ChatGPT应用背景下图书馆科研数据服务版权风险研究

闫宇晨

摘　要　随着人工智能技术的进步,ChatGPT因其在自然语言处理方面的强大能力而备受关注。在图书馆领域,ChatGPT与科研数据服务的结合具有广阔的应用前景和巨大的实践价值。然而,这种结合也带来了版权方面的法律挑战。本文深入探索了ChatGPT在科研数据的收集、分析和共享中的应用,并对其在图书馆传统的版权许可方式和合理使用制度下可能引发的版权风险进行了阐释。同时,借鉴国外科研数据使用版权例外的改革经验,按实施阶段提出两种解决方案:一是在司法实践中对合理使用条款的适用范围做扩张解释;二是增设人工智能科研数据使用著作权例外。图1。参考文献55。
关键词　ChatGPT　图书馆　科研数据服务　版权风险

Research on Copyright Risk of Library Scientific Research Data Service under the Background of ChatGPT Application

Yan Yuchen

Abstract: With the advancement of artificial intelligence technology, ChatGPT has garnered significant attention for its strong capabilities in natural language processing. In the library sector, the integration of ChatGPT with scientific research data services presents broad application prospects and substantial practical value. However, this integration also introduces legal challenges related to copyright. This article delves into the applications of ChatGPT in the collection, analysis, and sharing of scientific data, and elucidates the copyright risks that may arise from traditional library copyright licensing practices and fair use regime. Additionally, drawing on international experiences in copyright exceptions for the use of scientific data, the article proposes two approaches based on implementation phases: first, an expanded interpretation of the fair use provisions in judicial practice; second, the introduction of copyright exception for the use of AI in scientific research data. 1 fig. 55 refs.
Keywords: ChatGPT; Library; Scientific Research Data Service; Copyright Risk

近年来,人工智能技术不断进步,从苹果的Siri助手到AlphaGo围棋机器人,再到Google的垃圾邮件过滤器,弱人工智能在我们生活的各个领域大显身手。2022年11月,OpenAI公司发布了聊天生成预训练转换器——ChatGPT,并对其进行持续的优化和更新。2023年3月,OpenAI最新的大型语言模型GPT-4一经推出即引发热烈反响,该模型不仅支持与人类的自然对话交互,还能基于文本、图片和视频等内容生成准确且有创意的答案,《纽约时报》甚至盛赞其为“有史以来向公众发布的最佳人工智能聊天机器人”[1]。ChatGPT可能标志着通用人工智能的一个重要里程碑,预示着强人工智能即将到来[2]。尽管关于ChatGPT是否真正具备深度的自我意识和完整认知能力的问题仍然存在争议,但不可忽视的是,作为一款尖端的语言模型,ChatGPT在预训练期间已接受广泛的数据训练,从而使其能够理解、学习并执行多种领域的智能任务。这种预训练赋予其强大的问题解决能力,为其未来的应用开辟了广阔的空间。例如,在教育领域,ChatGPT能够帮助构建沉浸式的数字化学习体验,推动学生的学习方法创新,并为教师在教学与研究中提供创新支持[3]。在新闻和出版领域,ChatGPT正在改变内容的创作方式、读者的互动方式以及出版物的传播方式[4]。ChatGPT的推出也引起了图书馆界的广泛关注,相关研究主要集中在以下两个方面:

DOI:10.13666/j.cnki.jnlc.2024.0303

一是 ChatGPT 与图书馆服务的融合应用前景。例如,吴若航、茆意宏探讨了 ChatGPT 给图书馆服务带来的发展机遇[5];张慧等认为以 GPT 类技术为标志的 AI 2.0 可以为智慧图书馆注入新的活力,并有望实现智慧图书馆的 GPT 技术驱动创新[6];郭亚军等详细描述了 ChatGPT 在图书馆智能服务中的应用场景,如资源建设增值化、咨询服务智能化、社会教育均等化、科研服务专业化[7]。二是 ChatGPT 在图书馆服务中可能带来的风险和挑战。沈奎林认为,随着 ChatGPT 与图书馆服务的深度结合,可能会在技术、科学及伦理等多领域出现争议性问题[8]。蔡士林、杨磊进一步指出,由于 ChatGPT 的运行依赖于大量的数据,其中必然会涉及受版权法保护的数据,很可能引发版权风险问题[9]。相关研究提醒我们,ChatGPT 为图书馆业带来了新的发展机遇,但也带来了一定的风险,这就要求图书馆在实际应用中建立起适当的风险预防和管理机制,深入了解 ChatGPT 在各种服务场景中的使用方法及可能的风险,这对于推动 ChatGPT 在图书馆界的成功应用至关重要[10]。

科研服务是 ChatGPT 技术应用于图书馆服务的关键领域,也是有效提升智慧图书馆建设水平的重要方面[11]。在 E-Science 的大背景下,科学研究逐渐从实验、理论、计算机科学向数据密集型科研转变[12]。与此同时,图书馆因其专业的数据管理团队和丰富的数据资源库,不仅成为数据的存储中心,更是一个集深厚专业知识和丰富实践经验于一体的数据服务中心。多年来,图书馆在数据管理和服务领域持续努力,已经初步建立了一套较为完善的科研数据管理流程和服务体系,并使科研数据服务成为图书馆科研服务的一个重要组成部分[13]。科研数据服务是指包含数据导航与检索、数据存储与备份、数据发布与共享等一系列数据管理和服务的活动[14],在增强信息素养、评估科研发展趋势以及提升科研效率等方面具有显著作用。国际图书馆联盟(IFLA)的一项调研指出,ChatGPT 在图书馆科研数据服务领域具有广阔的应用前景,同时也面临着来自法律和伦理等方面的多重挑战[15]。2018 年国务院办公厅印发《科学数据管理办法》,第二十三条明确规定:“科学数据使用者应遵守知识产权相关规定”。这表明,若要实现 ChatGPT 与图书馆科研数据服务的有效融合,必须严格遵守知识产权法律法规,这是图书馆打造并推广新服务模式的重要前提。本文认为,确保科研数据的规范使用不仅是图书馆提供科研数据服务的重要基础,也是充分释放数据要素价值的关键途径。鉴于此,本文以 ChatGPT 在图书馆科研数据服务的应用为研究对象,深入分析其涉及的版权使用问题,揭示可能产生的版权风险,并结合我国《著作权法》相关规定,提出图书馆合法应用 ChatGPT 的版权法方案。

