【理论·探索】

生成式人工智能训练数据使用的著作权困境及其破解*

●谢星辰　宋　尧　李可心
最高人民法院知识产权司法保护理论研究(湘潭大学)基地,湘潭,411105

[摘　要]生成式人工智能训练数据作为技术创新的基础性资源,其合规使用对推动算法优化与产业迭代具有战略意义。然而,传统著作权框架下的授权使用、合理使用、法定许可等规则已捉襟见肘,生成式人工智能的海量数据需求与现行著作权制度形成冲突,并演变为制约人工智能产业创新的法律桎梏。文章通过规范分析与比较,详细阐述了生成式人工智能训练数据使用的著作权困境及原因;基于对美欧日制度实践的批判性考察,提出建构我国生成式人工智能训练数据著作权例外制度的三重路径:一是重构合理使用规则,将"信息分析型使用"纳入豁免范围并确立"无市场冲突"判断标准;二是创新准法定许可制度,通过"公告+异议排除"机制建立弹性授权路径;三是探索著作权集体管理组织路径,构建"默认许可+精准分润"的规模化授权体系。以消解权利保护与产业发展之间的矛盾,避免制度遏制创新,防止创新侵蚀权利。

[关键词]生成式人工智能　训练数据　授权使用　合理使用　法定许可
[中图法分类号]D923.4　　[文献标识码]A　　[文章编号]1003-7845(2025)02-0006-07
[引用本文格式]谢星辰,宋尧,李可心.生成式人工智能训练数据使用的著作权困境及其破解[J].高校图书馆工作,2025(2):6-12.

引言

在数字技术与内容产业深度融合的当下,ChatGPT、DeepSeek等生成式人工智能(Generative Artificial Intelligence,GenAI)通过对海量文本、图像等数据的深度学习,重塑知识生产与传播范式。党中央、国务院高度重视GenAI的发展,习近平总书记指出,"加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题"[1]。为了促进GenAI的健康发展和规范应用,我国已陆续出台《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》等政策文件。

伴随GenAI技术的突破性发展,其训练过程对大规模作品集合的依赖性不断增强,由此触发的著作权合规争议演变为制约行业发展的法律障碍。从汤森路透起诉Ross侵权案,到《纽约时报》起诉OpenAI与微软,再到环球音乐出版集团起诉技术开发公司Anthropic,系列案件揭示权利持有人和GenAI开发商正进行着火热的权利争夺战。我国学者围绕这一问题,已展开诸多探索,尤其注重解读GenAI训练数据合理使用的规则适用[2-5],或是GenAI训练数据使用对我国著作权法产生的冲击与应对[6-10],抑或是以域外为鉴讨论我国著作权法可能的调试方向[11-14]。尽管现有研究为理解GenAI训练数据的著作权争议提供了多维视角,但技术迭代与法律规制之间的张力仍在持续加剧,如何在维护权利人利益的同时促进GenAI产业发展,依然没有得到回答,而这需要著作权例外制度予以纾解。本文将从这一视角出发,分析GenAI训练数据使用的著作权困境,讨论授权使用、合理使用与法定许可面对新兴技术发展的制度窘境,并梳理美国、欧盟、日本等国家/地区的著作权例外对我国的启示,进而构建我国GenAI训练数据使用的著作权例外制度,为解决GenAI领域数据获取与著作权冲突提供新方案。

1　生成式人工智能训练数据使用的著作权困境

现行著作权制度在应对技术驱动的数据或作品规模化利用时,面临授权路径模糊、合理使用边界不清、法定许可适配困难等多重阻滞。

1.1　授权使用的困境

我国著作权法规定,使用受著作权保护的作品,

*本文系国家社会科学基金重大项目"总体国家安全观下产业知识产权风险治理现代化研究"(项目编号:21&ZD204)的研究成果之一。
作者简介:谢星辰,硕士,研究方向为数字版权;宋尧,硕士,助理研究员,研究方向为知识产权,通信作者,E-mail:songyao1108@126.com;李可心,硕士研究生,研究方向为人工智能知识产权。
收稿日期:2025-02-06　　责任编辑:谢伟祯

