油气大模型破局需从三方面发力

◎刘 合


视觉中国供图

在人工智能技术日新月异的今天，生成式人工智能的代表——ChatGPT的横空出世，不仅在短时间内吸引了全世界关注的目光，更激发了各行业对大型预训练模型的无限遐想。油气行业作为国民经济的支柱之一，同样期待它能为油气勘探、开发带来新变革。尤其是随着国内油气资源品质逐步劣质化，油气勘探开发难度逐渐加大，亟须运用新技术提质增效。

油气大模型应用面临挑战

数据、算力和算法是大模型发展的核心要素。其中，数据是大模型应用的基石，算力是大模型应用的保障，算法是大模型应用的工具。由于油气行业具有特殊性，在上述三个层面，油气大模型开发都面临着诸多挑战。

在数据方面，油气大模型应用面临数据稀缺、复杂和安全性的挑战。一是油气行业的数据涵盖了地质勘探、钻井、生产和运输等多个环节，每个环节的数据采集都伴随高昂的成本，样本数量少且获取非常困难，采集回的数据还具有多解性和不可验证性的特点。二是油气行业的数据类型非常多样化，处理这些不同种类、不同版本、不同结构的数据本身就充满挑战。三是大模型需要学习海量数据，但油气行业对数据的安全性和保密性有着极高的要求，数据不能上传公有云，且必须防止泄露，因此普遍存在“数据孤岛”问题。这一现状使得如何在保障数据安全前提下，整合分散的数据并训练行业基础模型，成为油气行业大模型应用的关键难题。

在算力方面，油气大模型的训练和优化也面临着算力资源不足的挑战。大模型的训练和优化通常需要巨大的算力资源，这往往伴随着相当高的投入成本。自建算力中心需要巨额资金投入，而租赁算力又存在数据安全和隐私保护的问题。目前，国内油气行业仅具备有限的微调算力，普遍不具备训练行业基础模型所需的高水平算力。此外，由于各种原因，国内油气行业在短期内很难建立起满足大模型需求的算力资源。这一问题进一步加剧了算力资源的短缺，使得大模型的应用和发展受到严重制约。

在算法方面，油气大模型也面临版权纠纷等挑战。算法的优劣直接影响大模型的实际应用效果。相较于传统深度学习等算法，大模型的技术门槛更高，目前的发展主要依赖少数高端算法人才推动。尽管许多开源大模型算法可以作为研发基础，但它们可能缺乏必要的技术支持和安全保障，存在商业机密泄露的风险，且其能力往往不如闭源算法。此外，开源算法的版权协议中存在诸多限制条款，使得基于开源算法进行研发时可能面临版权纠纷。如果选择使用闭源算法，则难以实现核心算法的自主可控。

从数据、算力和算法入手推动大模型应用

油气大模型应用并非坦途，需要在数据、算力和算法等方面破局。

首先，数据之困需破冰。面对数据采集高成本与复杂性并存的挑战，破解数据之困，要以大模型应用为契机，推动数据治理，确保数据的全面性、准确性和时效性。油气行业在大模型方面的核心竞争力是“行业数据”，要做好“训练样本库”的基本功。油气企业必须强化数据全生命周期管理，从数据源头、数据采集、数据清洗、数据融合和匹配、数据完整性增强、数据标注等环节严格规范，建立高质量的训练样本库，提升数据治理能力，为模型提供坚实的数据基础。同时，应通过数据脱敏、数据加密、访问控制和审计、合规性审查等方式加强数据安全和隐私性保护。如设置合适的权限和用户角色，限制用户对数据库的访问和操作，保护数据的安全性。还需进行数据库的维护和优化工作，包括定期备份、数据清理、性能监控等。在此前提下，构建一批高质量开源数据集，推动油气大模型研发生态建设。