1 ChatGPT 技术原理及其在图书馆科研数据服务的应用

1.1 ChatGPT 技术原理分析

早在 1950 年,计算机科学先驱阿兰·图灵就对机器学习的可能性进行了探索,并深入思考计算机是否有能力模仿人类的思维方式。然而,直到 21 世纪初,这一领域才实现了实质性的突破。在这一时期,大数据的迅速积累、算法的持续优化以及算力的显著增强三者相得益彰,共同推动了机器学习向深度学习的技术演进[16]。深度学习是一种模拟人类大脑神经网络结构的先进机器学习技术,它的设计灵感来源于人脑神经网络的工作机制[17]。在人类大脑中,无数的神经元通过互相连接来学习和传输信息。与此类似,深度学习的人工神经网络也是由多层模拟神经元构建而成,这些层级之间进行交互,以执行复杂的学习和信息处理任务。值得说明的是,虽然 ChatGPT 的设计和实现是基于深度学习的基本原理,但它也展现了一些独特的技术特点。其一,自注意力机制(Self-Attention Mechanisms)。自注意力机制是 ChatGPT 的一个核心工作机制,它使深度学习模型能够处理数据中长距离的依赖关系,并实现并行计算。简而言之,当深度学习模型生成一个

单词时,它会考虑文本中的所有其他单词,以更准确地捕捉上下文和复杂的语言结构[18]。这种机制不仅增强了模型的上下文理解能力,还提高了其数据训练效率。其二,大规模预训练(Large-Scale Pre-Training)。在此过程中,模型利用大量文本数据,通过分层训练与微调,提高其在特定任务上的适应性,从而在多种自然语言处理任务和对话场景中展现出卓越性能[19]。

ChatGPT 的基本工作流程可划分为三个核心阶段:输入层、隐藏层和输出层,如图 1 所示。(1)输入层:作为人工神经网络的起点,输入层负责接收和处理原始数据,将其转换为一种机器可解析的格式。例如,文本数据会被转换为数字序列,音频数据被转化为特定格式,而图像数据则被转化为像素信息,以便进行后续的分析使用。(2)隐藏层:隐藏层构成了神经网络的关键部分,主要负责对数据进行深度的处理和学习。这里的"深度"是指模型中多个叠加的隐藏层,它们从不同的维度解析数据,进而产出更精确的结果。在这种多层结构中,每一层的"输出"都会成为下一层的"输入",形成一种分层的学习机制[20]。以图像识别问题为例,不同的隐藏层可能会对图像中的不同特征进行识别,如颜色、形状或纹理,并通过一系列"是/否"的决策判断来提炼这些特征,以识别图像中的具体对象。(3)输出层:输出层位于人工神经网络工作的终点,根据隐藏层的处理结果,提供最终的预测或判决。在此阶段,输出层通过回复模型将答案转化为更加易于人类理解的形式,如文本描述、图表或其他形式。

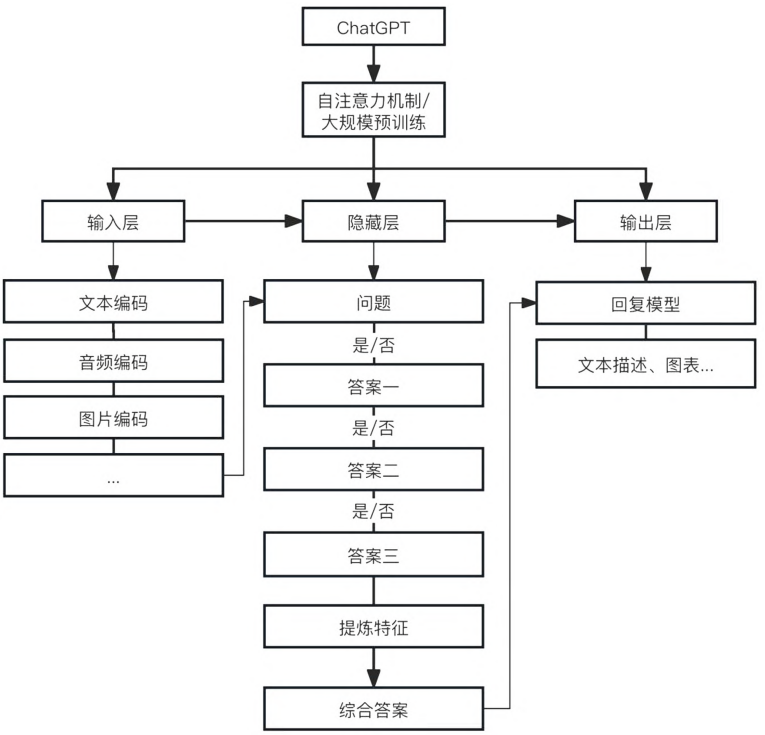

图 1　ChatGPT 技术原理分析

1.2　ChatGPT 在图书馆科研数据服务中的应用场景

从数据生命周期来看,图书馆科研数据服务涵盖了从数据的收集创建、检索分析到共享重用的各个环节[21]。基于此,本文认为 ChatGPT 可以从以下三个方面融入图书馆科研数据服务中。

