人工智能辅助科研要从可用走向可信

热点透视 rediantoushi

对于科研工作者来说，检索、阅读文献是一项费时费力的工作。在大模型发展如火如荼的今天，以其为代表的人工智能正渗透进人们工作生活的各个角落，科研领域也不例外。

日前，阿里巴巴发布了基于Transformer架构自主研发的千亿参数级夸克大模型。据介绍，该大模型可用于科研资料收集、文献快速阅读与翻译、创作润色等场景。

不仅是阿里巴巴，科大讯飞股份有限公司（以下简称科大讯飞）、腾讯等企业，也都推出了用于辅助科研的大模型产品。这一系列产品的问世，正悄然改变着科研工作者的工作方式。

大模型已进入科研领域

今年初，ChatGPT的走红掀起了语言大模型热潮。人们可以随心所欲地提出问题，大模型总会给出答案。这股风很快也吹到了科研领域。ChatGPT发布后不久，一款名为txyz.ai的应用插件在科研圈中受到追捧。

这是一款借助ChatGPT的强大理解能力，专门用来阅读科研文献的插件。用户可以直接将论文全文上传至该应用，并提出相应解读要求，它便能够以最快速度对用户提出的问题予以回答。

即使没有下载论文全文也没关系，txyz.ai支持对论文预印本网站进行检索。用户可以只提供一个论文ID序号，txyz.ai就会自动检索、学习该论文，并根据用户需求给出回答。不仅如此，用户还能以聊天的方式与其进行对话，就论文中的内容提出各种问题。

视频网站哔哩哔哩知名科普博主严伯钧是txyz.ai的忠实用户，他时常在各类科普视频中使用txyz.ai来协助解读论文。在他看来，txyz.ai给出的论文解读准确率已经非常高，具备很强的实际应用价值，可以帮助科研工作者更加高效地检索、阅读文献。

“txyz.ai无法解读的情况当然也会有。”严伯钧表示，以他的使用经验来看，向txyz.ai提出的问题必须是一个能被回答的“有效问题”，“如果问题问得太细、太深，或者过于刁钻古怪，那么它就会直接告诉你，无法回答”。

但必须承认的是，在大模型迅猛发展并逐渐进入千行百业的今天，专门针对科研领域的大模型产品仍然不算多，且大多数是试验性质的产品。

不久前，科大讯飞在发布最新版本的讯飞星火认知大模型V3.0时，也一口气发布了12个面向行业的专用大模型。其中便有联合中国科学院文献情报中心共同研发的、面向科研工作者的科技文献大模型，以及基于该大模型的应用产品——星火科研助手。这也是国内为数不多的专门为科研工作推出的大模型产品。目前，星火科研助手有成果调研、论文研读、学术写作三大功能。

浙江大学第一附属医院图书馆工作人员以“大语言模型”为关键词对星火科研助手进行了试用。在“成果调研”板块，星火科研助手在检索到的1251314篇文献中遴选了167篇文章进行分析，给出了关于大语言模型的概述。其还可以进一步从遴选的167篇文章中勾选最多30篇文章，据此生成综述。

星火科研助手的论文研读功能则采用当前大语言模型通用的问答模式，可默认生成论文摘要、方法、结论等主要信息；用户也可以就自己关心的论文内容进行提问，科研助手会根据文章内容进行回答。其学术写作功能则主要聚焦科研文章的翻译与润色，目前支持中英文互译，也可以对研究人员撰写的英文文章进行润色。

须保证内容真实且专业

由于技术原因，大模型有时会出现编造信息、“一本正经地胡说八道”的现象。这种现象在业内被称为AI幻觉。生活中，人们在和大模型聊天时，如果出现了AI幻觉，人们可能会一笑了之；但若AI幻觉出现在追求严谨精确的科研领域，后果可能就会很严重。

科大讯飞北京研究院执行院长、科技文献大模型研发负责人伍大勇表示，研发科技文献大模型，核心难点就在于保证其内容的可信性和专业性。“一方面，这要依靠高质量的论文数据；另一方面，在模型预训练和监督微调方面也需要下功夫。”伍大勇说。

他介绍，科大讯飞通过与中国科学院文献情报中心合作，在合规的情况下获取了丰富的科技文献数据，并对数据进行了去重、去噪等处理，以提升数据质量。“星火科研助手采用中国科学院文献情报中心提供的论文接口来进行论文检索。此外，我们还使用了基于论文知识库的检索增强和知识增强策略。这些都使大模型生成的结果有据可依。”伍大勇表示，这些措施从技术上保证了星火科研助手回答结果的准确性，也尽量避免了大模型出现AI幻觉。