必须履行事先授权与支付对价的法律义务。这使得以个案协商为基础的传统授权框架,无法有效适应基于巨量数据处理的技术研发需求。

1.1.1　须获得何种著作权授权尚存争议

目前,学界对 GenAI 使用训练数据需获得何种著作权(如复制权、改编权等)授权尚未形成共识。

复制权争议聚焦数据输入阶段对作品的数字化处理是否构成著作权法意义上的“复制”。支持者认为,我国著作权法第 10 条已明确将“数字化”纳入复制权范畴,GenAI 开发者未经许可将作品转码为机器可读格式的行为,实质上固定了作品的表达形式,构成复制权侵权[15]。然而,反对观点指出,我国立法未将“临时复制”纳入复制权控制范围,训练数据在模型中的短暂存储仅服务于技术目的,属于“技术性中间过程”,不构成实质性复制[16]。

在模型训练阶段,改编权的界定同样存在分歧。部分学者主张,GenAI 通过分析作品间的关系重构表达形式,若输出内容保留原作品核心表达,则可能构成对原作品的“改编”[17]。然而,反对观点认为,改编须具备“独创性改动”,而数据训练仅提取规律,未对原作品进行主观创作性修改,故不构成改编权侵权[18]。美国版权局发布的《版权与人工智能》提出折中立场,针对 GenAI 对艺术风格的模仿行为,可通过反不正当竞争法等其他法律路径实现权益维护;但若输出内容包含“可识别的原作品片段”,则可能侵犯原作者改编权[19]。

上述争议无疑将导致 GenAI 研发方既无法准确识别需向著作权人申请的权利类型,也难以构建清晰合规的授权路径,实质上形成了“授权标的无锚可定”的困境。

1.1.2　事先授权原则难以适用

授权许可模式是指数据使用者通过支付使用费的方式,获取著作权人的个别授权[20]。GenAI 呈现出显著的分散性数据利用特征,其效能实现依赖海量参数与超大规模数据库支撑,而非对个别作品的集中性使用[21]。因此,“先授权后使用”原则在 GenAI 训练数据领域将面临难以适用的现实困境。

其一,权利主体确认难。在技术层面,部分训练数据源自网络爬虫的自动采集,原始权利人信息存在天然缺失[22]。即使完成权利人识别,也可能因为权利人数量庞大而大大降低授权效应。如在一部电影生产过程中,所有参加创作的人,包括编剧、导演、音乐人等主体,都有可能成为法律意义上的作者[23]。这种大量授权主体带来的时空地域差异、沟通协调成本,将大大降低授权程序的完成度,更遑论权利溯源困境导致的法律授权链断裂问题。

其二,授权成本高昂。GenAI 训练需处理的数据规模远超传统授权机制承载力,逐一获得授权的成本极高。从时间成本来看,数据需求方对于相关作品的授权获取并非简单的“发出要约、达成合意”[10],数据需求方需要与版权主体展开法律条款的多轮协商,不可避免地降低了数据获取的整体效率。从金钱成本来看,精准定位权属人,本身就增加了数据收集成本,确定作品与权利人后,还需就授权价格、范围反复协商,若遇权属不清的情况,成本更会大幅攀升。

1.2　合理使用规则适用困境

1.2.1　规范体系中的制度缺位

我国现行法律规范体系未对数据训练过程中的合理使用问题作出具体指引,导致实践中仍需回归传统著作权法体系寻求解决方案。这一制度衔接的空白,与我国特有的“封闭式”合理使用立法模式密切相关。著作权法第 24 条明确列举的 12 类合理使用情形均难以直接适用于数据训练场景,而作为兜底条款第 24 条第 13 款的其他情形又因缺乏配套规定而形同虚设。即便通过法律解释途径尝试将合理使用适用于数据训练领域,但细究起来颇为牵强。

其一,就“个人使用”条款而言,其适用前提系基于自然人主体为满足学习、研究或欣赏之需求,GenAI 系统及研发机构对数据的使用是为了训练模型、优化算法,进而实现商业运营等目的,显然不符合主体适格要件。欧盟《数字化单一市场版权指令》(以下简称《指令》)对文本与数据挖掘(Text and Data Mining,TDM)主体的严格限定(仅限于科研机构与文化保护组织),更印证了当前法律中的“个人”概念无法扩张至 GenAI 领域[24]。

其二,“适当引用”规则要求实施行为需以介绍评论作品或阐释问题为目的,且使用程度符合“适当性”标准。在数据训练过程中对作品的利用属于“功能性消化”而非“表达性引用”,不具备明确引用目的,其大规模复制行为亦突破了“适当”的量化边界。

