

从“描述世界”到“理解科学”——探一探，大科学模型的“科学脑”

眼下,以大模型为代表的人工智能技术,引发了新一轮科技发展浪潮。其中,与DeepSeek等擅长处理日常文字类任务的通用大模型不同,大科学模型瞄准处理各个科学领域的种种专业问题,正在引发科研范式和方法的革命。

■ 本报记者 何冬健
通讯员 肖乐 盛汪淼苻

当你想让人工智能写份工作总结,它大概率能轻松搞定;可要是问它“蛋白质折叠的最低能量状态怎么算”,多数通用人工智能就只能支支吾吾了。

这就是当下人工智能领域的现实:通用大模型擅长处理日常文字类任务,像写讲话稿、整理会议纪要这类活儿,对它们来说不在话下。但面对专业的科学问题,它们就显得力不从心了。

与之不同的是,在之江实验室,由100多名计算专家组成的研究团队正聚力打造的科学基础模型,却瞄准了这些“硬骨头”。它不仅能处理公式、分子结构、基因序列等复杂数据,更能像领域专家一样,循着科学逻辑找到答案。比如研究黑洞时,它会整合光学观测数据一步步推导可能的位置;探索疾病奥秘时,能从蛋白序列一路追溯到基因缺陷。

“这让人工智能模型从‘描述世界’跃升至真正‘理解科学’的崭新境界。”之江实验室相关科学家介绍。而这正是科学基础模型与通用模型最核心的区别。

不止“写文章”，更会“科学推理”

今年年初,名为Humanity’s Last Exam(人类最后一次考试)的大模型测试项目诞生。其包含2500个问题,由数百位领域专家开发,用于追寻人类知识推理的边界。这场“测试”显示,那些看似无所不能的通用大模型实际准确率最高的仅有25.4%(数据库更新时间为2025年4月3日)。

2020年,人工智能系统AlphaFold横空出世。在此之前,解析一个蛋白质的折叠结构堪称科研界的“马拉松”。之江实验室科学模型总体部技术总师薛贵荣亲眼见过同事在超算中心守了3个月,盯着屏幕上原子运动轨迹一点点收敛,最后得出的结构还得靠冷冻电镜反复验证。这个专业模型把整个过程压缩到了分钟级。薛贵荣突然意识到:人工智能不仅能“做题目”,还能像科学家一样“搞推理”。

STEM教育是美国首先提出的人才培养计划,旨在提升学生的科学(Science)、技术(Technology)、工程(Engineering)和数学(Mathematics)能力,并迅速成为知识经济时代人才培养的重要标准。

薛贵荣认为,科学基础模型的出现,将颠覆STEM模式。“伽利略有一个著名的论断——宇宙是一本用数学语言写成的‘宏伟之作’。在STEM中,‘M’其实是支撑‘STE’发展的重要工具。人工智能时代,我们认为数学、人工智能将共同推动‘STE’这三门学科的科研范式变革。”

过去,科学家80%的时间耗在查文献、算数据、分析推理上,只剩20%的精力思考核心问题;现在,模型可以接手这些繁琐工作。

中国科学院院士陈润生在接受媒体采访时,感触尤深,这位参与过中国人类基因组研究的科学家,将投身大模型研究视为自己人生第三次重要抉择:我的人生至此经历了三次抉择:第一次是给沃森写信,相信人类一定要破译遗传密码,此后参加了中国的人类基因组研究;第二次是对人类基因组序列组装后发现,编码蛋白质的基因组序列只占人类基因组的一小部分(不超过5%),从而率先开展了非编码研究;第三次,就是我最近参与学习、研究和推动的大模型。

尽管研发难度高,挑战大科学基础模型目前已经取得许多突破,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。

日前,《北京市加快人工智能赋能科学研究高质量发展行动计划(2025—2027年)》正式发布,提出到2027年建成科学基础大模型。在浙江,科学基础模型作为重点,同样已提上日程。

如何从零开始打造一个前所未有的科学基础模型?薛贵荣最初也找不到突破口。但有一个问题显而易见:作

为为科学家打造的模型,科学家到底需要什么?