未来，人工智能或将帮助科研工作者跳过文献检索、粗读的过程，直接找到需要的文献，大幅提升科研工作者的文献阅读效率。

同时，伍大勇表示，针对科技文献服务的各个场景，星火科研助手研发团队还邀请专业团队，对大模型训练数据进行监督微调，以提升星火科研助手在科技文献服务上的性能表现。“例如在成果调研和论文研读功能上，我们采用大模型结合知识图谱和知识库的策略，以保证产品输出的内容有据可依。在学术写作上，我们针对学术翻译和学术英语润色专门进行了大模型监督微调，以达到比通用翻译和校对产品更强的专业性。”伍大勇说。

或能激发科研工作者更多灵感

虽然目前尚未有太多人工智能产品被应用于科研领域，但已有学者对人工智能进军科研提出了反对意见，认为这会让科研工作者变得懒惰。在严伯钧看来，科研工作者在应该“懒惰”的地方“懒惰”，反而可以节省出更多时间用在更有价值的工作上。

阅读文献前首先要进行文献检索。为此，科研工作者往往需要搜寻大量文献，在此基础上对部分感兴趣的文献进行粗读，以进一步判断哪些是自己真正需要的文献。这是实打实的“体力活”。严伯钧认为，在这种情况下，借助人工智能工具帮助科研工作者跳过检索、粗读的过程，以更高效的方式直接找到需要的文献，可大幅提升科研工作者的文献阅读效率。

虽然可以借助大模型等工具来检索阅读文献，但严伯钧也不否认读原文的价值。“原文当然要读，但并不一定是每篇都要读。更加精准地定位到需要的文献后再进行精读，是更加高效的方式。”

伍大勇同样表示，研发星火科研助手的初衷在于帮助用户快速了解论文核心内容，提高论文研读效率，让科研工作者能够把更多精力花在更为重要的实验验证等工作上。“辅助提升科研效率是科技文献大模型的关键和目标，但科研工作所需要的灵感、思路、逻辑推理、实验验证、创新与探索等仍离不开科研工作者发挥主观能动性。”

事实上，除了能够辅助阅读文献，人工智能已经在多个科学研究领域带来实际成果。例如在预测蛋白质结构方面，人工智能产生的成果已经远超人类过去工作的总和。严伯钧认为，这种需要大量计算、反复试错的工作，正是人工智能的强项，人类应与其形成合理分工，拥抱新技术。

谈及未来人工智能可能给科研工作带来的改变，严伯钧认为，目前的文献阅读、翻译润色等功能，可能只发挥了人工智能在科研工作领域潜力的1%。在他看来，当下科研发展正呈现出细分化的趋势，一位学者往往只深耕于某一科研领域，而人工智能的跨界思维模式未来或能给科研工作带来一些改变。“或许人工智能可给科研工作者带来更多跨领域、交叉学科的原创性启发，激发科研工作者更多想象力。”

都芃

算力紧缺背景下提升训练推理效率迫在眉睫

“技术升级+一站构建”助大模型降本增效

如何在算力紧缺的背景下提升大模型训练和推理的效率，并降低成本？这已成为一众大模型企业不得不面对的难题之一。

日前，腾讯披露，腾讯混元大模型背后的自研机器学习框架Angel再次升级。“自研机器学习框架升级后，腾讯大模型训练效率可提升至主流开源框架的2.6倍，用该框架训练千亿级大模型可节省50%算力成本，大模型推理速度提高了1.3倍。”11月30日，腾讯机器学习平台部总监陶阳宇向科技日报记者表示。

不只是腾讯，在提升大模型训练效率、加速大模型落地应用方面，一批中国企业交出了自己的“答卷”。

双管齐下节约算力成本

在大型模型训练和推理过程中，需要消耗大量算力资源。因此，提高硬件资源利用率，对国产大模型技术的发展至关重要。

陶阳宇介绍，面向大模型训练，腾讯自研了机器学习框架Angel。该框架针对预训练、模型精调和强化学习等全流程进行了加速和优化。据悉，它采用FP8混合精度训练技术，并深度优化了4D混合并行训练策略，还在ZeROCache技术基础上减少了冗余模型存储和内存碎片，提升了内存的利用率。同时，该框架还可兼容适配多款国产化硬件。

2023世界人工智能大会上，观众正在参观国内大模型落地案例。在提升大模型训练效率、加速大模型落地应用方面，一批中国企业交出了自己的“答卷”。

而据媒体披露，除了提高硬件资源利用率，针对通信策略、AI框架、模型编译等进行系统级优化，亦可大幅节约训练调优和算力成本。

此外，随着模型参数的增大，大模型推理的成本也随之攀升。陶阳宇介绍，腾讯自研的大模型机器学习框架Angel通过扩展并行、向量数据库、批处理等多种优化手段，提高了吞吐能力，达到了更快的推理性能，降低了成本。