其三,针对“科研少量使用”条款,虽然算法训练可纳入科研行为范畴,但现行法不仅将主体严格限定于公立教育及科研机构,更附加“少量复制”的实质要件。而 GenAI 技术发展需要海量的数据支撑,来保证模型的准确性和可靠性,这种对大规模数据的需求与“少量复制”的要求形成了根本冲突。

1.2.2 司法实践中的认定困境

当前司法领域对GenAI训练行为是否适用合理使用的判定同样面临挑战。尽管相关诉讼开始出现,但既有判例或未触及核心争议,或存在特殊因素,难以形成普遍性指引。

已决案例中,“奥特曼生图案”具有典型意义,但对合理使用条款的适用未能提供实质性指导。该案中,法院虽认定被告在生成阶段存在侵权行为,但对原告要求删除训练数据的主张,以“被告为GenAI使用方非训练方”为由未予支持①。本案虽对规范生成环节具有指导价值,却未能触及训练环节的侵权认定及合理使用规则的适用问题。

全国首例AI声音人格权侵权案中,法院创造性认定未经许可使用特定配音师声音进行模型训练构成人格权侵权,但法院也指出,著作权不包含对声音开展训练后生成相应模型的排他权利[25]。该案裁判特殊性在于训练对象为具有人身属性的声音特征,与通用型GenAI所需的海量数据训练存在本质差异,因而参考价值受限。

1.3 法定许可规则适用困境

针对GenAI的作品授权困境,不少研究主张以法定许可制度规制数据训练行为,并认为这既能够提升数据获取效率,也可为版权人提供合理经济补偿[26-28]。但是,法定许可在GenAI数据训练领域是否切实有助于实现激励创作与促进技术发展的平衡目标,仍存在理论与实践层面的双重不确定性。

1.3.1 立法价值错位

在探讨法定许可制度对GenAI数据训练的适用性时,应回归制度本质的立法价值判断,而非简单将其作为矫正市场失灵的工具。当前支持论者片面强调制度优势,却未系统论证GenAI技术特性与法定许可体系的结构兼容性。以美国版权法演进为镜,为平衡市场主体的各方利益,其法定许可实为产业迭代期的缓冲机制,该制度设计凸显两大特征:其一,作为技术变革期的过渡方案,其效力具有时空局限性;其二,通过增设严格要件强化权利保护,本质上并未突破市场自治框架[29]。

尽管法定许可可暂时缓解GenAI开发者与权利人间的授权僵局,但其公共价值内核在技术应用场景中面临双重消解:在技术层面,GenAI通过数据训练形成内容生成能力,其产出品实质构成对原作的替代性供给,这与法定许可促进并保护作品传播的公共目标背道而驰;在用户层面,公众需求聚焦技术赋能的创作辅助,而非直接获取训练数据中的作品内容,导致法定许可保障公众接触权的制度功能失去作用场域[15]。这种价值目标的错位,使得强制适用法定许可无法激活作品传播效能。

1.3.2 实践效能不足

法定许可制度在解决数据训练合法性问题上存在显著局限,从经济效率视角分析,该制度规制模式虽提升作品获取效率,却难以有效化解交易成本困境。

其一,经济可行性存疑。GenAI训练所需的海量作品规模,导致法定许可费用总额远超技术研发的预期收益。由于数据使用行为具有隐蔽性特征,权利人举证难度较大,这种制度设计反而可能加剧侵权行为的隐蔽化趋势。

其二,定价机制僵化。在数据训练场景下适用法定许可制度时,需构建费用动态调整体系,该体系的复杂性源于其需在多个维度上实现动态平衡:从价值评估维度看,费用基准的确定需综合权衡模型的商业价值、研发投入、运营成本等多重参数;从市场适配维度看,高速迭代的技术变革导致数据要素市场的供需关系与价格体系持续波动。为实现上述平衡,定价者既要保持费用机制的科学性,又须具备市场敏感性[30]。

可见,我国现行著作权法框架中设置的法定许可制度,在GenAI训练数据应用领域呈现出明显的制度适配困境。由于GenAI技术研发需对数以亿计的作品进行系统性分析加工,若简单套用现有法定许可机制应对版权合规问题,将面临立法与实践双重维度的难题。

2 域外GenAI训练数据的著作权例外及其启示

GenAI训练数据著作权例外规则,已成为全球著作权治理的核心议题。美国以合理使用司法判例、欧盟借助TDM例外规则立法、日本凭借“非欣赏性”使用目的构造,形成了各自的应对之道。

