

【理论探讨】

数智赋能的科研创新

——基于数智技术的创新辅助框架探析

陆伟 马永强 刘家伟 杨金庆 程齐凯

【摘要】以 ChatGPT(Chat Generative Pre-trained Transformer)为代表的人工智能大模型在文本生成、人机对话等方面展现出了优异的性能。在大模型背景下,大数据、人工智能等数智技术在赋能科研创新方面表现出重要的现实价值。当前的科技信息资源管理和知识服务能够为科研创新提供较为准确的信息以及常规的知识聚合服务,但是仍未能与科研创新活动形成深度融合。同时,科研人员在科研活动中也面临信息处理能力不足、认知能力有限等挑战。据此,本文首先对数智时代科研活动的新特点进行了剖析,然后提出了基于数智技术的创新辅助框架,并对所提出的框架进行了深入分析和探讨,阐述了其在创新全过程中的功能定位、服务模式和关键赋能路径。未来,随着大数据和人工智能技术的不断成熟和进步,数智赋能的科技信息资源管理将进一步嵌入科研创新活动全过程。基于数智技术的创新辅助服务能够为科研人员提供个性化、细粒度的知识和场景化的解决方案,如面向文献阅读、实验设计和论文撰写场景的创新辅助服务,从而更好地服务于科研创新活动。

【关键词】数智赋能;ChatGPT;人工智能大模型;科学智能;科研创新全过程;创新辅助框架

【作者简介】陆伟(1974-),男,教授,博士生导师,武汉大学信息管理学院,武汉大学信息检索与知识挖掘研究所,研究方向为信息检索、AI治理、人机协同,E-mail:weilu@whu.edu.cn(武汉 430072);马永强(1997-),男,博士研究生,武汉大学信息管理学院,武汉大学信息检索与知识挖掘研究所,研究方向为信息抽取、文档智能;刘家伟(1994-),男,博士研究生,武汉大学信息管理学院,武汉大学信息检索与知识挖掘研究所,研究方向为信息检索、信息安全(武汉 430072);杨金庆(1991-),男,博士,讲师,华中师范大学信息管理学院,研究方向为科技情报、学科知识演化(武汉 430079);程齐凯(1989-),男,博士,副教授,武汉大学信息管理学院,武汉大学信息检索与知识挖掘研究所,研究方向为文本挖掘、信息检索(武汉 430072)。

【原文出处】《情报学报》(京),2023.9.1009~1017

【基金项目】国家自然科学基金重点项目“数智赋能的科技信息资源与知识管理理论变革”(72234005);国家自然科学基金面上项目“基于机器阅读理解的科学命题文本论证逻辑识别”(72174157)。

0 引言

2022年年末,OpenAI发布的ChatGPT^[1]在人机对话、文本生成等方面展现出令人惊艳的性能,让人们看到了人工智能大模型在赋能科研创新方面的巨大潜力^[2-4]。以大数据、人工智能为代表的数智技术作为一种新的生产力正被广泛地应用于科学和工程等领域^[5-6]。2023年3月27日,科技部启动了“人工智能驱动的科学驱动”(AI for Science)专项部署工作^[7]。以大模型为主要特征的数智技术对资源获取、组织开发、传递利用、存储检索等信息资源管理学科核心任

务^[8]和文献情报工作^[9]都具有重要的影响。

ChatGPT系列模型的升级过程,正是面向大规模数据的知识获取能力不断提升的过程。科研创新是一项知识密集型的活动,高质量、高密度的知识输入是科研创新得以开展的基础,并且贯穿了科研创新的全过程。从供给侧来看,虽然科技信息资源管理与知识服务能够为科研创新提供较为准确的信息以及常规知识聚合服务,但是仍未能与科研创新活动形成深度嵌入。从需求侧来看,为了获得支撑科研活动开展的信息和知识,在科研创新过程中的各个

环节,科研人员面临着信息处理能力和认知能力有限等挑战。然而,如何为科研人员提供个性化、细粒度、场景化的知识仍未得到有效解决。为此,现有科技信息资源管理理论方法、技术手段和服务模式亟须推进“供给侧改革”。本文认为,以嵌入科研创新活动的形式,使用数智技术为科研人员提供创新辅助服务,面向文献阅读与梳理、科研问题发现和假设构建、实验设计与验证和论文撰写等主要环节,为科研人员提供个性化、细粒度、场景化的知识,实现对科研创新活动的赋能。

当前,已有研究人员试图将大数据与人工智能技术应用于科研创新场景,以解决科学研究中的现实问题。例如,基于大数据和自然语言处理技术对科技信息资源进行全面评价^[10-11]、细粒度理解^[12-14]和提供知识服务^[15-16]。其中,面向科研活动的人工智能模型开发通常是基于分治的思想,即模型仅仅面向特定的任务,如论文摘要自动生成^[12]、引用意图识别^[14]、论文版式识别^[17]等;面向单个学科,如面向生物医学和计算机科学领域的SciBERT^[18]、面向化学领域