其次，算力建设应灵活。面对资金投入与隐私保护之间平衡的挑战，破解算力之困，应以油气大模型为契机，推动融合算力建设。可采取租赁与自建相结合的方式，注重智算能力建设的同时，加强数据安全与隐私保护。例如，企业应根据自身业务需求、成本预算和技术实力，灵活选择算力获取方式。对于常规的计算任务，可通过租赁公有云资源快速响应；而对于涉及敏感数据或需长期稳定运行的任务，则可考虑自建或合作共建数据中心，确保数据安全与算力的可持续供给。在算力设施规划上，应注重长远，实现通用计算、智能计算和高性能计算的融合布局，通用计算满足日常运营的基本计算需求，智能计算侧重于深度学习、机器学习等智能算法的高效执行，高性能计算则针对大规模科学计算和复杂模拟，满足不同场景的需求，显著提高算力资源的利用效率。

此外，算法创新勿盲从。面对大模型训练周期长与迭代速度快的双重考验，破解算法之困，应量身定做适合行业特性的算法模型，避免盲目跟风。应理性认识大模型的价值，优先实施场景模型和数据质量优良的L2行业基础模型，重点应放在微调和适配下游任务上，避免盲目投入通用基础模型的研发，确保技术栈的自主可控。油气行业应秉持资源优化配置的原则，聚焦油气主营业务，从投入成本、产出效益、技术成熟度，以及稳定性、行业聚焦、核心竞争力等方面慎重考虑。在岩心分析、地震资料处理解释、测井数据分析等特定领域，大模型能发挥显著作用，但不可过度依赖，应明确模型的适用范围。

为了推动大模型技术的自主可控，还需加强“AI+能源”复合团队的建设。训练、应用大模型不能闭门造车，要打破传统行业壁垒，注重联合研发生态的建设，例如推动油气行业与互联网企业、高校等的合作，促进跨学科人才整合，形成产学研用紧密结合的创新体系，为油气大模型应用构建可持续的人才保障。同时，可通过项目合作、人才培养、共建研发平台等方式加强大模型算法等方面的合作交流，并明确合作目标与分工，以及知识产权分配与管理、数据保密和隐私保护等制度和规范。

大模型必将推动油气行业新质生产力发展，未来可期，但道阻且长。油气行业要充分认识油气大模型的特殊性，从数据、算力、算法等方面做好工作，稳扎稳打，逐步推进，让AI成为推动油气行业转型升级的重要驱动力。

（作者系中国工程院院士、中国石油勘探开发研究院正高级工程师）

创新预测理论与方法 更好服务经济民生决策

◎杨翠红 田开兰

党的二十届三中全会提出，科学的宏观调控、有效的政府治理是发挥社会主义市场经济体制优势的内在要求。进行科学合理的宏观经济预测可为预判经济形势、开展宏观调控、制定经济发展规划等提供重要依据。经济系统是一个不确定性很强、影响因素众多的复杂系统，当前我国宏观经济预测仍然存在准确度不够高等问题，亟待科学改进预测理论与方法，提高预测精度，有力服务经济民生决策。

经济预测精准度有待提高

宏观经济预测是根据经济学的基本原理和经济发展规律，基于主要经济指标之间的因果联系，结合经济波动周期，在一定信息资料的基础上，利用科学的方法和手段，对未来一定时期内主要经济指标走势进行预测，发现风险和问题，变事后应对为事前调控，以提高政策执行的科学性和有效性。例如，提前半年以上时间，高精度预测当年的粮食产量能够帮助政府和农业企业准确研判农业生产形势，科学安排粮食生产、储备和进口；相反，如果没有进行预测，当因歉收而需要进口粮食时，国际市场的粮价很可能已大幅上涨，进口成本大幅上升。

由于经济社会系统的不确定性，直接导致经济预测经常面临准确度不高等问题。

一是经济预测的结果会影响预测对象——人的预期，产生“俄狄浦斯效应”，反过来影响预测结果的精度。例如，如果预测未来几年会发生经济危机，企业可能会减少投资，家庭会降低消费，使得经济下滑提前到来，之前的预测便可能出现“不靠谱”的情况。