2.1 美国:限制训练数据合理使用认定

美国合理使用制度对GenAI训练数据的合法性认定呈现严格化倾向。美国著作权法第107条以四要素奠定了传统合理使用判断基准,随着Campbell案戏仿行为合理使用的认定,美国合理使用判断逐步转向以是否具有“转换性”为基准[31]。尽管在“谷歌数字图书馆案”中,法院认定批量复制图书制作检索摘要属于高度转换性合理使用[32],但GenAI使用作品的模式与之存在显著差异,相关判例中也有向传统四要素回归的倾向。在“汤森路透起诉Ross侵权案”中,法院初步否定了Ross公司复

制法律摘要训练竞争性产品的合理性[33]。这表明当 AI 训练行为与权利人核心市场形成直接竞争关系时,司法实践更倾向于严格解释何为转换性,实质抬高合理使用认定的证明标准。

然而美国多起有关训练数据合理使用争议的案件尚未裁决。GenAI 训练数据是否构成合理使用尚未形成明确统一的裁判标准。目前,仅美国特拉华州联邦地区法院作出否认 GenAI 训练数据合理使用的判决,这种司法保守的态度折射出美国司法实践应对新技术带来的著作权冲击的审慎态度。

2.2　欧盟:明确文本与数据挖掘例外制度

欧盟《指令》创设的 TDM 例外制度实质上是著作权法在数字技术冲击下对利益平衡机制的适应性重构。首先,TDM 例外制度通过法定豁免突破传统授权许可限制,为 GenAI 训练数据获取作品提供了合法性基础。根据《指令》第 3 条和第 4 条,TDM 行为在符合“合法获取”“限于特定目的”等条件下可以豁免著作权侵权责任。其次,TDM 例外制度设计反映出欧盟著作权激励创新与保护私权的平衡。《指令》区分“科学研究”与“其他用途”两类例外,前者为强制性例外,后者则允许权利人通过技术措施或者合同声明保留权利[13]。此分层设计表明立法者对公共利益与私权保护的价值选择:科学研究因具有更强的正外部性,其数据挖掘自由优先于权利人的控制权;而商业性的 GenAI 开发则需要兼顾权利人的利益,允许权利人保留控制权。最后,TDM 例外制度的适用暗含对作品使用伦理风险的管控意图。指令要求 TDM 以合法获取数据来源为前提,且不得损害作品正常利用。这意味着 GenAI 研发者需要确保训练数据来源符合法律规定、合同约定等。若爬取受技术措施保护的数据库或者违反网站服务条款获取数据,仍有可能构成侵权,而不能援引 TDM 例外主张合理使用。

《指令》第 3 条和第 4 条对 TDM 例外适用的主体、客体、目的以及行为都有着明确的要求,被认为对 TDM 有着严苛的条件限制[34]。不可否认的是,TDM 例外制度赋予了 GenAI 数据挖掘合理使用作品的空间。尽管《指令》将主体划定为“科研机构和文化遗产机构”,但是其允许欧盟成员国自行决定具体机构类型,避免因主体类型僵化阻碍技术发展;另外,《指令》第 25 条还允许欧盟成员国通过或者维持更为广泛的例外规定[35],此举既在欧盟层面确立最低保护标准,又为成员国根据数字产业发展需求拓展范围预留制度接口,为 GenAI 训练数据合法使用作品构筑了动态平衡的法律基础。

2.3　日本:构建信息分析合理使用规则

在 GenAI 技术快速迭代的背景下,日本政府于 2016 年将人工智能确立为“社会 5.0”建设的核心驱动力[36]。这种战略性布局不仅体现在技术研发层面,更在法律制度层面呈现出显著创新性特征。2018 年,日本著作权法修正案的颁行标志着日本在人工智能法律规制领域取得突破性进展。该次修订亮点在于,针对 AI 机器学习过程中不可避免的作品利用行为,增设以第 30-4 条为核心、辅之第 47-4 条和第 47-5 条共 3 个条款,构建了 GenAI 训练数据合理使用制度框架。

其中,日本著作权法第 30-4 条以“欣赏性使用”和“非欣赏性使用”区分作品使用目的,前者受著作权排他权约束,后者则纳入合理使用范畴。这一区分从“使用行为是否实现作品实质价值”出发,为技术应用提供了创新性规范路径。通过“欣赏性”,第 30-4 条将 GenAI 技术应用中普遍存在的信息分析行为纳入合理使用制度安全港。

