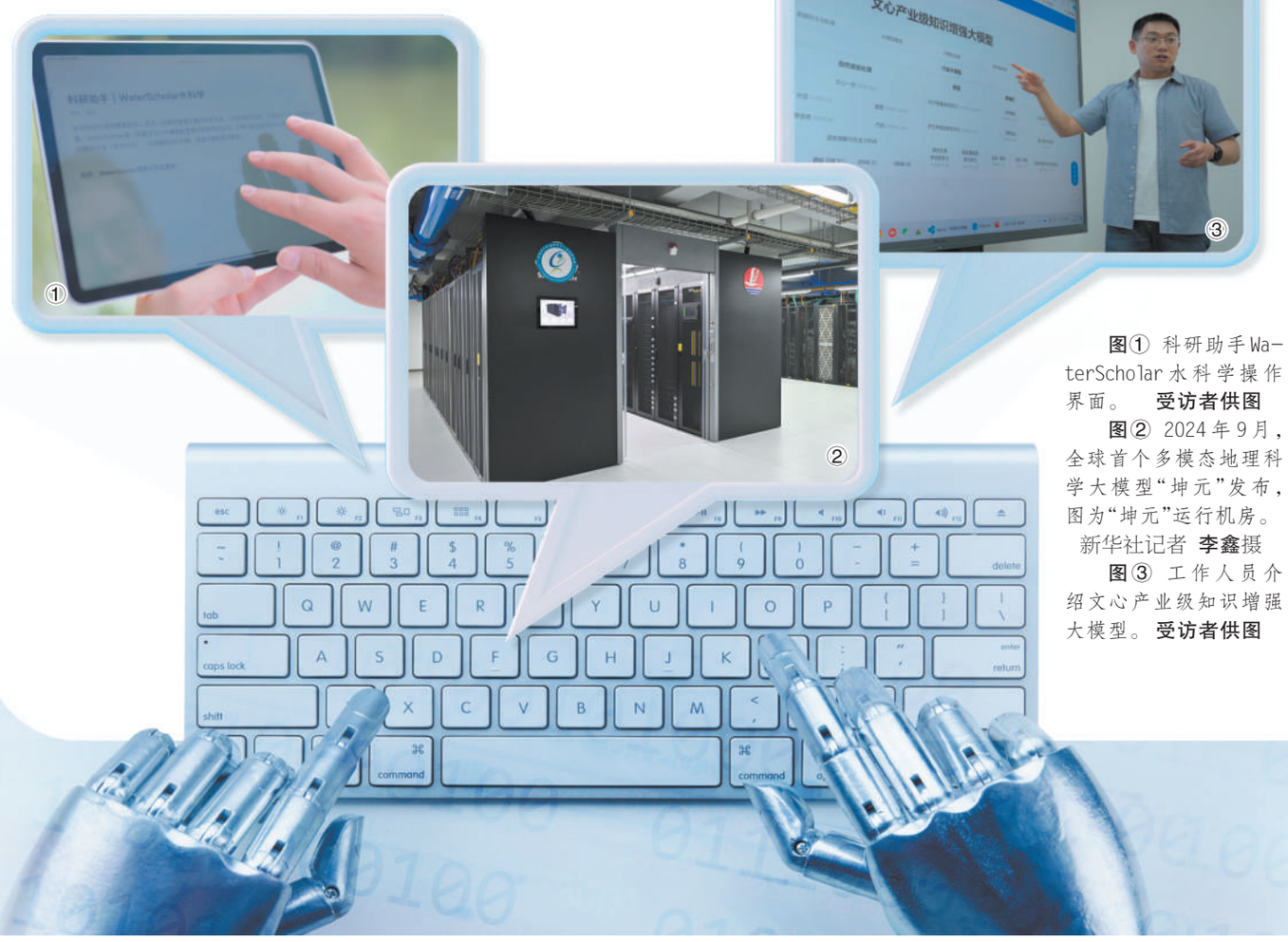
许锦波认为，培养兼具科研和产业能力的创新者，关键在于紧密贴合产业实际需求，全力促成跨领域协同创新。同时，一支汇聚多学科知识背景、兼具产业实操本领与科研攻坚实力的复合型人才团队，是持续创新的源泉。