二是经济系统中总有超预期事件发生，蕴含很大的不确定性，使得经济预测常常需要及时修正。经济系统由数量庞大的、具有自由思维的人组成，其想法和行为往往相互影响，使得经济系统自由度、随机性很大。同时，当前全球经济系统是一个相互依存、相互联系的整体，国际经贸环境变化和重大国际政治事件都会给相关经济体带来外部输入的不确定性，对经济预测造成干扰。此外，经济系统可能遭受重大的外部冲击。这些系统内部的随机性和外部的重大冲击都可能会给经济预测带来意想不到的结果，因此在有超预期重大事件发生时，往往需要及时修正预测结果。

三是经济预测结果的准确度与数据的可获取性和数据质量密切相关。由于所处发展阶段以及统计制度和数据管理制度的不完善，我国可能存在部分数据难以获取以及部分统计数据弄虚作假的问题，使得预测模型无法捕捉真实的经济情况。随着“统计造假”被纳入《中国共产党纪律处分条例》以及统计技术和方法的日益完善，我国的统计数据质量日益提高，统计数据质量问题给经济预测带来的干扰将越来越小。

多措并举促进预测精度提升

宏观经济预测是一项复杂而富有挑战性的工作。它需要全面分析和掌握经济社会现象之间的内在联系，以准确可靠的数据资料为依据，根据研究对象特点综合运用多种科学预测方法对经济的各个方面进行分析，需要不断改进、创新预测理论与方法，提高预测精度。

第一，需要创新经济理论与方法以更好地刻画经济主体预期的动态变化及其对经济预测的影响。虽然行为经济学对经济预期有一定研究，但是目前的经济预测模型还无法较好地量化静态预期、非理性预期、理性预期和适应性预期等不同类型的经济预期对预测结果的影响，这是经济预测领域急需攻克的重要问题之一。

第二，进一步发挥大数据预测方法和实时预测方法在经济预测领域的作用。基于大数据的预测方法，通过挖掘及时性强、数据量大、获取成本低的海量数据中隐藏的模式和关联关系，发现变量之间的相互作用，从而进行预测。由此，可以克服传统预测方法所需的统计数据相对滞后、数据集较小、获取成本高等不足。例如，中国科学院数学与系统科学研究院研究员洪永淼提出的基于大量微观股票价格数据的机器学习预测方法很好地改进了通货膨胀率的预测精度。近年来，实时预测方法在GDP等宏观经济指标预测中取得了良好效果。其基本思想就是利用所有可获得的大数据信息，包括公布时间不同步的数据、有缺失值的数据、截面维度高和频率不一致的数据等来获得对经济状态的预估，提供动态的、及时的经济预测。大数据预测是对传统预测方法的有益补充，将在宏观经济预测中发挥越来越重要的作用。

第三，集成多种效果好的预测方法，采取组合预测方法提高预测结果的可靠性。组合预测即采用多种预测方法建立多个预测模型，得出多个预测值，然后对这些预测值进行科学分析，通过加权平均组合法或递推回归系数组合法等方法将多个预测结果组合成最后的预测结果，降低单个模型预测结果的不可靠性或者系统性偏差，从而提高预测结果的准确度。例如，中国科学院数学与系统科学研究院的多项预测工作均利用了组合预测：陈锡康研究员提出的预测粮食产量的以投入占用产出技术为核心的系统综合因素预测法，在多因素分析基础上，建立了多个预测方程进行组合预测，1980年以来已连续44年提前半年预测当年全国粮食产量，年均误差仅为1.6%，达到很高预测精度，在服务农业决策中发挥了重要作用；汪寿阳研究员提出的TEI@I方法论是一种兼具传统统计技术与人工智能技术的方法论，系统地融合了文本挖掘技术、经济计量模型、人工智能技术与系统集成技术，在外汇汇率与国际原油价格波动预测等领域取得了很好预测效果。

当前，宏观经济预测虽面临一些挑战，但其意义在于描绘经济社会未来可能的走向以及背后的逻辑，探索最可能发生的情景，为当前决策提供指导。相信在未来实践中，通过不断优化预测理论与方法，将会提高预测结果精度，更好地为经济民生决策服务。