团队作了一个决定,先从领域模型开始着手。“通过测试我们发现,通用模型在回答专业领域的问题时,表现并不好,无论是准确性,还是深度和广度。”于是,基于目前主流的几款通用大语言模型,团队打造了地学领域模型GeoGPT和天文领域模型OneAs-tromony。

在打造GeoGPT的过程中,研发团队与美国普渡大学詹姆斯·奥格教授团队合作,共同攻克了全球权威巨著《无脊椎动物专著》的数据化难题。这部涵盖50卷、10万化石属的“数据金矿”,因纸质载体限制和复杂数据结构长期无法被有效利用。研究团队创新性地提出了“AI批量抽取+专家验证+模型迭代”的协同技术路线:GeoGPT团队负责研发核心的自动化数据抽取引擎,实现了化石属名、地质年代、生物地理分布等关键信息的结构化转换;奥格教授团队则主导数据标准制定、结果验证及可交互数据库的构建。

团队仅用4个月便完成了3卷数千万化石属的精准提取,时间成本降低75%。“GeoGPT将过去视为‘不可能’的化石大数据工程变为现实。”奥格教授摩挲着屏幕上自动生成的生物演化树,眼里闪着兴奋的光。基于当前成果,团队正规划将技术拓展至生物演化树等更高维度的图表数据挖掘领域,进一步推动古生物学研究的数字化转型。

在与领域科学家的不断交流碰撞中,需求不再脱钩,团队也锤炼出了将科学数据token化(即将连续文本、图像等数据分割到具有独立语义或功能的最小处理单元的过程)的能力。构建科学基础模型,似乎有了眉目。

从地基开始“造房子”

从领域模型到基础模型,科学数据是关键。

“科学模型本质上还是三件事,算力、token和深度学习架构,我们熟悉的AlphaFold、ChatGPT都是在这个框架下训练出来,还有很多人在用这样一个框架做很多其他类似的模型,比如说天气预报、基因模型、蛋白质模型。”薛贵荣称,科学不仅是用语言来表示的,还有很多是用公式、分子式、图像等等来表示的。之江实验室要做的就是,把所有这些科学数据全部token化,在一个模型框架里面训练。

“现在的AlphaFold已经到了第三代,把token化对象从蛋白质扩展到了DNA、RNA、小分子配体、抗体,尽管它们的表达都不一样,但将它们to-ken化之后,就可以放在同一个空间下去训练,所以AlphaFold不仅仅能够高精度地预测蛋白质结构,同时也可以预测其他生物分子形成的复杂结构。”薛贵荣表示,“这个时候模型的魅力就出来了。”

把基因序列、蛋白质结构、天文数据、材料分子结构等科学知识都变成统一的“数据语言”,放进一个模型里。这个看起来略显疯狂的想法,让这群计算专家度过了无数通宵达旦的日子,实验室的科学基础模型021 Large Sci-



之江实验室科学模型成果亮相2025年人工智能造福人类全球峰会。

本版图片均由之江实验室提供

ence Model(简称“021 LSM”)逐步有了雏形。

融入天文、地学、数学、物理、化学、生物等多学科的科学数据,科学基础模型相当于从地基开始“造房子”。

2024年,之江实验室与国家科技图书文献中心(NSTL)签订了“共建科技语料库战略合作框架协议”。基于该中心文献,团队构建了科学文献语料处理流水线,表格、图像及学科标签均被精准解析,确保“喂”给模型的都是科研精华。

2025年初,021 LSM正式启动覆盖174个学科的科学语料生产计划,由具备相关学科专业背景的科研人员对34万个网站开展系统性质量评估,确保语料在准确性、专业性等方面达到高标准要求。