第一,科研数据收集服务。首先,高效检索在线资源。得益于在线联网功能的推出,ChatGPT

不仅能实时更新并捕捉学科领域内的最新研究进展和数据动态,还能自动化检索各大学术资源平台(如 JSTOR、SSRN、PubMed 等),以快速定位科研人员所需的相关论文、研究报告和数据集。此外,ChatGPT 也可以从学者的个人主页、社交媒体平台和其他在线渠道获取与研究相关的数据。其次,个性化收集数据。ChatGPT 能够充分理解科研人员的具体需求,并实现对数据的高度个性化搜索和筛选,从而呈现与研究主题密切相关的科研数据。最后,初步整合数据。在收集阶段,ChatGPT 可以对获取到的数据进行初步的整合和精细化处理,例如,ChatGPT 能够整合无序的数据、去除重复项,并修正数据中的不准确之处,从而构建一个干净、精确且适于进一步分析的数据集供后续研究使用。

第二,科研数据分析服务。图书馆的科研数据分析服务致力于为教师、学生及其他研究者提供全面的数据分析援助。这些服务涵盖了从数据筛选、整合、分类到可视化的全过程,目的是协助用户更高效地运用数据,进而得到更为准确和更有深度的研究发现。以哈佛大学图书馆为例,其数据馆员与科研人员合作,利用专业管理软件将科研数据与相关文献关联,增强了检索结果的丰富性,提升了研究数据的价值[22]。将 ChatGPT 融入科研数据分析服务,能进一步提升图书馆的服务能力。这主要得益于:(1)ChatGPT 具有极强的语言理解能力。ChatGPT 不仅具备出色的自然语言处理能力,能够理解复杂的查询意图并提供基于深度学习的研究建议和新视角,而且对多种机器学习方法(如决策树、支持向量机和神经网络等)都有深入的了解,这使其能够深度挖掘并利用数据中的潜在价值和关系。(2)ChatGPT 具有优秀的语言表达能力。ChatGPT 在对用户问题进行理解和数据分析的基础上,可将分析结果以直观和易懂的方式呈现给用户。以科研数据分析服务中的数据可视化为例,ChatGPT 可以根据数据的特点为用户推荐合适的可视化工具,例如 Matplotlib、Tableau 和 Power BI 等,同时提供关于数据可视化的设计方式和最佳实践的建议。

第三,科研数据共享服务。在科研领域中,数据获取技术的提高、数据存储技术的进步以及数据高速传输网络的发展,使得海量数据的积累和传播成为可能,也使得科研数据的开放共享成为全球共识。图书馆作为科研数据管理者,在帮助实现科研数据共享方面发挥着重要作用[23]。例如,在国际上,以剑桥大学图书馆为代表的学术型高校图书馆,建立了专门的科研数据服务团队管理和共享数据,旨在推进科研数据更广泛的共享和更高效的应用,从而进一步促进学术领域的创新与合作[24]。在国内,复旦大学社会科学数据平台[25]、北京大学开放研究数据平台[26]等数据管理平台,在促进科研数据共享方面都发挥了积极的作用。

随着智能化时代的到来,图书馆有望通过 ChatGPT 进一步推动科研数据的共享和合作。具体而言,这种合作主要体现在以下几个方面:(1)图书馆可以利用 ChatGPT 为科研人员推荐最适合其需求的数据共享平台。首先,ChatGPT 能够为研究者展示各种数据共享平台的特点和优势,帮助他们更全面地了解数据共享的各种可能性。其次,ChatGPT 可以分析科研人员的具体需求,例如,计划发表的科研领域(自然科学、人文科学、社会科学等)、预计产出的成果类型(课题项目、学术论文、行业规范、知识产权等)。最后,ChatGPT 可以结合上述特定内容为科研人员提供个性化的平台建议。(2)图书馆可以借助 ChatGPT 来协助科研人员满足数据共享相关标准和规范的要求。为了确保数据的可靠性、可验证性,数据共享通常需要满足一系列标准和规范的要求。一方面,ChatGPT 可以为科研人员解释这些标准中的专业术语和要求,如文档编制方式、元数据描述规则等。另一方面,ChatGPT 还可以根据研究者的数据处理需求提供具体修改建议,如进行数据格式优化、数据清洗或增加必要的元数据,以确保数据满足相关标准。(3)图书馆可以利用 ChatGPT 构建一个专门针对科研数据共享的交流平台。这一平台应主要由以下两个核心内容组成:一是科研数据共享论坛。该论坛为科研人员提供一个开

放的平台,他们可以在此提问、寻找答案、分享自己的经验或观点,并与同行进行深入的讨论。这不仅可以促进知识的传播和分享,还有助于加强科研人员之间的合作与交流。二是科研数据共享互动工具。通过 ChatGPT 的强大功能,图书馆能够为科研人员提供实时的咨询和解答服务。同时,该工具还能够捕捉科研人员在数据共享中可能面临的挑战和需求,从而为图书馆创新和完善服务提供工作优化方向。

综上,ChatGPT 可以在图书馆科研数据服务中发挥重要的作用。它在迅速并准确地进行科研数据收集方面具有巨大优势,科研人员可利用其强大的数据分析、计算能力获取学术分析和解答。在数据共享方面,ChatGPT 也能够高效地传递科研数据,为科研活动提供强大的数据支撑。

2 ChatGPT 应用于图书馆科研数据服务的版权使用行为

科研数据是科研人员对技术、自然和社会等领域观察、解读和预测的结晶,蕴含着丰富的创新性信息,而版权法作为保护创意类成果的专门法律制度,会与科研数据使用行为产生密切联系。有研究者强调,当图书馆使用 ChatGPT 来处理和组织信息时,必须高度重视版权问题,确保其在合法的框架内运用[27]。本文将结合 ChatGPT 在图书馆科研数据服务中的应用场景,进一步探讨其与版权使用行为的内在联系。