的ChemBERT^[19]和面向生物学领域BioGPT^[20]。然而,创新活动实际上包含一系列的过程,是一项系统性的工作。单靠某一个人工智能模型并不能有效解决科研人员创新活动中遇到的复杂问题。因此,本文基于数智时代科研活动的新特点,总结了数智时代科研人员面临的信息处理能力不足、认知能力有限等挑战,针对数智时代科研活动的新特点和科研人员面临的新挑战,提出了基于数智技术的创新辅助框架,并进一步阐述了其功能定位、服务模式和在科研创新全过程中的关键赋能路径。

1 数智时代的科研活动的新特点

科研人员是科研创新活动的主体。科研人员先通过阅读科学文献获取相关的观点和事实等信息,以从中获得需要的知识;然后,通过相关仪器设备发现新的事实,从而得到新规律、新发现;最后,科研人员对新规律、新发现进行分析和综合,形成科学成果,以学术论文、专利的形式进行发表。通过上述过程的不断循环,新知识得以源源不断地产生,具体过程如图1所示。知识生产的循环过程中伴随着新的

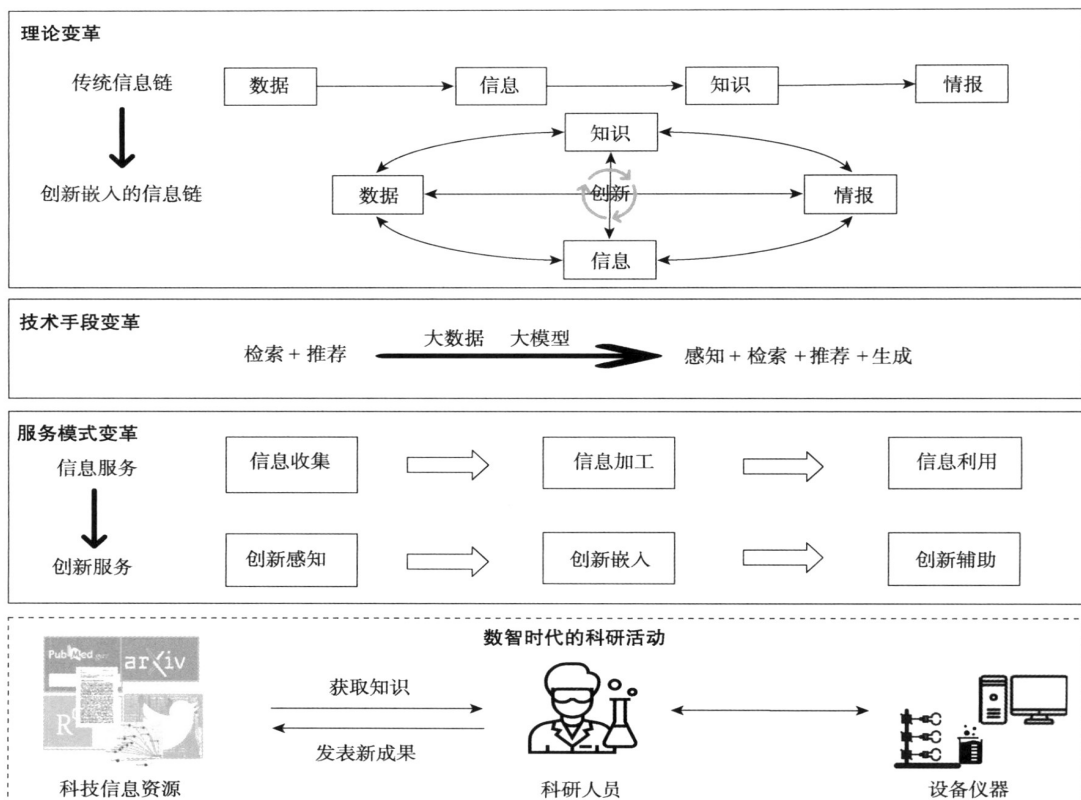


图1 数智时代科技信息资源管理和科研创新变革

科技信息资源的生成,这些资源可以作为数智时代人工智能大模型的知识源泉,实现对人工智能大模型的增强。

1.1 数智时代科技信息资源管理理论、技术与服务模式变革

在大数据背景下,科学交流系统也发生了新的变化。其中,科学知识的载体形式更加多样,交流渠道也更趋多元。借助社交媒体等工具,科学交流系统中的科研人员之间的信息流动速度更快^[21]。依托互联网平台(如社交媒体、预印本网站、问答社区等)、在线会议工具(如Zoom、腾讯会议等),科研人员可以更加高效、快捷地获取领域前沿的信息。同时,科学需要解决的问题也更加复杂,学科间交流合作更加频繁。人工智能技术作为一种新的生产工具,也越来越多地参与到科研创新活动中,例如,使用 scite.ai、LitMaps 进行文献发现,使用 AlphaFold 进行蛋白质3D结构预测等。科研人员在获得便利的同时,也面临着信息处理能力不足和认知能力有限等问题。