在GeoGPT的研究中,为了让模型“学透”地学,团队与国内外25家机构、400余位地学专家共建全球最专业的地学领域数据集,覆盖8个地学二级学科。与此同时,之江实验室还通过与中国科学院国家天文台共同举办大模型种子班等方式,推动天文专业数据和人工智能算法深度融合,进一步提升模型语料的科学密度。

021 LSM没有在现有通用基础模型上“搭房子”,而是“从0到1”打地基“建房子”。这极具挑战。

“基础模型决定了模型能级的上限。”薛贵荣感叹,“就好比1升的瓶子装不下3升的水,在他人的通用模型的框架里做研究就很难有所突破。”

目前,021 LSM模型训练流程每个阶段都可能耗时数月。让人工智能模型从“描述世界”跃升至真正“理解科学”的崭新境界,团队正跋涉在一条前人未至之路上。

让世界换一种玩法

今年6月,021 LSM作为一名浙江的虚拟考生参与了全国高考。满分150的全国新高考I卷,它能考143分。它还有“成长空间”——针对几何类的题目,人类使用视觉的辅助线可以高效快速地解决问题,但是若使用纯文字的解法需要更多的逻辑推理。

8月,021 LSM亮相2025年人工智能向善全球峰会,吸引了全球参观者的目光,大家对这个“理科生”模型感到好奇——它与通用大语言模型究竟有什么不同,非英语母语者能否用它来解决科研问题?

了解科学家的共性、加深对科学问题的理解、覆盖更系统全面的科学领域知识……“还有太多的科学问题等着我们去发现和解决。”薛贵荣感慨。

中国工程院院士、之江实验室主任王坚曾说:“基础模型是人工智能的皇冠,是人工智能发展的技术底座。”基础模型突破带来的将是巨大的想象空间。

“或许以后能发一个火箭派一堆机器人去外太空做实验。”薛贵荣畅想,“只要有电的地方,人工智能就能工作,地底下也是一样。机器人就能够完成一系列的科学实验。而人作为指挥者,要思考的是未来的问题。”

这并非空想。王坚近期在接受媒体采访时分享了一位美国高中生发现未知天体的故事:马特奥·帕兹是一名18岁的美国高中生,他借助人工智能技术,在美国航空航天局积累的2000亿条观测数据中,发现了150万个此前未被发现的天体。那些藏在数据洪流里的异常信号,曾让专业团队望而却步,却被模型敏锐捕捉——这恰是科学基础模型的神奇之处:它能让人类突破专业壁垒与认知边界,实现从前不敢想象的科学发现,即便发现者并非该领域的科研工作者的,也能叩开未知世界的大门。

现在,之江实验室太空计算星座已经能让人工智能直接在轨处理遥感数据,不用把图片传回地球。深地探测、新药研发……未来,这些曾依赖“人海战术”的领域,或许会因为科学基础模型的加入而加速突破。而当科学基础模型深度嵌入材料、生物等产业创新链条,高效调用、深度分析海量科学知识和数据背后的联系,或许将从根本上重塑产业发展格局。

就像望远镜延伸了人类的视野、显微镜扩大了人类的微观世界,科学基础模型正在延伸人类的“思维”。它不是要取代科学家,而是要成为那个“最懂科学家”的搭档——帮你算公式、画图纸、找规律,让人类能更自由地探索那些“想都不敢想”的未知。

就像薛贵荣在采访的尾声,俏皮的下一下:“让世界换一种玩法,还挺有趣的。”

链接

王坚和他的计算范式

■ 本报记者 何冬健 通讯员 肖乐

近日,穿着一件灰白格子衬衫的“码农”王坚,站上了联合国的2025年人工智能造福人类全球峰会的发言席。随即,掌声响起。

这是发生于瑞士日内瓦的一幕。台下,来自世界各地的人工智能科学家们身着休闲西装或格子衬衫,聚精会神;台上,中国工程院院士、之江实验室主任王坚,将实验室研发科学模型与三体计算星座,践行开放科学的故事娓娓道来。

人工智能的“之江实践”何以登上联合国讲台?