GenAI 模型对文字作品进行词频统计、对音乐作品进行波形分析、对美术作品进行像素解析等信息分析行为时,并不是为了获取情感上的体验或共鸣,属于非欣赏性使用,未触及作品本质性价值的技术利用,并不会影响著作权人的市场利益。信息分析使用作品过程中剥离了作品的形式美感与情感要素,此种处理并不具有人文内涵,不会减损著作权人的市场收益[12]。

日本著作权法第 47-4 条和第 47-5 条在第 30-4 条的基础上,将计算机信息分析过程中附随性的使用作品以及通过信息分析向公众提供结果时少量使用作品的行为作为合理使用,有效弥补了第 30-4 条的不足[12]。例如,在论文检测中,服务提供者需要将有关的文献、期刊等作为训练数据使用,所提供的检测结果报告中会使用属于他人作品的内容与论文内容进行比较[37]。提供信息分析结果时使用作品,本质是对作品文字结构、语言表达等方面的利用,属于享受性使用。此时,当使用作品表达无法援引第 30-4 条时,可依据第 47-5 条认定此时对作品表达的使用在合理范围内而构成合理使用。

需要指出的是,日本著作权法虽为 GenAI 训练数据构建合理使用框架,但该制度仍旧存在适用争议。如“少量使用”标准在 GenAI 输出内容的量化界定上存在模糊性,可能引发司法实践同案不同判的风险;再如当前日本国内认为此规定对 GenAI 过

于宽松包容,不利于著作权人保护等[38]。

3 我国 GenAI 训练数据的著作权例外制度构建

美国的合理使用规则依赖个案,目前司法不确定性强,且偏向否认合理使用的认定;欧盟 TDM 例外制度允许成员国自行转化存在规则碎片化风险,且商业性使用受制于权利人的保留条款,无法真正促使 GenAI 发挥效用;日本立法虽创新了训练数据合理使用规则,但其适用范围限于国内法,国际协作支持不足。这些制度恐难以适应全球性的 GenAI 模型训练行为。

在构建我国 GenAI 训练数据著作权例外制度时,不宜将视野局限于合理使用制度之中,而要辩证看待其优劣。合理使用可为非商业性、非竞争性的 GenAI 研究提供兜底,避免过度限制技术探索自由;而对于更深层次的数据利用效率与利益平衡问题,需要通过准法定许可与著作权集体管理组织(CMOs)的协同机制予以系统性解决。

3.1 重构合理使用规则

我国现行著作权法采取“封闭式”合理使用体系,其第 24 条所列 12 项具体情形以传统使用场景为导向,主要涵盖教育、科研、时政、司法等人类社会交往与信息传播背景下的作品使用。第 13 项兜底条款虽形式上保留合理使用的开放性,但因缺乏明确的法律标准与有效的判例支持,难以为实践中规模化的数据训练行为提供稳定预期和行为指引。随着 GenAI 技术迅速发展,该制度设计已明显滞后。因此,亟须通过立法解释、司法指引乃至制度重构,明确将“信息分析型使用”纳入合理使用体系,推动以“使用目的”为核心的适用逻辑变革。

首先,立法上应引入“信息分析目的”作为新类型的合理使用情形。在此方面,可借鉴日本著作权法第 30-4 条的成功经验。该条款明确指出,以“数据分析”为目的复制作品不构成侵权,无论使用者是企业还是个人,只要不以“供人类感知的方式”使用作品,即属于合理使用范围[12]。立法解释中应当进一步明确“欣赏性使用”与“非欣赏性使用”之间的实质区别:前者以人类感知作品之思想与美感为核心;后者则以数字技术为前提和支撑,聚焦通过作品实现某种技术功能,未触及作品的精神价值与经济利用[39]。对于后者,若不提供免责保护,将使大规模模型训练陷入灰色地带,不利于技术创新与权利人利益的均衡保障。