2.1 ChatGPT 与科研数据收集:大量的作品复制行为

在机器学习中,数据爬取是获取有价值信息的关键步骤。从获取信息的方式来看,数据爬取可以直接从网上抓取或通过 API 接口获取[28]。ChatGPT 主要通过以下两种方式进行数据收集活动:其一,爬取互联网的公共数据资源。尽管 OpenAI 公司并未透露 ChatGPT 训练数据集的具体来源,但一些技术专家已经通过反向工程技术对其进行了分析。研究表明,ChatGPT 所使用的数据主要来源于各种学术期刊、书籍、在线百科、社交媒体以及像 Common Crawl 这样的公共数据集[29]。其二,利用 API 接口直接与数据库进行交互。随着 ChatGPT 开放 API 接口,开发者能够将其无缝集成到自己的应用或系统中,实现与本地或实时数据的深度交互。例如,ChatGPT 可与图像生成工具形成交互,用户通过 ChatGPT 描述他们心中的画面,图像生成工具(如 Midjourney)会根据这些描述生成相应的图像。

从技术角度看,ChatGPT 的数据爬取实际上是从网页或 API 接口提取内容并将其保存到本地的过程,这本质上是一种数据复制行为。从版权角度看,这些潜在的数据来源中可能包含了大量受版权保护的内容。因此,ChatGPT 在爬取数据时,可能触及到著作权法中的复制权问题。《民法典》第一百二十三条明确指出,知识产权是权利人基于法律对特定对象所享有的专有权利。其中,复制权是著作权人对其作品进行复制的专有权利[30]。在图书馆使用 ChatGPT 收集科研数据时,如果满足特定条件,这种行为可能会被视为著作权法下的复制行为。首先,图书馆通过 ChatGPT 收集的科研数据必须能够在某种形式的载体上呈现,例如电子文档、数据库或云存储服务等。根据《著作权法》第十条规定,复制权指的是通过印刷、复印、录音、录像、数字化等各种方式制作作品副本的权利。这意味着,任何使作品能够被复制并在这些载体上再现的行为均可视为复制行为。其次,图书馆通过 ChatGPT 所收集的数据需要在物质载体上“固定”,并达到一定的稳定性和持久性。不同于为了保证计算机程序运行而存储于计算机内存储器(RAM)的临时复制,这种“固定”意味着作品被以某种方式保存,并且不会因计算机的关机、重启或其他操作而消失。例如,当科研人员询问 ChatGPT 关于某部受版权保护的书籍的内容时,ChatGPT 可能会搜索图书馆的电子文献数据库并引用书中的部分内容作为答复。这些内容在 ChatGPT 的答复中被“固定”,并可能被科研人员下载、存储或分享。

2.2 ChatGPT 与科研数据分析:作品创作与“二次创作”过程

数据分析是信息时代决策的基石,ChatGPT 正是凭借其出色的数据分析和生成能力得以迅速应用[31]。相较于传统的数据查询和检索方式,ChatGPT 具有显著的优势:它不是仅仅从固定的数据库中抽取和复制数据,而是利用其基于深度学习的高级架构,深入解析用户的输入,并生成与查询紧密相关的全新内容,这种独特的文本生成技术为 ChatGPT 在科研数据分析中创造了新的应用机会。《自然》期刊的一项调查研究表明,ChatGPT 不仅能将复杂的科学语言转化为更易理解的内容,还能在解决特定科研问题时引入之前未被考虑的新方法或术语[32]。这不仅拓宽了研究人员对现有数据的解读,还会促使他们探索新的文献领域,从而丰富现有的研究内容。此外,在科研数据分析中,ChatGPT 通过重新解析和组织数据,能够生成与原始数据密切相关但表达方式全新的内容。例如,ChatGPT 能深入分析大量科研文献,从中提取关键信息,并基于这些信息为研究者生成针对性的摘要、引言或结论[33]。虽然关于人工智能生成物是否能被视为著作权法下的作品仍存在争论,但无可否认的是,数据分析生成的结果在形式上确实满足了作品创作的要求[34]。然而,这也带来了著作权相关的问题。根据《著作权法》规定,对原作品的改编、翻译、注释、整理或汇编等形式的活动均属于“二次创作”,并被归类为演绎行为。

综上,在图书馆利用 ChatGPT 进行科研数据分析的过程中,除了会产生全新的作品创作之外,还可能基于已有作品进行创新、拓展,从而产生受著作权法保护的演绎作品。

2.3 ChatGPT 与科研数据共享:作品的公开传播行为

如前述,图书馆可以将 ChatGPT 作为科研数据共享的核心工具,为科研人员提供交流、协作的平台。这样的平台不仅便利了科研人员分享研究经验、探讨数据相关议题,还为他们提供了一个公开展示研究成果的窗口。然而,从著作权的角度审视,以不涉及转移作品有形载体所有权或占有的方式使公众获取作品的行为,被视为作品的公开传播行为[35]。

根据世界知识产权组织《版权条约》(*WIPO Copyright Treaty*,WCT)第 8 条规定[36],采用交互式向公众提供作品的行为被明确为著作权人的专有权利,即所谓的“公开传播权”。该权利可以从以下两个方面进行解释:首先,“交互式”是指公众可以实时或按需选择、控制或更改内容的传播方式。例如,用户可对在线视频进行播放、暂停或关闭等操作。其次,“向公众提供作品”这一表述的核心意义在于,某传播行为使得公众具备了在网络途径下获取某一作品的可能性,并且也达到了“交互式”获得的状态[37]。为了符合 WCT 的相关要求,我国《著作权法》在第十条中明确将作品的网络传播行为纳入信息网络传播权的保护范围。具体来说,这包括通过有线或无线方式向公众提供作品,从而使得公众能够在个人选择的时间和地点获取该作品。本文认为,当图书馆利用 ChatGPT 为科研人员提供数据共享服务时,这种行为可能被视为著作权法下的公开传播行为,其原因在于:首先,在科研数据共享服务中,科研人员可以实时与 ChatGPT 互动,按需查询、获取或修改科研数据,这种互动方式与 WCT 所描述的“交互式”传播方式高度一致。其次,图书馆通过在线数据库进行科研数据共享,使科研人员或其他用户可以在任何时间、任何地点通过网络访问这些数据,这实际上满足了“向公众提供作品”的法律要件。