面对数智时代科研活动发生的新变化和科研人员面临的新挑战,现有科技信息资源管理理论方法、技术手段和服务模式同样面临变革。在理论方法方面,传统的“数据-信息-知识-情报”信息链转变为创新嵌入的信息链。借助于大数据、人工智能等数智技术,数据与知识、数据与情报、信息与情报之间的跨越式转换成为可能。在技术手段方面,现有的大型预训练语言模型(如对话语言模型 ChatGPT)展现了强大的信息加工、荟萃、整合和生成能力。对话式人工智能大模型可以为用户提供比传统的搜索引擎、推荐系统更加自然的交互方式,实现由“检索+推荐”到“感知+检索+推荐+生成”的技术范式转变^[22]。在服务模式方面,现有的以信息为中心的服务模式转变为面向创新过程的创新服务,包括创新感知、创新嵌入和创新辅助等。

1.2 数智时代知识生产的新变化

科学创新活动可以视为一种知识生产活动^[23-24]。作为一种创造性的思维活动,科学知识生产的最初动力源于社会发展中人们所提出的各种各样的问题,当前科研创新越来越多地与问题情境紧密结合,数智技术也越来越多地与其他学科进行深度嵌入。

此外,为了解决复杂问题,各学科之间从各自独立发展走向跨学科合作。

1.2.1 知识生产与问题情境紧密结合

问题情境是科研创新的切入点。随着科学的进步和所需解决问题复杂性的不断增加,以知识生产为目的的科研创新活动与一连串的问题情境之间的互动也越来越密切^[25]。为了更好地解决问题,科研人员首先需要深入了解和刻画所面对的问题情境,然后才能更好地确定研究方向,发现解决问题的途径。在问题情境的驱动下,科研人员需要探索 and 开发新的途径和方法来解决问题。以分子合成问题为例,传统“试错式”的研究模式存在周期长、成本高的问题。为此,研究人员尝试使用人工智能技术实现化学合成路线自动化设计^[26],以提升分子合成路线设计的效率。

1.2.2 数智技术与科研创新活动深度嵌入

数智技术作为一种新的生产工具,越来越多地与科研创新活动进行深度结合与嵌入。一方面,数智技术越来越多地被用于解决具体的科学问题。例如,基于主动学习的量子实验设计^[27]、蛋白质结构预测模型 AlphaFold^[28]、基于 AI(Artificial Intelligence)的天气和气候基础模型 ClimaX^[29]。另一方面,数智技术作为一种新的科研创新辅助工具,越来越多地被嵌入科研创新全过程的某些任务场景中。例如,Auer 等^[30]构建了开放研究知识图谱,用于对解决特定研究任务的最新技术进行简要概述;Pankratius 等^[31]和 Pyzer-Knapp 等^[32]相继对计算机辅助自动科学发现和人工智能辅助新材料发现进行了探索。

1.2.3 智能技术推动多学科合作走向深入

在产业界,企业之间通过专业化协作来共同完成产品的生产。作为一种知识生产活动,科研创新活动中学科之间的合作也广泛存在,例如,化学学科的科研成果已经逐渐渗透到生物学,形成了化学生物学。打破了学科藩篱,促进了学科交叉,实现不同学科间科研人员的深度合作成为加快科技创新的关键。以解决科学问题、社会问题为中心的多学科合作,成为推动基础研究、应用研究等各个领域发展的主要动力。2020年,国际纯粹与应用化学联合会(International Union of Pure and Applied

Chemistry, IUPAC)将人工智能纳入了“2020年化学领域十大新兴技术”(Top Ten Emerging Technologies in Chemistry)^[33],人工智能技术和化学研究的跨学科合作变得越来越紧密。

1.3 数智时代科研人员面临的挑战

以问题解决为导向的科研活动可以视为一种认知的过程。为了解决设定的问题,需要问题解决者(即科研人员)运用并重组已有的信息、知识,朝着问题的目标状态进行认知操作和运行。然而,在处理科研创新过程中的复杂任务时,科研人员自身能力也存在着诸多不足,主要表现在记忆能力、信息编码能力和信息存储能力等方面。上述诸多不足可以分为两大方面:一是信息加工处理能力不足,无法有效应对任务环境中快速增加的信息;二是认知能力受限,无法以并行的方式解决复杂的科学问题。

1.3.1 信息加工处理能力不足

随着数字基础设施的发展,科学交流系统中的信息和知识载体已经从单一的纸质形式,扩展到了纸质和数字形式。科技信息检索工具,如谷歌学术,可以帮助科研人员找到相关的文献,但是检索到的文献并不能直接向科研人员提供其所需的知识。因此,科研人员需要花费时间来人工阅读文献,以获取需要的知识。单就学术论文而言,其年发文量增长速率正在逐年上升,同时预印本网站上的论文数量也呈指数增长。面对不断增加的文献,科研人员只能通过逐篇阅读的方式从大量的学术文献中获取所需的信息和知识。然而,有研究表明,科研人员每年阅读的论文数量趋于平稳,平均每年只有250篇左右^[34-35]。本文认为,“急剧增长的科技信息资源规模与科研人员有限的信息加工处理能力的矛盾”是阻碍科研效率提升的关键因素之一。