其次,应确立“数据利用行为不影响原作市场”作为判断合理使用的重要标准。根据《保护文学和艺术作品伯尔尼公约》第 9 条第 2 款以及《世界知识产权组织版权条约》第 10 条的基本精神,各国在设定著作权限制与例外时,应确保此类使用不得与作品的正常使用相抵触,亦不得无故侵害作者的合法权益。GenAI 训练行为的主要目的是抽取语言规则、图像结构等信息特征,往往不涉及作品本身的可感知内容,也不会影响作品原有的销售市场或许可渠道。实践中,作品用于训练模型生成与原作表达不同、功能定位不同的新型内容,难以构成对原作的市场替代。因此,以“无市场冲突”作为判定合理性的中轴标准,不仅具有国际法基础,也契合我国当前司法需求。

3.2 创新准法定许可制度

严格的法定许可制度需以法律明文规定为前提,且并未给予著作权人退出许可的自由,在 GenAI 训练数据场景下缺乏必要的灵活性,反而是“准法定许可制度”更具可适用性。

在我国著作权法框架下构建适用于 GenAI 训练数据的准法定许可制度,须立足现行法律体系,结合技术发展需求与利益平衡原则进行制度创新。相较而言,《信息网络传播权保护条例》第 9 条确立的准法定许可模式,通过“公告+异议排除”机制形成的弹性化授权路径,为破解 GenAI 训练数据著作权困境提供了可资借鉴的制度范式。

一方面,应当以《信息网络传播权保护条例》第 9 条的制度架构为基础进行适应性改造。该机制的核心在于建立“双向透明化”的作品使用规则:首先,要求 GenAI 开发者履行强制性公告义务,通过指定平台公示拟纳入训练数据的作品清单、使用目的、报酬计算标准等关键信息,设置不少于 30 日的公告期供著作权人行使异议权;其次,赋予著作权人“选择退出”的权利,对于在公告期内明确表示反对使用的作品,开发者负有删除义务;最后,确立标准化报酬支付体系,参考行业惯例制定阶梯化付酬标准,建立由 CMOs 参与的报酬收转机制。该制度设计既通过公告程序降低了权利确认成本,又通过保留退出权保障了著作权人的控制力,在促进技术发展的同时维护了创作激励机制。

另一方面,为保障该制度的有效实施,还需通过司法解释或行政法规进行配套规则的细化。在实体规范层面,应规定开发者必须建立完整的训练数据溯源系统,对已行使退出权的作品建立动态过滤机制。在程序规则层面,可授权国家版权局搭建统一的训练数据使用公告平台,制定标准化的报酬计算

指引,并建立争议快速处理通道。同时,应当设置必要的监督机制,要求开发者定期提交训练数据使用报告,接受著作权行政管理部门的合规性审查。通过构建“法定框架+行业自治+行政监管”三位一体的实施体系,既确保准法定许可制度的规范运行,又为 GenAI 技术创新保留必要的制度弹性。

3.3　探索著作权集体管理组织路径

相较于传统作品使用,GenAI 训练对作品数量、种类及使用频率的要求呈指数级上升,单纯依靠个别授权模式在实践中面临极高的交易成本和效率瓶颈。我国 CMOs 制度原本旨在解决音乐、文字、摄影等领域大量小额作品使用的授权问题,然而当前体系在应对“机器使用”“非公众展示”“非线性引用”等新兴场景时,制度功能与市场能力均面临挑战。因此,推动 CMOs 制的功能重构与制度升级,已成为实现数据合规与产业发展的关键路径。

CMOs 制可在全球制度缺位下构建更高效的解决方案。对于商业性大规模训练,强制通过 CMOs 获取授权,形成“默认许可+精准分润”的主流通路。此架构既避免欧盟 TDM 例外制度因“商业/非商业”二分法导致的制度割裂,亦克服美国合理使用个案裁判的不可预见性,为我国参与全球 AI 数据治理规则竞争提供规范性范本。其一,通过集中授权破解海量数据交易“反公地悲剧”,依托集体协商形成标准化许可费率与使用范围,避免 GenAI 企业陷入逐项谈判的高成本困境。其二,借助“默示许可+退出”机制平衡效率与私权保护,在 AI 训练数据收集中,默认已纳入 CMOs 管理的作品均允许非表达性数据利用,权利人未明确反对即视为同意;权利人可通过 CMOs 平台声明其作品禁止用于 AI 训练,CMOs 需建立实时数据过滤系统确保退出作品及时从训练集中移除。此机制既满足数据训练对优质作品的利用需求,又为权利人保留自主控制空间。