3 图书馆利用 ChatGPT 开展科研数据服务所面临的版权风险

3.1 图书馆传统版权许可方式的固有风险

著作权法以赋予创作者财产权与人身权作为激励手段鼓励文化创新,通过保障创作者对其作品使用的控制权,使其从作品利用中获得经济

利益。因此,著作权法一直遵循“先授权、再使用”的原则。然而,随着 Web2.0 的兴起,用户在内容创作、合作和分享方面发挥了越来越重要的作用,导致网络上的版权作品数量迅速增加。这种开放的创作环境虽然促进了内容的快速传播,但也增加了确定版权归属和获取版权许可的难度。如今我们步入以智能化为核心特征的 Web3.0 时代,新技术正不断塑造新的作品创作和传播模式,这无疑为传统版权许可模式带来了前所未有的挑战[38]。在这个变革过程中,传统的版权授权方式可能无法适应新的形势,同时,著作权保护与采用 AI 等新技术的行业之间的冲突可能会逐渐加剧。

首先,ChatGPT 可能会在未经授权的情况下使用作品。一般而言,图书馆在其核心服务如图书采购、读者借阅和数字化工作中,会通过购买或订阅的方式获得书籍、期刊、数据库和其他资料的使用权。为了确保这些资源被合法利用,图书馆会事先与供应商签订许可协议,明确作品的使用权限、用户访问权及服务提供范围。这种传统的版权许可策略旨在与著作权人预先明确作品的使用范围、方式和条件,从而最大程度地避免侵权风险。然而,当 ChatGPT 被应用到图书馆科研数据服务中时,可能会引发一些新的版权挑战。尤其是在 ChatGPT 生成答案时,它无法总是严格遵守图书馆与著作权人的许可协议。例如,作为一种自适应学习的人工智能,ChatGPT 可能会不当引用、节选或以创新方式呈现受版权保护的内容。这种行为有可能超出了原许可协议的约定,从而带来潜在的版权侵权风险。

其次,ChatGPT 可能未经授权许可进行数据爬取。支持机器学习,通常需要大量数据,而数据爬取作为一种常见的数据获取方式,涉及到大量的内容复制。“如果输入的数据中包含着大量未经著作权人授权使用的作品,那么这种行为可能会构成对著作权人复制权的侵犯”[39]。现实中,AI 数据爬取已经引发了版权争议。在“Andersen 等人集体诉 Stability AI 公司案”中[40],画家 Andersen 和其他艺术家发现他们的作品被 Stability AI 公司用于 AI 技术的训练,但该公司并没有得到这些艺术家的明确授权或为此支付费用。在审理过程中,法院根据 Andersen 在公开网站的搜索结果确认其作品被包含在训练数据集中。尽管部分索赔被驳回,法院仍认为现有证据足以支持继续审理对 Stability AI 的直接版权侵权指控。

最后,ChatGPT 在未获得授权情况下可能会生成侵犯版权的作品内容。著作权法中的“接触+实质性相似”原则是判断侵权行为的关键标准。虽然 ChatGPT 在创作新内容时可能并未直接使用某个特定的原创作品作为训练数据,但只要它有可能间接地接触或了解该作品,并产生了与之高度相似的内容,那么就可能会面临版权侵权的风险。例如,在“作家 Mona Awad、Paul Tremblay 诉 OpenAI 公司案”中[41],作家 Mona Awad 和 Paul Tremblay 发现 ChatGPT 能够生成与他们的书籍摘要极为近似的内容。虽然两位作家不能直接证明 OpenAI 公司将他们的作品纳入了数据训练范围,但是可以推断出 OpenAI 公司可能从公开的在线资源中间接获取了相关作品信息,并据此生成了与原作品“实质性相似”的内容。而 OpenAI 公司是在未经他们授权的情况下使用了这些相关信息,也没有标明内容的来源或支付相关费用。因此,两位作家将 OpenAI 公司告上法庭,声称该公司侵犯了他们的作品复制权、演绎权①。

3.2 图书馆版权合理使用条款的潜在风险

利益平衡是著作权法制度设计的关键价值

① 该案仍在审理中,尚未有审判结果。目前,该案件已与关联案件合并处理,并计划进行进一步的法庭调查。参见 Tremblay v. OpenAI, Inc., 3:23-cv-03223-AMO(N. D. Cal. Feb. 16, 2024)及 Two OpenAI book lawsuits partially dismissed by California court[EB/OL]. [2024-04-25]. https://www.theguardian.com/books/2024/feb/14/two-openai-book-lawsuits-partially-dismissed-by-california-court.

要素,而“合理使用”是著作权法中平衡著作权人利益与公众利益的重要制度规范。合理使用是指在特定情况下,法律允许他人自由使用作品而不必征得著作权人许可的合法行为[42]。为实现文化保护与知识传播,各国著作权立法普遍为以图书馆为代表的公共文化机构设置了合理使用条款。例如,美国《版权法》明确规定图书馆在未获得著作权人许可的情况下,可以为了替换已损坏、丢失或被盗的作品而制作复制品[43]。我国《著作权法》也将图书馆“为陈列或者保存版本的需要,复制本馆收藏的作品”的行为认定为著作权合理使用。此外,我国《信息网络传播权保护条例》还进一步明确,图书馆在特定情况下,如作品已损坏、丢失或其存储格式已过时,且市场上难以购买或价格过高时,可以数字化方式复制该作品,无需著作权人的许可。可见,这些立法主要针对“图书的保存需求”而为图书馆设定了合理使用条款。但在人工智能时代,这种法律框架可能无法充分满足图书馆对版权作品的数字化利用需求,也无法呼应图书馆朝向智能化发展的必然趋势[44]。因此,近年来,图书馆界开始对现有的合理使用条款提出质疑,并呼吁进行相应的法律改革,建议增设图书馆作品使用著作权例外条款,即“出于科学研究或其他合理目的,可以在必要限度内使用已经合法接触的作品开展信息分析”[45]。