（作者杨翠红系中国科学院数学与系统科学研究院副院长、研究员，田开兰系中国科学院数学与系统科学研究院副研究员）

学报观点要览

从ChatGPT演进历程看未来发展趋势

文章：《ChatGPT的工作原理、关键技术及未来发展趋势》
期刊：西安交通大学学报，2024年第1期
作者：秦涛、杜尚恒、常元元、王晨旭
评荐：管晓宏（中国科学院院士、西安交通大学电子与信息学部主任）

大数据时代，如何有效挖掘和利用海量数据中蕴含的知识，是我们当前亟须解决的难题。自然语言处理技术作为连接数据与知识之间的桥梁，是解决这一难题的重要工具。近年来，生成式人工智能逐渐成为自然语言处理的利器。ChatGPT作为生成式人工智能的典型代表，极大地推动了自然语言处理技术的发展，引发了自然语言处理研究范式的转变，使自然语言处理技术能够更加高效、智能地适应多样化的应用场景。

通过大规模的预训练，生成式人工智能具备了强大的上下文理解与自然语言文本生成能力，可以完成对话问答、信息检索等任务，与人类交互更加自然和灵活，成为当前自然语言处理领域的重要工具之一。

从整体架构来看，ChatGPT遵从“基础语料+预训练+微调”的基本范式。海量高质量的基础语料是技术突破的关键，预训练是构建大规模语言模型的基础，微调是实现模型实际应用的保障。GPT-4在上述架构的基础上进行了多模态升级，多模态输入能力对语言模型至关重要，使其可以获得除文本描述外的常识性知识，为多模态感知与语义理解的结合提供了可能性。这一新范式可归纳为“预训练+提示+预测”。

该文梳理了ChatGPT的模型架构和技术演进过程，重点讨论了提示学习、指令微调、思维链、人类反馈强化学习等关键技术，并结合运行原理分析了其面临的缺陷与挑战，探讨了进一步改进的着力点，为自然语言处理领域的深入研究提供有益参考。

该文提出，虽然ChatGPT的性能优越，但它是完全基于数据的模型，具有一些不可避免的局限性。例如，对垂直领域知识挖掘不充分不深入而造成的生成内容不够合理准确、训练数据偏差导致的生成内容存在潜在偏见、容易受到对抗性攻击从而诱导模型产生偏离事实的输出、无法高效在线更新近期知识等问题。

展望未来，ChatGPT还有很大的提升和改进空间，包括采用迁移学习、引入机理和领域知识等方法提升领域适应能力，增强多模态处理能力、筛选训练语料等方法提升模型的泛化性。此外，在进一步增强模型人机交互协同能力、提升模型可信度，构建新的可信测试基准等角度，都有提升空间。

开源生态建设助力国产大模型创新发展

文章：《推动我国大模型开源创新生态建设的挑战与建议》
期刊：中国科学院院刊，2024年第8期
作者：温馨、张超、郭锐、陈凯华、冯泽、朱其罡
评荐：杨柳春（院刊执行主编）

大模型基础资源门槛高、产业集群效应强、潜在垄断性大等特点，成为国产大模型快速形成行业积累、实现迭代发展与赶超的制约因素。大模型开源创新生态是指具有相同开源理念的多元创新主体，围绕开放数据、开源框架、开源软硬件等数字基础设施，实现价值共创的复杂系统。其旨在以开放、协作、共享的精神，整合大模型创新链各环节基础资源、降低研发门槛，从而激发群体智慧，促进大模型技术持续创新、广泛传播和产业化应用。

通过开源方式降低研发门槛，是美国大模型以往取得领先优势的基础，各国也正在通过开源创新生态发展大模型，我国应积极应对。开源是全球公认的突破科技垄断或制约的有力手段，能帮助汇聚全球开发者智慧，推动大模型技术进步，并激发社会创新活力，加快大模型应用落地。开源创新生态建设是我国人工智能技术和产业发展的突破口，有助于我国企业摆脱对具有封闭知识产权技术的依赖，提升科技话语权、化被动为主动，对促进国产大模型技术迭代与产业化落地、推动潜在国际合作破除垄断壁垒、培育未来产业竞争优势等具有决定性意义。