1.3.2 认知能力受限

认知是指人们获得知识或应用知识的过程,包含感觉、知觉、记忆、思维、想象和语言等。在科研创新活动中,科研人员通过自身的认知能力来发现、理解和解决问题,这些活动包括文献阅读与梳理、科研问题发现和假设构建、实验设计与验证和论文撰写等。传统的科学研究主要是基于“假设构建-实验验证”的模式,其本质上是一种不断试错的方法。这种

方法不仅依赖于科研人员的个人经验,而且需要花费大量的时间进行验证。科学研究是一个在科学知识空间中进行搜索的过程^[31,36],其中科学知识空间包括了理论空间、概念空间、实验空间和设备空间等。即使在狭窄的领域内,也有巨大的潜在方向可供探索。受限于人类大脑有限的认知能力^[5],科研人员只能处理科学知识空间中的小部分。此外,科研人员在科学知识空间中进行搜索的过程,也会受到自身认知偏见的影响^[37]。

2 基于人工智能大模型的创新辅助框架构建

当前,面向科研创新的应用和服务通常是面向创新全过程中的单个任务,应用服务之间缺乏联动协调,形成了一个数据孤岛。这阻碍了科研活动中知识的有效循环和流动。为此,基于人工智能大模型,本文构建了一个面向科研创新活动的创新辅助框架(图2)。该框架可以推动科技信息资源管理与知识服务向智能化转型。针对当前科研人员所面临的信息加工处理能力不足和认知能力受限的问题,本节首先阐述了创新辅助框架在科研创新中的定位,即作为智能知识转换器、智能科研助理和智能服务平台;其次,基于数智时代科技信息资源管理理论、技术与服务模式变革,提出了对话式的信息服务模式 and 基于API(Application Programming Interface)接口的插件式的智能服务模式,用于提升科研人员的信息加工处理能力和认知能力,从而更好地辅助科研人员开展科研活动;最后,针对科研创新活动的主要环节——文献阅读与梳理、科研问题发现和假设构建、实验设计与验证和论文撰写,阐释了数智时代创新辅助框架的赋能路径。

2.1 数智时代创新辅助框架的定位

Krenn等^[38]指出,人工智能在科学研究中发挥作用的三个维度:①作为工具的人工智能;②作为灵感和概念来源的人工智能;③作为理解代理人的人工智能。本文提出的创新辅助框架在科研创新中的主要定位是:作为智能知识转换器、作为智能科研助理和作为智能服务平台。

2.1.1 作为智能知识转换器

借助人工智能和大数据技术可以实现对大规模科技信息资源的高效利用,因而,数智技术具有提升

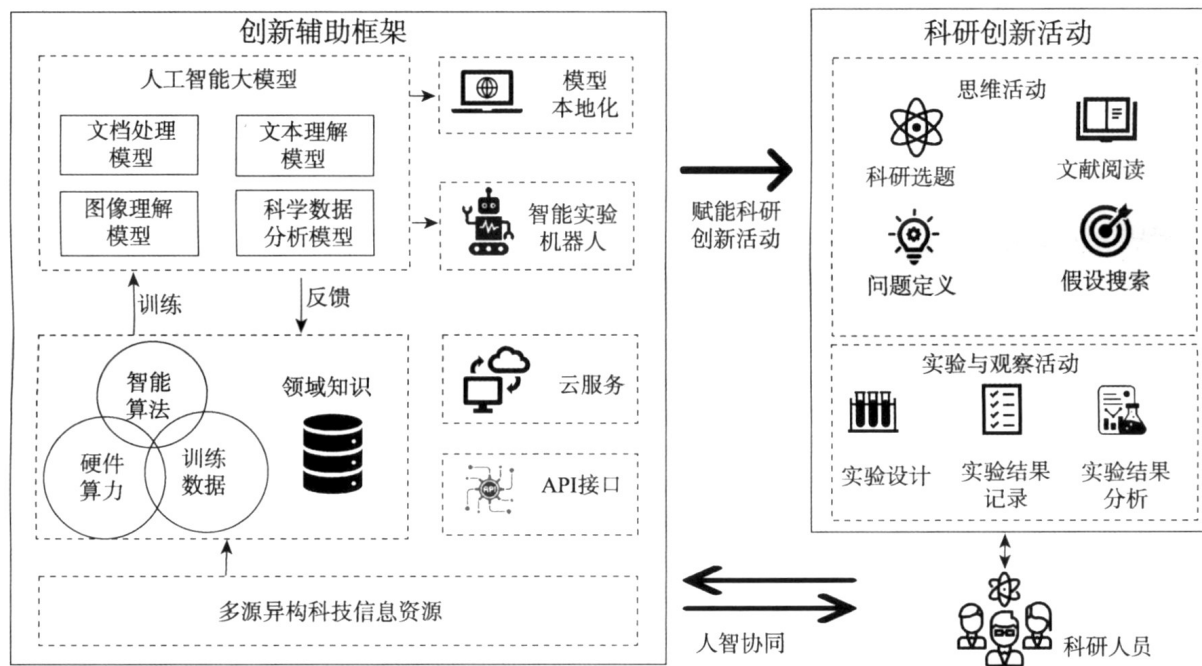


图2 基于人工智能大模型的创新辅助框架示意图

科研活动中知识的流动速度的功能。传统的科学研究范式中,科研人员需要花费大量的时间和精力来分析数据、阅读和梳理文献。基于大规模科技信息资源、高性能计算和人工智能算法,构建面向科研创新活动的人工智能大模型成为可能。数智时代创新辅助服务可以基于学术文档理解大模型和科学数据分析大模型等工具,实现数据与知识之间的快速转换,从而弥补科研人员在信息加工处理能力和认知能力上的不足。