不只是腾讯，在第二十届中国计算机大会上，百度首席技术官王海峰就公开透露，文心大模型4.0从今年3月发布至今，其训练算法效率已提升3.6倍；通过百度飞桨与文心大模型的协同优化，文心大模型周均训练有效率超过98%，推理性能提升50倍。

此外，据公开资料显示，阿里云通义大模型则聚焦于规模定理，基于小模型数据分布、规则和配比，研究大规模参数规模下如何提升模型能力，并通过对底层集群的优化，将模型训练效率提升了30%，训练稳定性提升了15%。

让大模型“开箱即用”成为可能

不难看出，调整和优化模型的训练和推理方式，其最终目的都指向使模型更好地适应实际应用场景、降低在终端应用中的额外成本。“大模型的应用和研发同样重要。”腾讯机器学习平台专家工程师姚军说，只有提供方便、强大的接入平台，才能让大模型真正走向应用。

百度创始人、董事长兼首席执行官李彦宏也曾表示，大模型本身是不直接产生价值的，基于大模型开发出来的应用才是大模型存在的意义。然而，很多大模型落地的难度很大，因为一个大模型往往会对应着很多不同种类的应用，这需要大量的接口和流量支持。

如何破解这道难题？据悉，基于自研机器学习框架Angel，腾讯打造了大模型接入和应用开发的一站式平台，让针对业务场景的数据处理、模型微调、评测部署和应用构建等多个环节，从以往“散装”的多团队协作方式，转化成流水线平台上自动化生产方式，让大模型的“开箱即用”成为可能。“开箱即用”的关键在于预训练基础模型的泛化能力，高性能框架提供的微调或扩展工程能力，以及应用平台的灵活构建能力等支撑。据媒体披露，目前腾讯会议、腾讯新闻、腾讯视频等超过300个腾讯产品及场景均已接入腾讯混元大模型进行内测，数量相比10月份翻了一倍，覆盖文本总结、摘要、创作、翻译、代码等多个场景。比如，腾讯混元大模型就可支持智能化的广告素材创作，满足“千人千面”的需求。

《北京市人工智能行业大模型创新应用白皮书（2023年）》数据显示，截至2023年10月，我国10亿参数规模以上的大模型厂商及高校院所共计254家，分布于20余个省市/地区。

“未来大模型产品的发展趋势可能是通用大模型与垂直领域细分模型的结合。”中国人民大学数字经济研究中心主任李三希此前表示，这不仅需要具备坚实的技术基础，如大规模、高质量、多样化的语料库，创新的大模型算法，自研的机器学习框架和强大的算力基础设施等，也需要大模型产品具有坚实的基于场景的应用。未来，从实践中来，到实践中去的“实用级”大模型将成为趋势。

罗云鹏

创新杂谈 chuangxinzatan

加快构建基础研究长周期支持机制

吴月辉

前不久，国家自然科学基金委员会官网发布消息，将从2024年起，对上一年底资助期满的国家杰出青年科学基金项目（以下简称“杰青项目”）开展分级评价，同时择优遴选不超过20%的优秀项目给予第二个五年滚动支持，资助强度加大至800万元；资助期满后再择优遴选不超过50%的优秀项目给予第三个五年1600万元的资助，通过十五年近3000万元的高强度支持，集中优势资源培养造就高水平领军人才。

设立于1994年的杰青项目重点支持基础研究优秀人才，五年内给予稳定的经费资助，在科技界广受好评。此次杰青项目改革，旨在探索构建对优秀基础研究人才的长周期支持机制，让他们在基础研究领域更加心无旁骛地自由探索。

优秀的基础研究人才队伍是强化基础研究能力的基本保障。基础研究特别是原创性基础研究难度大、周期长、风险高，长期稳定的经费支持至关重要。对优秀科研人员、高水平研究团队等给予长期稳定支持，有利于激励他们聚焦重大基础科学问题和需要长期积累的研究方向，也有助于打造原始创新策源地和基础研究先锋力量。

对优秀科研人员、高水平研究团队等给予长期稳定支持，是国内外基础研究资助的成功经验。德国政府给予马普学会长期稳定的经费支持，截至目前，该学会的科研人员已获得多项诺贝尔科学奖。在科技部和北京市长期支持下，北京生命科学研究所取得多项重大原创性突破，培养了一批优秀科技人才。