王坚说,当前我们进入了“计算密集、数据驱动、基于模型”的第三范式,可以称之为“计算范式”。

在之江实验室,“计算范式”正成为新事物的“加速器”。

地学大模型GeoGPT,初始于云栖工程院,由之江实验室主导打造,受深时数字地球国际大科学计划使命愿景启发。它包含了深度检索和分析理解、文献阅读和数据抽取、地质图问答和生成、知识图谱提取与构建等一系列科研工作,以及智能体开发平台。

在王坚的讲述中,该大模型通过人工智能技术,将化石孢子属种数量从68个扩展到3878个,实现了前所未有的数量级飞跃。英国地质调查局前首席科学家迈克·斯蒂芬森认为,“GeoGPT在资源获取困难地区具有特殊价值”。截至目前,GeoGPT已吸引来自135个国家的4万余名地球科学研究者注册使用,帮助该领域众多科研人员提升了研究效率。

“没有人工智能技术,科学家可能一生都无法实现某些发现。”王坚补充说,“不要把人人工智能当成一个简单的工具,它不仅改变科学研究的方法,甚至改变你提问题的方法。”

“计算范式”更是一种思考方式的改变——每周一、三、五,之江实验室6号楼的会议室里,大模型“种子班”的青年科学家们聚在一起,开展头脑风暴。在之江实验室,“干好智能计算一件事”已成为共识。若你偶遇一位行色匆匆的科研人员,并出于好奇提出任何关于人工智能的疑问,他定会停下脚步,热情洋溢地向你比划上半天。

这群科学家怀揣着一个更为宏大的梦想:从仅仅“描述世界”跃升至真正“理解科学”的崭新境界。随着人工智能不断向深层认知范式迈进,之江实验室在开拓GeoGPT等前沿领域模型的同时,也在全力聚焦、深耕科学基础模型领域,打造“021”大科学模型。

“所有早期的先驱者都在思考,人工智能和人类智能之间的关系是什么?在我看来,人工智能并不直接对应



王坚在2025年人工智能造福人类全球峰会发言。

人的智能,作为一种技术,人工智能帮助拓展的是人类创造力。”王坚说。

太空中的算力与人工智能,又将这一想法延伸到了更广阔的宇宙范围。今年杭州“2050大会”的探索空间,有一个主题展区叫做“人类距离太空只有100公里”。王坚觉得总提“人类”略显沉重,无端地为这一触手可及的事物增添了距离感。他更倾向于以一种轻松愉悦的方式表达,比如“你的家园,距离太空不过100公里”。

三体计算星座,是他为世界各国科学家带来的另一个借助“计算范式”突破的案例。今年5月14日,由之江实验室牵头研发的太空计算星座成功发射,12颗计算卫星成功入轨,是实现算力上天、在轨组网、模型上天的太空计算的第一步。

演讲中的王坚兴致勃勃:“为了探索距离地球约1.5亿公里的日地第五拉格朗日点,我们亟需太空计算技术的强有力支撑。”通过构建三体计算星座,能够实现跨卫星编队的数据转发。

“计算范式”,既是一种可资借鉴的案例,又是一条前行的路径。在联合国的2025年人工智能造福人类全球峰会上,《人工智能向善创新实践案例集》发布,之江实验室两项科学模型成果GeoGPT及OneAstronomy收录其中,展示了大模型等人工智能技术在地学科学研究、海洋生态监测与保护以及宇宙探索中的应用效果。GeoGPT从全球200多个评选案例中脱颖而出,获评国际电信联盟颁发的优秀创新实践案例奖。

之江实验室正聚焦智能计算,通过算力、数据、模型体系化的创新,为推动“人工智能+科学”提供核心动力,并期待与全球科学家携手,以人工智能延展人的创造力,解决人类社会面临的共性挑战。