“我们在做的事情既需要‘从0到1’研究和解决科学问题，也需要将技术落地于产业实践。我们需要懂计算、懂人工智能、懂生物科学的复合型人才。”许锦波表示。

面对人才瓶颈，胡晓光认为，打造开放的科研生态，降低大模型应用门槛是关键。

百度发起的“飞桨AI for Science”共创计划，通过提供算力支持、资源与服务，共同推进AI技术在科学计算领域的创新与发展。截至2024年末，“飞桨”产业级深度学习开源开放平台已在服务43万企事业单位，创建模型超100万个。

胡晓光介绍，“飞桨”通过由参与单位和个人共同建设模型库和场景范例、提供免费算力、为优秀科研方案和重点项目提供框架、模型资金支持、开发套件以及推出全方位课程资源技术合作支持等方式，和科研人员一起开展科研工作、研制前沿模型、建设场景范例、取得科研成果。“科学大模型的开发、落地和推广，需要大量的跨领域科研人才，并且实现人工智能与传统科学计算工具链的协同。这需要我们搭建稳定、优质的科研生态，把资源和机会凝聚起来，共同打破目前遇到的瓶颈。”胡晓光说。



图① 科研助手WaterScholar水科学操作界面。 **受访者供图**

图② 2024年9月，全球首个多模态地理科学大模型“坤元”发布，图为“坤元”运行机房。新华社记者 **李鑫**摄

图③ 工作人员介绍文心产业级知识增强大模型。 **受访者供图**

延伸阅读

为大模型研发营造创新生态

◎本报记者 孙明源 华 凌

如何营造科学大模型创新生态，以确保高效的科研产出？

“不以论文论英雄。”智源研究院代表性的创新做法之一，就是摒弃了“以论文论英雄”的传统考核标准，转而以科研成果在学界和产业界产生的实际影响作为评判依据。

由科技部和北京市支持创办的北京智源人工智能研究院（以下简称“智源研究院”）成立于2018年，是我国最早开始做大模型研发的科研机构之一。“在北京市的支持下，我们还突破了传统的申请—答辩制度，采取了包干制，由科研人员自己决定做什么项目。”王仲远介绍。

当前，智源研究院正在积极探索新型研发机构建设模式创新，建立了“青年人才挑大梁”的人才评价及培养机制，打造“代表作文化”，通过“小同行评议”，遴选拥有学术代表作的一流人才；在“有组织科研”机制创新上，探索“集中力量办大事”的跨机构、跨领域、大团队的新型科研组织机制。

“此外，我们很重要的一个理念，就是接纳失败。”王仲远说，“创新必然面对风险，失败也有重要的价值。我们会组织专门的顾问委员会、技术委员会去评估研究成果，分析失败的工作是否还有进行的可能，以及给了我们哪些技术路线上的启示。”

智源研究院不仅构建起一套开放的内部生态，也打造了一套面向整个科研系统的外部生态。王仲远介绍，智源研究院的许多数据、模型都是开源的。同时，智源研究院积极搭建学术界和企业界的桥梁，并努力邀请全世界科学家参与到中国的人工智能科学生态当中，这些做法不仅在我国，在世界范围来看都是比较罕见的。

除了开放数据集和大模型，智源研究院还在持续完善覆盖模型、数据、算法、评测、系统的大模型全栈开源技术基座，并打造面向大模型、支持多种异构算力的智算集群软件栈，为整个行业提供支持。

王仲远认为，科学研究的探索之路往往曲折而漫长，作为学者的后盾，研究机构需要给予自由度，以鼓励创新和促进进步，但同时也需要进行一定的引导，避免“迷失方向”。

“我们要做的就是通过协作，打造用于未来的研究平台，集结最优秀的同行，专注可能产生原始创新与长期影响的领域，让创新系统更高效地运行，通过我们的努力让中国出现突破性成果的概率增加。”王仲远说。

本版图片除标注外由视觉中国提供