从我国法律修订的实际情况来看,2020 年修订通过的《著作权法》中与图书馆合理使用密切相关的第二十四条未作大幅修改,并未就此问题作出学界期待的回应。由此产生的问题是,ChatGPT 在图书馆科研数据服务中的复制行为、演绎行为、公开传播行为都难以纳入合理使用范畴,这会给图书馆带来多方面的版权侵权风险,具体而言:ChatGPT 对科研数据进行大量复制时,可能会触及著作权人的复制权;ChatGPT 对科研数据进行分析并生成新的内容时,产生的创新性表达可能被认为侵犯了原作者的演绎权;若 ChatGPT 被用于在线共享科研数据,可能会侵犯著作权人的公开传播权。

4 ChatGPT 应用于图书馆科研数据服务的版权风险防范

4.1 国外科研数据使用版权例外的立法改革及启示

随着人工智能和机器学习技术的迅速发展,图书馆服务在版权领域遭遇了前所未有的挑战。为了有效地应对这些问题,许多发达国家和地区已经重新审视并调整了关于科研数据使用版权例外的制度、政策。

(1)日本为数据使用制定较为宽松的著作权例外条款。在 2009 年的《著作权法》修订中,日本在第 47 条中首次明确规定,合法获得作品的实体可以出于计算机信息分析的目的进行复制或改编。值得注意的是,这一规定并未对计算机信息分析是否具有商业目的进行限制[46]。2018 年再次修订《著作权法》时,日本立法者认为计算机信息分析的概念局限于“统计分析”范围,会对人工智能深度学习所依赖的几何分析或代数分析的运用造成法律限制,从而阻碍人工智能产业发展。为此,修订后的第 47 条用“为了提供新的知识或者信息”替代了“计算机信息分析”的表述。“为了提供新的知识或者信息”是指“开展信息处理的人(包括执行部分任务的人员)可以对他人作品进行必要限度内的复制和向公众提供,并将整理后的作品在实现此目的的必要限度内向公众提供,但同时强调该复制和向公众提供行为不得不合理损害权利人的利益”[47]。根据这一修订,日本将因数据使用引起的作品复制、演绎和公开传播行为都纳入了著作权例外条款的适用范围,这反映了日本在数据使用著作权例外问题上的法律政策相对宽松和开放。修订后的日本《著作权法》为数据使用以及人工智能的研发活动提供了更广阔的自由空间,以支持其持续创新和发展。

(2)英国引入针对非商业目的的数据使用版权例外。英国在 2014 年对《版权、设计与专利法 1988》进行了修订,在第 29 条 A 款规定,只要公众通过合法形式获取了作品,他们就可以为非商

业目的进行数据使用。在这种情况下,与数据使用相关的复制行为不会被视为侵权[48]。

(3)法国于2016年对《知识产权法典》进行了调整,修订后的第L.342-3条款明确规定,科研人员在非商业背景下使用其合法获取的科研数据和相关数据库属于著作权例外[49]。

(4)为了更好地满足数字化时代下科学研究对数字作品和数据库的使用需求,德国在2018年对《著作权与邻接权法》进行了修订。该法第44条b款和第60条d款规定,科研机构为非商业研究目的进行数据使用,被视为著作权例外。此外,为了促进科学合作或评估科研质量,将数据使用的结果传播给特定的第三方也被纳入了这一例外范围[50]。

(5)欧盟为促进科研活动统一设定了数据使用的著作权例外条款。欧盟成员国拥有丰富的数字资源却受到著作权法的限制,科研人员往往面临着需要向出版商申请许可以获取付费内容进行研究的困境。为解决这一问题,2019年欧盟正式推出并实施了《数字化单一市场版权指令》(*Directive on Copyright in the Digital Single Market*),该《指令》第3条明确规定了科研数据使用的著作权例外。根据这一规定,科研机构和文化遗产机构在进行科学研究(包括兼具商业目的的研究)时,可以复制和提取数据,而无需获得原作品著作权人的明确许可[51]。

综上,随着对科研数据在现代科研和技术发展中关键作用的认识不断加深,很多国家和地区已经对其版权法进行了修订,为科研数据的合法利用创造更多空间。虽然各国在这些条款的具体细节和覆盖范围上有所不同,但一个普遍的趋势是,将非商业性的科研数据使用视为著作权例外,从而更好地促进知识的创新和传播。

4.2 ChatGPT应用于我国图书馆科研数据服务的版权法进路

近年来,我国图书馆界和法学界对于是否应将数据使用纳入著作权例外进行了深入的研究和讨论。尽管学者们普遍支持将科研目的下的数据使用行为视为著作权例外,但在实施细节上存在分歧。一种观点是,我国应参考欧盟的立法经验,在《著作权法》的修订中引入数据使用的著作权例外条款[52]。另一种观点认为,对著作权合理使用制度进行全面的修订,从数据使用的主体、目的、客体和行为四个维度出发,构建一个完整的规范体系[53]。还有学者提议,我国可以借鉴美国的合理使用立法模式,明确数据分析的“合法要素”,并通过对具体案例的判断来确定数据使用的合法性,从而使我国的合理使用制度更为开放和灵活[54]。上述改革建议都是在《著作权法》进行第三次修订期间提出的。鉴于我国《著作权法》经过十多年的修订刚刚完成,短期内再次进行修订的可能性较小。因此,持续探讨如何改革著作权法例外制度可能并无助于解决当前ChatGPT在图书馆科研数据服务中的版权问题。目前,有两种方案可以应对ChatGPT应用于图书馆科研数据服务中的版权风险,具体可以分为两个实施阶段。