该文梳理了国际上大模型开源创新生态的成功经验和做法，重点讨论了如何构建稳固完善的上游供应生态、丰富多元的下游应用生态和公开有效的治理协调生态。同时，指出了我国大模型开源创新生态建设面临的诸多挑战，如系统协同政策架构设计缺失、技术能力制约生态形成、数据算力显著限制技术发展、创新主体无序竞争制约整体发展速度、开源支持体系建设水平较低等。

为此，该文指出，应加强顶层设计，坚持系统观念，统筹谋划开源技术生态，以数据、算力和算法为抓手补短板、固底板，推动产学研持续投入大模型开源技术研发。同时，要打造共享的大模型研发基础体系、强化全产业链开源开放体系、完善开源创新治理体系等。

展望未来，我国应充分吸收开源创新生态构建经验，秉持开源开放的理念，构建大模型开源创新生态，推动大模型全产业链的繁荣有序发展。一方面，要处理好打造大模型开源生态过程中政府和市场之间的关系。另一方面，要建立起对开源的合理认知，探索构建符合大模型产业特性的开源治理体系，推动形成涵盖大模型上下游全产业链的健康开源创新生态。

AI生成内容如何获得专利法保护

文章：《生成式人工智能时代发明人制度的结构性改革》
期刊：北京邮电大学学报（社会科学版），2024年第4期
作者：张惠彬、王怀宾
评荐：罗璨（西南政法大学人工智能法学院副教授）

党的二十届三中全会提出，完善生成式人工智能发展和管理机制。完善的法律法规是生成式人工智能安全治理的基石，然而现行专利制度中的发明人只能是自然人，人工智能既不是自然人，也不是法人，因此被排除在发明人行列之外。随着社会发展，这种做法是否符合专利法立法鼓励发明创造、促进科学技术进步和经济社会发展的目的，也是学界讨论的热点问题。

当前，人工智能技术在各科学领域得到了深入而广泛的应用，促进了相关科学领域的发展，例如在医药分子药物合成领域的应用大大加速了新药研发进程。该文认为，加强知识产权保护可以激励更多创新者投入到人工智能领域的研究和开发中，从而推动新质生产力的不断发展。知识产权的明确界定和有效保护，可以为人工智能技术的商业化应用提供有力保障，促进相关产业的繁荣和发展。

此外，随着生成式人工智能技术的蓬勃发展，发明人资格的界定、专利权的归属规则、专利审查的基准以及诚实信用原则均面临新的挑战与考验。传统专利法体系下的发明人制度有可能引发专利申请领域的诚信缺失问题，并可能诱发由人工智能驱动的低质量专利泛滥现象。因此，发明人制度的改革已成为迫切需求。

该文指出，应适应现代发明活动的集体化特征以及人工智能产业的快速发展需求，进行发明人制度改革，从而激发创新投资活力，推动技术成果的有效公开，并坚守诚信原则的底线。

在理念上，应实现从单一激励发明人向全面激励创新投资转变；在制度形式上，应制定人工智能署名指导原则，明确人工智能在发明创造过程中的贡献与地位；在制度实质上，则需构建针对人工智能自主发明的审查指南，依据科学严谨的标准，为人工智能自主发明的审查工作提供专属的制度框架。

该文提出，通过改革传统发明人制度，服务技术发展与产业创新，具备相当的参考价值。鉴于当前人工智能技术日益呈现专业化、智能化的发展趋势，各行业均将其视为实现降本增效的重要工具。因此，确立清晰明确、科学合理的专利规则，对于产学研各界更好地保护其智力劳动成果，具有至关重要的意义。

专栏主持人：刘若涵
电话：010-58884176
邮箱：liurh@stdaily.com