4　结语

GenAI 训练数据的著作权合规,本质上是数字时代创作生态与技术创新的价值再平衡过程。本文通过解构传统著作权制度在应对技术变革时的系统性失能,揭示出法律滞后性引发的三重著作权困境。而后,通过对域外经验的批判性借鉴,为构建适配本土国情的著作权例外制度提供了新方案。当下,GenAI 产业正处在高速发展的关键节点,法律制度的及时调整与完善,不仅关乎产业的可持续发展,更关乎我国在全球科技竞争格局中的地位。唯有在尊重创作、激励创新、保障权益的多元价值维度上寻求平衡,才能推动 GenAI 产业在法治的轨道上稳健前行,驶向技术与法律和谐共生的未来。

注释:

① 广州互联网法院(2024)粤 0192 民初 113 号。http://ahsbqj.anhuinews.com/bq/202402/t20240227_7410093.html/。

参　考　文　献

[1] 新华社. 习近平:推动我国新一代人工智能健康发展[EB/OL]. [2025-02-06]. http://cpc.people.com.cn/n1/2018/1031/c64094-30374719.html?mc_cid=2c65101867&mc_eid=86e1c4303b.

[2] 关春媛. 生成式人工智能训练版权合理使用探究:国际趋势、本土发展与规则构建[J]. 出版发行研究,2024(12):91-97.

[3] 张平. 人工智能生成内容著作权合法性的制度难题及其解决路径[J]. 法律科学(西北政法大学学报),2024(3):18-31.

[4] 刘晓春. 生成式人工智能数据训练中的“非作品性使用”及其合法性证成[J]. 法学论坛,2024(3):67-78.

[5] 魏远山. 生成式人工智能训练数据的著作权法因应:确需设置合理使用规则吗?[J]. 图书情报知识,2025(1):78-88.

[6] 张吉豫,汪赛飞. 大模型数据训练中的著作权合理使用研究[J]. 华东政法大学学报,2024(4):20-33.

[7] 黄玉烨,杨依楠. 论生成式人工智能版权侵权“双阶”避风港规则的构建[J]. 知识产权,2024(11):37-58.

[8] 吴汉东. 人工智能生成作品的著作权法之问[J]. 中外法学,2020(3):653-673.

[9] 王迁. 论人工智能生成的内容在著作权法中的定性[J]. 法律科学(西北政法大学学报),2017(5):148-155.

[10] 张平. 人工智能生成内容著作权合法性的制度难题及其解决路径[J]. 法律科学(西北政法大学学报),2024(3):18-31.

[11] 马一德,汪婷. 人工智能训练数据版权侵权风险规制:欧盟实践、本土困境与解决路径[J]. 德国研究,2025(1):82-99,150-151.

[12] 李可心,肖冬梅. 日本生成式人工智能训练数据合理使用规则及其启示[J/OL]. 图书馆论坛,1-9[2025-03-06]. https://link.cnki.net/urlid/44.1306.g2.20250224.1351.004.

[13] 包赛君,肖冬梅. 生成式人工智能训练数据的著作权法因应:欧盟版权例外规则及其对我国的启示分析[J/OL]. 图书馆论坛,1-11[2025-02-06]. https://link.cnki.net/urlid/44.1306.G2.20250115.1117.002.

[14] 张笑尘. 人工智能生成物的可版权性问题——日本经验与中国镜鉴[J]. 现代日本经济,2025(1):81-94.

[15] 曹新明,范晔. 生成式人工智能数据训练的合理使用规则研究[J]. 中国版权,2024(4):20-35.

[16] 郭德忠,张云蔚. 生成式人工智能训练数据侵权风险与法律应对[J]. 湘潭大学学报(哲学社会科学版),2024(5):78-86.

[17] 焦和平. 人工智能创作中数据获取与利用的著作权风险及化解路径[J]. 当代法学,2022(4):128-140.

[18] 李安. 机器学习的版权规则:历史启示与当代方案[J]. 环球法律评论,2023(6):97-113.