2.1.2 作为智能科研助理

数据-技术-人之间的关系正在被重新定义^[39],通过机器智能和人类智慧的深度结合,可以形成高效的人智协作(Human-AI Cooperation, HAIC)^[40-41]科研创新模式,以实现对科研人员认知能力的增强。在科学数据的理解方面,人工智能如同一个“智能显微镜”,可以让科研人员更好地理解复杂数据中隐含的关联和模式。在研究结论的构建方面,依托人工智能算法对数据强大的分析和生成能力,基于数智技术的创新辅助服务,可以为科研人员提供对科学数据精确的分析和预测结果。在智能科研助理的辅助下,科研人员可以从依赖自身的先验知识的研究模式,转换到依托人工智能大模型进行研究的

新模式,集中更多精力在机理机制的发现和analys上。

2.1.3 作为智能服务平台

一方面,人工智能本身具有跨学科属性^[42],这使得人工智能的相关技术可以对其他学科产生影响,如物理、化学、材料和生物等领域。因此,基于人工智能技术的创新辅助服务天然地具有促进跨学科研究的功能。另一方面,当前大型语言模型可以出色地理解人类语言和完成人类指令^[43-44]。因此,基于领域知识设计相应的指令,可以直接让大模型生成指令对应的结果。这种以自然语言为信息交互媒介的使用方式极大地降低了大模型的使用门槛,使得不具有人工智能背景知识的用户也可以快速掌握使用方法。因此,基于数智技术的创新辅助服务可以发挥智能技术服务平台的作用,推动其他学科对智能技术的使用。

2.2 数智时代创新辅助框架的服务模式

2.2.1 基于生成模型的对话式信息服务模式

在数智时代背景下,与微软推出的新型检索工具 New Bing^[45]思路类似,本文所构建的面向科研创新活动的创新辅助框架,突破了现有的“检索+推荐”服务模式,转换为“感知+检索+推荐+生成”的模式^[46]。基于数智技术的创新辅助框架能够以更加自然的人

机对话形式,为用户提供其需要的知识。如图3所示,智能助手针对用户的问题,首先,自动调度和请求相关API接口,从互联网资源和机构内资源中获取相关文献;其次,对获取的文献进行语义理解,得到用户问题相关的资料片段;最后,借助人工智能大模型的信息整合和文本生成能力,为用户提供其需要的答案。基于科研人员提出的问题,智能助手就能够从输入的问题描述中理解真实的意图,为科研人员提供其需要的知识。

2.2.2 基于API接口的插件式智能服务模式

为了赋予大模型解决复杂任务的能力,OpenAI通过API接口将ChatGPT与外部应用(如Wolfram)进行连接^[47]。类似地,本文所构建的面向科研创新活动的创新辅助框架,通过向科研人员和情报服务机构提供面向科技信息资源的大模型API,可以让用户便捷地调用相关服务。一方面,基于数智技术的创新辅助框架,能够以插件的形式集成到其他的科技信息资源应用中,也可以与创新活动的具体环节进行嵌入,实现对创新全过程的赋能;另一方面,基于数智技术的创新辅助框架,作为一个智能中枢系统可以将复杂任务分解为简单的子任务,其他垂直领域的应用和API接口可以作为子任务解决者,从而实现科研创新全过程中多样化任务的解决。

2.3 数智时代创新辅助框架的赋能路径

从形式上看,科研创新活动主要分为两种基本类型:①观察和实验;②思维^[48]。这两类活动都是对象性的。其中,科学家通过观察具体的事物和现象,获得相关的经验和事实;通过思维活动构建相关理

论和规律,形成科学知识;最终以学术论文的形式进行发表。从流程上看,科研创新活动则包括文献阅读与梳理、科研问题发现和假设构建、实验设计与验证和学术论文撰写等主要环节。

2.3.1 文献阅读与梳理

阅读文献是科研人员获取领域最新知识的关键途径。科研人员的文献阅读过程是单线程的,因此,面对海量的学术文献,科研人员面临巨大的信息加工处理的负担。基于人工智能模型的创新辅助框架需要具备对海量学术文献进行批量化、并行化、细粒度的信息抽取、分析、归纳和总结能力。借助问答式检索技术^[15],数智时代创新辅助框架作为一个智能科研助手,能够为科研人员提供获取科技信息与知识的新途径。通过创新辅助框架中的对话式信息服务模式,科研人员可以为自己的文献阅读需求添加上下文并与系统响应交互。在人机对话的过程中,智能科研助手可以为科研人员提供精细化、个性化的内容,帮助科研人员厘清自身的信息和知识需求。