第一阶段,在司法实践中对《著作权法》中合理使用条款的适用范围做扩张解释。根据《著作权法》第二十四条第二款规定,为介绍、评论某一作品或者说明某一问题可以适当引用他人已经发表的作品。这里的“介绍、评论作品”意味着对原作品进行新的价值解读,而“说明问题”则强调基于原作品的新创作。法律中的“适当”不是仅指作品使用数量上的限制,而应当解释为作品使用方法上的适当性,由此可为科研数据的合理使用提供法律依据[55]。因此,当图书馆利用ChatGPT提供科研数据服务时,以下情形应被纳入“介绍、评论作品”或“说明问题”的合理使用范畴:首先,如果图书馆的科研数据服务通过ChatGPT为用户提供新的洞见、建议或结论,对原作品的再利用增加了新的价值,则该行为应被视为“介绍、评论作品”的合理使用。其次,如果图书馆的科研数据服务旨在支持、证实或反驳某一观点,或进一步深入探讨某议题,并适当地使用ChatGPT进行作品的再创作,那么这种对原作品的再创作属于“说明问题”的合理使用。此外,在司法实践中,应积极探索科研数据使用著作权例外的法律边界,不断明

确图书馆科研数据服务的合法范围,逐步增加“第二十四条第二款”法律适用的稳定性和可预测性,为图书馆科研数据服务的智能化发展提供一个更清晰、更公平的司法环境。

第二阶段,借《著作权法实施条例》修订之机,增设人工智能科研数据使用著作权例外。有学者指出,《著作权法实施条例》仍在修订过程中,这为我们提供了一个良好的契机来明确人工智能在科研数据使用中的法律地位[28]。增设人工智能科研数据使用著作权例外具有双重意义:首先,为科研人员的工作提供充足的法律保障,使他们在大规模数据使用时免于著作权侵权风险,从而促进知识的自由流动和创新;其次,为图书馆等公共文化机构的服务提供法律支持,使其能够更好地利用人工智能工具,不断创新服务方式。结合现有的《著作权法》规定,笔者建议对相关制度做如下设计:当图书馆、档案馆、纪念馆、博物馆、美术馆等公共文化机构出于科研目的采用人工智能工具处理数据时,属于著作权例外情形,但数据使用行为不应妨碍原作品的正常使用,也不应损害著作权人的合法权益,同时还应适当注明原作品的版权信息。

5 结语

在技术快速发展的当下,自然语言处理技术已经取得了突破性的成果。其中,ChatGPT 作为一种领先的自然语言处理工具,因其卓越的问题解决能力而备受瞩目,在各领域不断开辟新的应用途径。在图书馆领域,将 ChatGPT 与科研数据服务相结合展现出了前所未有的价值和潜力,但在实现技术融合的同时也带来了版权方面的挑战。一些发达国家已经认识到这一点,并开始调整其科研数据使用版权例外制度,以适应人工智能的快速发展。我国应对相关改革做法树立正确的认识,积极研究制定符合我国著作权立法形势的风险应对方案。具体而言,应考虑扩大司法实践中图书馆合理使用条款的适用范围,并在《著作权法实施条例》的修订中增加人工智能科研数据使用著作权例外。上述措施可以为图书馆科研数据服务提供更为明确和宽松的法律环境,有效避免因 ChatGPT 融合应用而导致的版权风险。

参考文献

1 The New York Times. The Brilliance and Weirdness of ChatGPT[EB/OL]. [2023-07-26]. https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html.

2 ChatGPT 是通用人工智能的奇点,强人工智能的拐点[EB/OL]. [2023-07-26]. https://finance.cnr.cn/ycbd/20230224/t20230224_526164037.shtml.

3 姜华,等. 生成式 AI 在教育领域的应用潜能、风险挑战及应对策略[J]. 现代教育管理,2023(7):66-74.

4 秦艳华,等. 媒介可供性视角下生成式人工智能 ChatGPT 应用于出版业的对策研究[J]. 出版与印刷,2023(3):20-30.

5 吴若航,茆意宏. ChatGPT 热潮下的图书馆服务:理念、机遇与破局[J]. 图书与情报,2023(2):34-41.

6 张慧,等. AI 2.0 时代智慧图书馆的 GPT 技术驱动创新[J]. 图书馆杂志,2023(5):4-8.

7 郭亚军,等. ChatGPT 赋能图书馆智慧服务:特征、场景与路径[J]. 图书馆建设,2023(2):30-39,78.

8 沈奎林. ChatGPT 的进击及其对图书馆的影响[J]. 大学图书情报学刊,2023(4):10-17.

9 蔡士林,杨磊. ChatGPT 智能机器人应用的风险与协同治理研究[J]. 情报理论与实践,2023(5):14-22.

10 周旭. 机遇与挑战:ChatGPT 普及背景下图书馆的应对分析[J]. 图书馆,2023(6):34-41,48.

11 张强,等. ChatGPT 在智慧图书馆建设中的机遇与挑战[J]. 图书馆理论与实践,2023(6):116-122.

12 The Fourth Paradigm[EB/OL]. [2023-10-08]. https://www.microsoft.com/en-us/researc

h/wp-content/uploads/2009/10/Fourth_Paradigm.pdf.

13 Coates H L, et al. How Are We Measuring Up? Evaluating Research Data Services in Academic Libraries[J]. Journal of Librarianship & Scholarly Communication, 2018, 6(1): 1-33.

14 陈媛媛. 高校科研数据管理服务能力研究[J]. 情报杂志, 2020(6): 203-207.

15 IFLA. ChatGPT in Libraries? A Discussion[EB/OL]. [2023-10-08]. https://blogs.ifla.org/cpdwl/2023/05/14/chatgpt-in-libraries-a-discussion.

16 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 10-13.