近年来，我国在构建有效的长期稳定支持机制方面不断加大力度。2018年国务院印发的《关于全面加强基础科学研究的若干意见》指出，强化稳定支持，优化投入结构。2020年，科技部、财政部等6部委共同制定并发布的《新形势下加强基础研究若干重点举措》明确提出，完善基础研究投入机制，加大对长期重点基础研究项目、重点团队和科研基地的稳定支持。2021年，中国科学院与财政部共同试点开展“稳定支持基础研究领域青年团队”，遴选了100个基础研究领域青年团队，给予5年为周期的稳定支持，积极培养优秀青年科技人才。

同时也要看到，基础研究长周期支持机制的构建是一项长期任务，不可能一蹴而就。目前，支持机制仍存在稳定支持经费比例较低、稳定支持项目较少等问题，还需要进一步完善。为此，要进一步明确稳定支持机制和各类竞争性项目机制的定位，加强统筹协调，优化资助端供给结构；要提高稳定支持经费中的研究经费及个人薪酬占比，增加基础研究人员的获得感。同时，也要依据基础研究的特点，优化实施过程中的绩效评估，确保“好钢用在刀刃上”。

当前，新一轮科技革命和产业变革突飞猛进，基础研究转化周期明显缩短，国际科技竞争向基础前沿前移，对基础研究能力快速提升提出更加迫切的需求。加快构建和完善有效的长周期支持机制，将进一步激发更多优秀基础研究人才和团队的创新活力，夯实高水平科技自立自强的根基。

2023世界5G大会主打跨界融合

（上接A1版）从5G元年起步，世界5G大会以高规格、国际化、专业化的特征成为国际各方高度重视、业界积极参与的重要国际会议，大会内容逐年丰富，这是5G有效驱动实体经济数字化、智能化转型和升级，进一步推动融合应用新业态、新模式蓬勃兴起的最好印证。

大会组委会表示，走入第五年的世界5G大会与往届相比，可浓缩为一个关键词，就是“跨界融合”。

随着5G惠民兴业的应用广泛落地，越来越多的跨界融合成为现实。会前经过数月征集的5G融合应用揭榜赛已经涌现出很多扎根各行各业，有效解决实际问题的科技创新项目，而2023世界5G大会期间更将通过会、展、赛集中展示众多跨界融合优秀成果，比如，5G助力数字孪生工厂、数据采集控制、高精度定位等赋能生产制造领域；5G辅助自动驾驶、无人物流、智慧出行等为交通领域提效；5G在急诊救治、远程诊疗、医院管理等医疗健康领域发挥的作用日益显现；在农业领域，“新一代数字科技与生物技术深度融合”迎来智慧育种的时代；在生态环保领域，5G成为绿水青山的守护者……

“筹备过程中我们深刻感受到河南赋予5G和世界5G大会更多的可能性，在这里我们看到千行百业从传统快步走向未来，深厚文化和创新科技融合相拥。”世界5G大会总策划人、未来移动通信论坛副秘书长富军表示，2023世界5G大会将举全国之力，聚全球资源，为河南探索新技术、新应用、新业态献计献策，为河南打造中国5G创新高地赋能。

刘艳

智联世界 共享健康

（上接A1版）采用开源的JAVA开发语言，搭建自主可控的J2EE平台架构，拥有独立自主知识产权，全面兼容国产化计算机客户端、服务器、国产数据库，可以为国内医院用户提供稳定可靠的服务。

当医疗资源分布合理、人才按需流动、就诊次序良性循环，缓解看病贵、看病难问题才会成为可能。据了解，该平台实现了机构间数据信息互联共享、业务协同，促进医疗资源合理配置、基本医疗卫生服务均等化，降低运营成本，提高服务效率。使县域内的医疗卫生资源配置更加科学，基层的医疗卫生服务能力、效率和活力进一步提升，医保基础性作用进一步发挥，医疗服务价格动态调整机制基本形成，做到县强、乡活、村稳，构建县乡一体、以乡带村、上下联动、信息互通的新型基层医疗卫生服务体系，实现“五升三降”。

“例如，定襄县的医共体信息化建设就连接了全县医疗卫生机构15个，服务人口20余万。”张俊亮讲道。定襄县连通县域内各层级医疗机构，构建“分级分层”“多级多层”的医共体一体化服务体系，不断提升医疗服务质量和效率，切实增强人民群众就医获得感、幸福感、安全感。

公司负责人韩温表示，2023年是医药健康行业重塑、企业转型升级的重要窗口期。智杰软件在抓住政策机遇的同时，将持续加大科技创新投入，促进技术和产品迭代升级，为医疗卫生机构和广大群众不断提供更多便捷舒心的智慧医疗产品和优质服务。