2.3.2 科研问题发现和假设构建

科研选题包括研究问题的确定和假设的构建,是创新活动的关键环节。科研选题对科研活动的成功与失败具有决定性的影响。传统的科研选题过程中,科研人员主要是基于自身的先验知识和有限的领域知识,认知能力面临巨大的负担。本文提出的基于数智技术的创新辅助框架,通过构建面向学术文献语义理解的大模型,能够对海量科技信息资源进行分析,从而辅助科研人员更全面地了解本学科的研究空白和痛点、热点;通过对科研人员研究经历

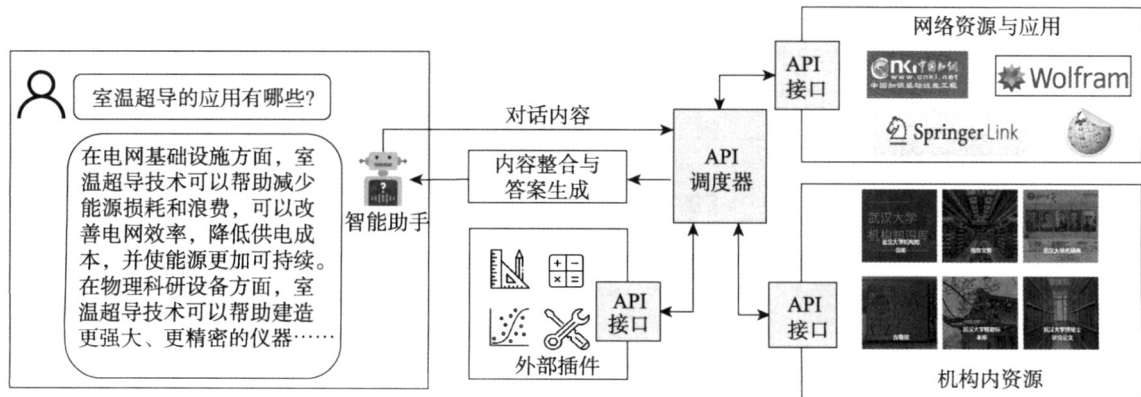


图3 数智时代创新辅助服务的服务模式示意图

和个人爱好的建模,可以更好地感知科研人员的创新需求,为其推荐未完全解决的问题或者尚待解决的问题。此外,当前大模型自身仍存在局限性,如生成内容存在知识性错误的问题;因此,当前大模型所生成的研究问题和研究假设仍只能作为科研人员开展研究的参考。

2.3.3 实验设计与分析

传统的科学研究主要是基于“假设构建-实验验证”模式,其实质上是一种不断试错的方法^[49],这需要研究人员花费大量的时间和精力。科技信息资源中包含大量的实验相关信息,如实验手册和学术论文中的实验过程以及实验结果等。基于数智技术的创新辅助框架可以在对实验相关信息资源的大数据分析和综合的基础上,提供面向特定研究问题的实验方案,同时对实验结果进行预判,为科研人员提供全面的参考。通过创新辅助框架中的基于API接口的插件式智能服务模式,可以把科研人员的实验设计方案转换为操作指令,然后通过外部接口与智能实验机器人,如IBM RoboRXN进行交互,实现自动设置反应条件、自动记录实验结果。

2.3.4 学术论文写作

学术论文是科研人员科研创新活动的结晶,其中包含科研人员发现的科学机理。当前,学术论文的写作主要是由科研人员来完成。学术论文包含引言、相关研究、研究方法、实验结果与分析和结论等章节,总篇幅通常达数十页。在前述的文献阅读与梳理、科研问题发现和假设构建、实验设计与分析等环节中,已经产生了大量与研究相关的内容。有研究表明,以ChatGPT为代表的大型语言模型可以生成媲美人类水平的文字内容^[50]。因此,依托大型语言模型在文本生成方面的能力,基于数智技术的创新辅助框架可以为科研人员生成论文相关的文本内容。此外,通过对画图工具、统计分析工具API接口的调用,还可以基于科研人员的指令和相关实验数据绘制出论文中需要的图表。

3 结论

AlphaFold预示着基于人工智能技术的科学智能所具有的广阔前景和应用价值。2022年年末发布的ChatGPT让人们看到了通用人工智能的曙光。当

前,人类社会已经到了新一轮科技革命——智能革命的前夜。科研创新活动是一项知识密集型的工作,为了从大规模科技信息资源中获得需要的知识,仅仅依靠科研人员个人的信息处理能力和认知能力将会越来越难以应对。人工智能技术作为一种高效的“知识转换器”和“智能助理”,可以快速、并行化地处理海量的科技信息资源,从中萃取出有价值的知识,增强科研人员的信息处理能力和认知能力。在系统思维的指导下,本文对数智时代的创新辅助框架进行了探讨,并提出了基于数智技术的创新辅助框架,以嵌入科研人员创新全过程中关键场景和任务的形式,可以实现对科研人员创新活动的赋能。