17 Schmidhuber J. Deep Learning in Neural Networks: An Overview[EB/OL]. [2024-04-25]. https://arxiv.org/abs/1404.7828v4.

18 Ghojogh B, Ghodsi A. Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey[EB/OL]. [2023-10-08]. https://osf.io/preprints/m6gcn.

19 Floridi L, Chiriatti M. GPT-3: Its Nature, Scope, Limits, and Consequences[EB/OL]. [2024-04-25]. https://link.springer.com/article/10.1007/s11023-020-09548-1.

20 Budzianowski P, Vulić I. Hello, It's GPT-2——How Can I Help you? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems[EB/OL]. [2023-10-08]. https://arxiv.org/abs/1907.05774.

21 尹怀琼, 等. 国内外高校图书馆科研数据管理服务分析及启示[J]. 图书馆学研究, 2020(11): 33-41.

22 俞德凤. 哈佛大学图书馆科研数据管理服务实践与启示[J]. 图书馆, 2021(10): 47-53.

23 Kim J. Data Sharing and Its Implications for Academic Libraries[J]. New Library World, 2013, 114(11/12): 494-506.

24 王晓鹏. 剑桥大学科研数据管理实践及启示[J]. 图书馆, 2022(9): 47-52.

25 复旦大学社会科学数据平台[EB/OL]. [2024-04-25]. https://dvn.fudan.edu.cn/home/index.jsp.

26 北京大学开放研究数据平台[EB/OL]. [2024-04-25]. https://opendata.pku.edu.cn.

27 Lund B D, Wang T. Chatting About ChatGPT: How May AI and GPT Impact Academia and Libraries?[J]. Library Hi Tech News, 2023, 40(3): 26-29.

28 张惠彬, 肖启贤. 人工智能时代文本与数据挖掘的版权豁免规则建构[J]. 科技与法律(中英文), 2021(6): 74-84.

29 钱力, 等. ChatGPT的技术基础分析[J]. 数据分析与知识发现, 2023(3): 6-15.

30 王迁. 知识产权法教程[M]. 北京: 中国人民大学出版社, 2014: 130-131.

31 Ray P P. ChatGPT: A Comprehensive Review On Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope[EB/OL]. [2023-10-08]. https://www.sciencedirect.com/science/article/pii/S266734522300024X.

32 Stokel-Walker C, Van Noorden R. What ChatGPT and Generative AI Mean for Science[J]. Nature, 2023, 614(7947): 214-216.

33 王一博, 等. AI生成与学者撰写中文论文摘要的检测与差异性比较研究[J]. 情报杂志, 2023(9): 127-134.

34 杨利华. 人工智能生成物著作权问题探究[J]. 现代法学, 2021(4): 102-114.

35 王迁. 著作权法中传播权的体系[J]. 法学研究, 2021(2): 55-75.

36 世界知识产权组织版权条约[EB/OL]. [2024-04-25]. http://treaty.mfa.gov.cn/Treaty/web/detail1.jsp?objid=1531876075988.

37 闫宇晨. 新闻聚合平台深度链接技术法律规制研究[J]. 新闻与传播评论, 2023(4): 40-47.

38 曹博. 著作权法如何应对Web3.0挑战: 以视听内容为样本[J]. 东方法学, 2023(3): 85-97.

39 徐小奔,杨依楠.论人工智能深度学习中著作权的合理使用[J].交大法学,2019(3):32-42.

40 Loeb & Loeb LLP. Andersen v. Stability AILtd.[EB/OL].[2024-04-25]. https://www.loeb.com/zh-hans/insights/publications/2023/11/andersen-v-stability-ai-ltd.

41 Authors File A Lawsuit Against Open AI For Unlawfully "Ingesting" Their Books[EB/OL].[2023-07-23]. https://www.theguardian.com/books/2023/jul/05/authors-file-a-lawsuit-against-openai-for-unlawfully-ingesting-their-books.

42 吴汉东.知识产权法[M].北京:中国政法大学出版社,2004:89.

43 柴会明.图书馆数字资源长期保存过程中复制行为的法律边界研究[J].图书情报工作,2020(20):46-53.

44 闫宇晨.我国智慧图书馆文本数据挖掘侵权风险与对策研究[J].国家图书馆学刊,2022(1):106-113.

45 王文敏,高军.人工智能时代图书馆信息分析的著作权例外规则[J].图书馆论坛,2020(9):60-68.

46 董凡,关永红.论文本与数字挖掘技术应用的版权例外规则构建[J].河北法学,2019(9):148-160.

47 张金平.人工智能作品合理使用困境及其解决[J].环球法律评论,2019(3):120-132.

48 Copyright, Designs and Patents Act 1988[EB/OL].[2023-07-13]. https://www.legislation.gov.uk/ukpga/1988/48/section/29A.

49 France: Law No. 2016-1321 of October 7, 2016, for a Digital Republic[EB/OL].[2023-07-13]. https://www.wipo.int/news/en/wipolex/2016/article_0014.html.

50 The Federal Ministry of Justice. Act on Copyright and Related Rights[EB/OL].[2023-10-08]. https://www.gesetze-im-internet.de/urhg/__60d.html.

51 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights In the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.)[EB/OL].[2023-10-08]. https://eur-lex.europa.eu/eli/dir/2019/790/oj.

52 阮开欣.欧盟版权法下的文本与数据挖掘例外[J].图书馆论坛,2019(12):102-108.

53 杨娟.文本与数据挖掘合理使用例外规范的体系化设置[J].图书馆论坛,2020(4):141-150.

54 林秀芹.人工智能时代著作权合理使用制度的重塑[J].法学研究,2021(6):170-185.

55 闫宇晨.科学数据出版著作权适用机制研究:保护、归属与利用规则[J].出版发行研究,2023(1):36-42.

(闫宇晨　副教授　安徽工业大学)

收稿日期:2023-09-01