[19] United States Copyright Office. Copyright and artificial intelligence part 1:digital replicas[EB/OL]. [2025-02-06]. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Dig-

ital-Replicas-Report. pdf.
[20] 王健宇. 生成式人工智能版权补偿金制度的原理及建构[J]. 出版与印刷,2025(1):37-47.
[21] 王文敏. 人工智能对著作权限制与例外规则的挑战与应对[J]. 法律适用,2022(11):152-162.
[22] 陈锐,江奕辉. 生成式AI的治理研究:以ChatGPT为例[J]. 科学学研究,2024(1):21-30.
[23] 孙云霄. 版权制度演进与文化产业变革的关系——基于中国电影版权制度的分析[J]. 重庆社会科学,2022(11):127-139.
[24] 孙嘉宇. 数据产权:生成式人工智能训练行为版权争议的规制路径[J]. 中国编辑,2024(8):63-71.
[25] 慕宏举. 全国首例AI生成声音侵权案一审宣判[EB/OL]. [2025-02-06]. https://www.chinanews.com.cn/sh/2024/04-25/10205621.shtml.
[26] 刘友华,魏远山. 机器学习的著作权侵权问题及其解决[J]. 华东政法大学学报,2019(2):68-79.
[27] 高阳,胡丹阳. 机器学习对著作权合理使用制度的挑战与应对[J]. 电子知识产权,2020(10):13-25.
[28] 张润,李劲松. 利益平衡视角下人工智能编创使用行为的法律定性与保护路径研究[J]. 出版发行研究,2020(11):72-79.
[29] 熊琦. 互联网产业驱动下的著作权规则变革[J]. 中国法学,2013(6):79-90.
[30] 刘禹. 机器利用数据行为构成著作权合理使用的经济分析[J]. 知识产权,2024(3):107-126.
[31] 相靖. Campbell案以来美国著作权合理使用制度的演变[J]. 知识产权,2016(12):82-90.
[32] 阮开欣. 美国版权法新发展:谷歌数字图书馆构成合理使用——评作家协会诉谷歌公司案判决[J]. 中国版权,2014(1):58-60.
[33] 李律编. AI数据训练的"合理使用"——版权&反不正当竞争视角:以Thomson Reuters诉Ross Intelligence案为例[EB/OL]. [2025-03-01]. https://mp.weixin.qq.com/s/iQo_XaP5IwHK1OXm0ae4jw.
[34] Directive(EU)2019/790 of the European parliament and of the council of 17 april 2019[EB/OL]. [2025-02-06]. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:3 2019L0790.
[35] 王胤嗣. 世界首例创建数据集侵犯著作权案:如何适用"文本与数据挖掘"例外条款?[EB/OL]. [2025-02-06]. https://mp.weixin.qq.com/s/uVczdzYH3HeKfStd1442Ow.
[36] 内閣府. Society 5.0とは[EB/OL]. [2025-02-06]. https://www8.cao.go.jp/cstp/society5_0/.
[37] 郑重. 日本著作权法柔性合理使用条款及其启示[J]. 知识产权,2022(1):112-130.
[38] 新清士. 赤松健氏「画像生成AI、珍しく日本が勝つチャンス」[EB/OL]. [2025-02-06]. https://ascii.jp/elem/000/004/122/4122855/.
[39] 袁帅. 数字化背景下作品非表达性使用的著作权法应对[J]. 知识产权,2024(9):110-126.

The Copyright Dilemma in the Use of Training Data for Generative Artificial Intelligence and Its Resolution

Xie Xingchen Song Yao Li Kexin

Research Base for Judicial Protection of Intellectual Property Rights of the Supreme People's Court (Xiangtan University), Xiangtan, 411105

Abstract Training data for generative artificial intelligence (AI) functions as a foundational resource for technological innovation, and its lawful utilization holds strategic significance for algorithmic advancement and industrial transformation. However, traditional copyright frameworks—centered on authorized use, fair use, and statutory licensing—are increasingly inadequate. The enormous data demands of generative AI conflict with existing copyright regimes, resulting in legal constraints that hinder AI-driven innovation. Through normative and comparative analysis, this paper examines in detail the copyright challenges associated with using training data in generative AI and the structural causes underlying these issues. Drawing upon a critical review of practices in the United States, European Union, and Japan, the study proposes a tripartite approach to building a copyright exception framework for generative AI training data in China: first, restructuring fair use by including "analytical use of information" within its scope and establishing a "no market harm" criterion; second, developing a quasi-statutory licensing system through a "public notice plus objection exclusion" mechanism to enable flexible authorization; and third, exploring a collective management approach to establish a scalable system based on "default licensing plus precise revenue sharing." These proposals aim to reconcile the tension between rights protection and industrial development, mitigating the risk of regulatory suppression of innovation while safeguarding copyright interests.

Keywords Generative artificial intelligence; Training data; Authorized use; Fair use; Statutory license