当前,大模型在生成内容的真实性和可靠性上仍存在不足,相信随着数智技术的不断成熟和进步,数智技术在科研创新中的应用会变得像数值模拟或微积分在物理学的应用一样普遍,从而更好地赋能科技信息资源管理与知识服务。大数据、人工智能具有推动各个科学领域和各个经济部门的创新和经济增长的潜力,正逐渐成为国家科研创新生态系统的关键驱动力,数智赋能的科技信息资源管理与知识服务则可以更好地服务于国家重大科技发展战略和科技创新。

参考文献:

- [1]OpenAI. ChatGPT: optimizing language models for dialogue[EB/OL].(2022-11-30)[2023-02-09]. <https://openai.com/blog/chatgpt>.
- [2]Hutson M. Could AI help you to write your next paper?[J]. Nature, 2022, 611(7934): 192-193.
- [3]Stokel-Walker C, van Noorden R. What ChatGPT and generative AI mean for science[J]. Nature, 2023, 614(7947): 214-216.
- [4]陆伟,杨金庆.数智赋能的情报学学科发展趋势探析[J].信息资源管理学报,2022,12(2):4-12.
- [5]赵志耘.论复杂信息环境下的科技情报卓智赋能[J].情报学报,2022,41(12):1229-1237.
- [6]许勇,黄福寿.人工智能哲学研究述评[J].上海交通大学学报(哲学社会科学版),2020,28(1):116-123.
- [7]新华社.科技部启动“人工智能驱动的科学”专项部署工作[EB/OL].(2023-03-27)[2023-03-28].http://www.gov.cn/xinwen/2023-03/27/content_5748495.htm.
- [8]马费成.守正创新,继续推进信息资源管理学科的发展

[J]. 情报资料工作, 2023, 44(1): 13-14.

[9]张智雄, 于改红, 刘熠, 等. ChatGPT对文献情报工作的影响[J]. 数据分析与知识发现, 2023, 7(3): 36-42.

[10]罗卓然, 陆伟, 蔡乐, 等. 学术文本词汇功能识别——在论文新颖性度量上的应用[J]. 情报学报, 2022, 41(7): 720-732.

[11]胡志刚, 章成志. 悄然兴起的全文计量分析[J]. 图书馆论坛, 2021, 41(3): 1-11.

[12]Cachola I, Lo K, Cohan A, et al. TLDR: extreme summarization of scientific documents[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2020: 4766-4777.

[13]程齐凯, 李鹏程, 张国标, 等. 学术文本词汇功能识别——基于标题生成策略和注意力机制的问题方法抽取[J]. 情报学报, 2021, 40(1): 43-52.

[14]Cohan A, Ammar W, van Zuylen M, et al. Structural scaffolds for citation intent classification in scientific publications[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 3586-3596.

[15]陈博立, 鲜国建, 赵瑞雪, 等. 科技文献问答式智能检索总体设计与关键技术探析[J]. 中国图书馆学报, 2023, 49(3): 92-106.

[16]孙坦, 刘峥, 崔运鹏, 等. 融合知识组织与认知计算的新一代开放知识服务架构探析[J]. 中国图书馆学报, 2019, 45(3): 38-48.

[17]Xu Y H, Li M H, Cui L, et al. LayoutLM: pre-training of text and layout for document image understanding[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 1192-1200.

[18]Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2019: 3615-3620.

[19]Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction[OL]. (2020-10-23). <https://arxiv.org/pdf/2010.09885.pdf>.

[20]Luo R Q, Sun L A, Xia Y C, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining[J]. Briefings in Bioinformatics, 2022, 23(6): bbac409.

[21]Stokel-Walker C. Twitter changed science—what happens now it's in turmoil?[J]. Nature, 2023, 613(7942): 19-21.

[22]陆伟, 刘家伟, 马永强, 等. ChatGPT为代表的大模型对信息资源管理的影响[J]. 图书情报知识, 2023, 40(2): 6-9, 70.

[23]王兴成. 科学经济学的对象[J]. 国外社会科学, 1982(1): 71-73.

[24]约翰·齐曼. 元科学导论[M]. 刘珺珺, 译. 长沙: 湖南人民出版社, 1988.

[25]迈克尔·吉本斯, 卡米那·利摩日, 黑尔佳·诺沃提尼, 等. 知识生产的新模式: 当代社会科学与研究的动力学[M]. 陈洪捷, 沈文钦, 译. 北京: 北京大学出版社, 2011.

[26]Segler M H S, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI[J]. Nature, 2018, 555(7698): 604-610.

[27]Melnikov A A, Nautrup H P, Krenn M, et al. Active learning machine learns to create new quantum experiments[J]. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(6): 1221-1226.

[28]Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[29]Nguyen T, Brandstetter J, Kapoor A, et al. ClimaX: a foundation model for weather and climate[OL]. (2023-07-10). <https://arxiv.org/pdf/2301.10343.pdf>.

[30]Auer S, Oelen A, Haris M, et al. Improving access to scientific literature with knowledge graphs[J]. Bibliothek Forschung und Praxis, 2020, 44(3): 516-529.

[31]Pankratius V, Li J, Gowanlock M, et al. Computer-aided discovery: toward scientific insight generation with machine support[J]. IEEE Intelligent Systems, 2016, 31(4): 3-10.

[32]Pyzer-Knapp E O, Pitera J W, Staar P W J, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics[J]. NPJ Computational Materials, 2022, 8: Article No. 84.

[33]Meyers F. IUPAC announces the 2020 top ten emerging technologies in chemistry[EB/OL]. (2020-10-25) [2022-02-02]. <https://iupac.org/iupac-announces-the-2020-top-ten-emerging-technologies-in-chemistry/>.

[34]Johnson R, Watkinson A, Mabe M. The STM Report: an overview of scientific and scholarly publishing[R]. Fifth Edition. The Hague: International Association of Scientific, Technical and Medical Publishers, 2018.

[35]van Noorden R. Scientists may be reaching a peak in reading habits[J]. Nature, 2014.

[36]Simon H A. The scientist as problem solver[M]//Complex Information Processing: The Impact of Herbert A. Simon. Hillsdale: Lawrence Erlbaum Associates, 1989: 375-398.

- [37]Hope T, Downey D, Weld D S, et al. A computational inflection for scientific discovery[J]. *Communications of the ACM*, 2023, 66(8): 62-73.
- [38]Krenn M, Pollice R, Guo S Y, et al. On scientific understanding with artificial intelligence[J]. *Nature Reviews Physics*, 2022, 4(12): 761-769.
- [39]马费成,张帅.我国图书情报领域新兴交叉学科发展探析[J].*中国图书馆学报*,2023,49(2):4-14.
- [40]罗威,罗准辰,雷帅,等.智能科学家——科技信息创新引领的下一代科研范式[J].*情报理论与实践*,2020,43(1):1-5,17.
- [41]Li J, Huang J S, Liu J X, et al. Human-AI cooperation: modes and their effects on attitudes[J]. *Telematics and Informatics*, 2022, 73: 101862.
- [42]Leeming J. How AI is helping the natural sciences[J]. *Nature*, 2021, 598(7880): S5-S7.
- [43]Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 27730-27744.
- [44]Chung H W, Hou L, Longpre S, et al. Scaling instruction-fine-tuned language models[OL].(2022-12-06). <https://arxiv.org/pdf/2210.11416.pdf>.
- [45]Microsoft. The new Bing: our approach to responsible AI [R/OL].(2023-02-01)[2023-03-19]. <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/02/The-new-Bing-Our-approach-to-Responsible-AI.pdf>.
- [46]Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 9459-9474.
- [47]Stephen W. ChatGPT gets its "Wolfram superpowers"[EB/OL].(2023-03-23)[2023-03-27]. <https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/>.
- [48]薛守义.科学性质透视[M].济南:山东人民出版社, 2009.
- [49]谢莹莹,马鹏宇,冯凡,等.2022 AI4S全球发展观察与展望[R].北京:北京科学智能研究院,深势科技,2022.
- [50]Ma Y Q, Liu J W, Yi F, et al. AI vs. human—differentiation analysis of scientific content generation[OL].(2023-02-12). <https://arxiv.org/ftp/arxiv/papers/2301/2301.10416.pdf>.

Data Intelligence Empowered Innovation: An Exploration of the Innovation Assistance Framework Based on Data Intelligence Technology

Lu Wei Ma Yongqiang Liu Jiawei Yang Jinqing Cheng Qikai

Abstract: Large Language Models (LLMs), such as ChatGPT (Chat Generative Pre-trained Transformer), have shown excellent performance in text generation and human-machine dialog. Under the background of LLMs, such technologies as big data and artificial intelligence have demonstrated important practical value in empowering scientific research and innovation. Although current Science and Technology (S&T) information resource management and knowledge services can provide relatively accurate information and routine knowledge aggregation services for scientific research and innovation, they have not yet been deeply integrated with scientific research and innovation activities. Researchers also face challenges such as insufficient information processing capabilities and limited cognitive abilities during the scientific research and innovation process. This article analyzes the new characteristics of scientific research activities in the data intelligence era. Furthermore, an innovation assistance framework based on data intelligence technologies is proposed, and its functional positioning, service mode, and key empowerment path in the entire innovation process are analyzed and discussed in depth. In the future, with the continuous maturity and progress of big data and artificial intelligence technologies, data-intelligence-technology-enabled S&T information resource management will be further embedded in the entire process of scientific research and innovation activities. Innovative assistance services based on data intelligence technologies, such as innovative auxiliary services for literature reading, experiment design, and article-writing scenarios, can provide researchers with personalized, fine-grained knowledge and scenario-based solutions to better serve scientific research and innovation activities.

Key words: Data intelligence empowerment; ChatGPT; Large language models; AI for science; Entire innovation process; Innovation assistance framework