大语言模型训练过程中作品数据复制行为的法律定性

刘自钦，赵璇
北京工业大学经济与管理学院，北京 100124

摘要
立足大语言模型训练数据复制行为的法律定性核心，梳理并对比中美欧日等国家、地区在相关法律制度、司法实践与理论层面的差异化立场，进而探索侵权判定标准的改进方向。研究得出，在现行著作权法框架内，对个人著作的表达进行的直接模仿构成著作权侵权，而对内容和思想的总结提炼在满足必要性原则和“市场替代性测试”的条件下可纳入合理使用范畴；同时明确应当将侵权判定标准从形式标准转变为功能标准，以转换性使用作为构建侵权免责事由体系的核心，改进侵权判定标准，优化数据训练著作权侵权规制路径。

关键词
大语言模型；数据复制；合理使用；著作权；复制权

中图分类号：D923.4 文献标志码：A doi：10.11959/j.issn.2096–0271.2026xxx

Legal characterization of copyrighted works data reproduction in large language model training

LIU Ziqin, ZHAO Xuan
College of Economics and Management, Beijing University of Technology, Beijing 100124, China

Abstract
Centering on the legal characterization of data copying behavior in large language model training, this paper sorts out and compares the differentiated positions of countries and regions such as China, the United States, the European Union, and Japan in terms of relevant legal systems, judicial practices and theoretical research, and then explores the improvement direction of infringement determination standards. The research concludes that within the framework of the current copyright law, direct imitation of the expression of individual works constitutes copyright infringement, while the summary and refinement of content and ideas can be included in the scope of fair use on the condition that the necessity principle and the "market substitution test" are satisfied; it is also clarified that the infringement determination standard should be transformed from a formal standard to a functional standard, with transformative use as the core of constructing the system of infringement exemption grounds, so as to improve the infringement determination standard and optimize the regulatory path for copyright infringement in data training.

Key words
artificial intelligence, data reproduction, fair use, copyright, reproduction right

0 引言

近年来，大语言模型的迅猛发展正在深刻改变人工智能技术的应用格局，更颠覆人类的生产、生活方式。大语言模型是“一种由包含数百亿个及以上参数的深度神经网络构建的语言模型”[1]。从GPT-3到GPT-4，从豆包到Deepseek，这些模型的应用量与参数量呈指数级增长，其性能不断提升，展现出强大的语言理解和生成能力，为人机交互、内容创作、信息生成等领域带来了颠覆性的改变[2]。大语言模型的训练是数据密集型的计算过程，包括海量文本数据的抓取、存储、预处理、参数优化等环节[3]。在此过程中，包括文章、书籍、图画等在内的作品数据会被存储复制并用于模型的迭代训练，这在著作权法层面产生的问题是：此类自动化、临时性的数据处理行为是否构成著作权法意义上的“复制”行为？传统著作权法中的复制权主要规制的是将作品固定在有形载体上的持久性复制行为，而大语言模型训练中的数据处理往往具有临时性、技术性和不可见的特征，现有著作权法律制度在适用上面临挑战。

当前，全球已出现多起涉及大语言模型训练数据复制的著作权纠纷。在国外，多件案件是版权人起诉OpenAI、StabilityAI、Meta、Alphabet等生成式人工智能研发企业，控诉后者未经授权使用版权人的作品训练模型。在国内，以“正版青团子”为代表的四位画师在2023年11月集体维权，指控小红书及其旗下Trik软件在开发AI绘画功能时擅自使用其原创作品作为训练数据，导致生成的图片与原作者作品存在明显相似性；爱奇艺公司在2025年1月对AI企业MiniMax提起诉讼，主张后者未经许可使用平台版权内容用于模型训练，索赔10万元赔偿。这些案件的争议焦点在于训练数据的合法来源、模型训练过程中对作品数据的复制行为（以下简称“作品数据复制行为”）的性质认定，以及该行为是否构成合理使用、法定许可使用等侵权免责事由。虽然我国《著作权法》规定为科学研究目的可以不经许可使用已发表作品，但这一例外条款是否适用于具有商业性质的大语言模型训练，仍有待探讨。随着AI产业在数据采集和使用环节的著作权合规风险日益凸显，有必要深入探讨大语言模型训练过程中的作品数据复制行为定性问题，因为这不仅关系到人工智能技术创新与著作权保护之间的平衡，也影响着未来相关立法和司法实践的发展方向。本文拟从学理争议与案例裁判规则两个维度出发，分析模型训练中数据复制行为的法律定性，在我国现有法律框架下寻求妥帖的解决方案，为促进人工智能产业发展和技术创新提供保障。

1 大语言模型训练作品数据复制行为的“使用”定性

确定作品数据复制行为是否属于著作权法意义上的“使用”，是判断该行为是否构成侵权的首要步骤。这里的“使用”可理解为对作品内容的任何形式的利用，包括阅读、存储、复制、分析等。只有该行为落入著作权法的专有控制范畴，才有必要进一步展开后续的探讨和研究。学界以思想/表达二分法为基础，提出“非表达使

用论”与“表达性使用论”这两种对立观点。

1.1 非表达性使用论

“非表达性使用”理论最初由美国学者Matthew Sag提出，是指当作品被用于技术性、非表达性目的时，虽涉及物理复制，但未实质利用作品的“表达价值”，故不应视为著作权法意义上的“使用”[4]。该理论认为，著作权保护具有二元性，应当注重保护个人作品的表达价值，即艺术性、文学性等人类可感知的创造性表达，而其中的功能价值则不受保护，例如作品作为数据载体的机器可读属性。

持非表达性使用论的学者认为，AI训练本质上是一个数学优化过程，其目的是从海量数据中提取出具备功能性的统计规律、语法结构、文体风格等。这些内容被视为不受版权保护的“思想”，而非受保护的“表达”。该观点严格区分“目的”和“手段”，承认在技术上确实发生复制行为，但强调其最终目的和结果在于提取“思想”。正如人类阅读书籍后吸收的是其中的思想和知识，而非直接复制原文句子。而AI通过复制进行学习，但最终掌握的是类似“写作风格”等思想性元素，而非具体的表达性语句。

据此，大语言模型训练过程中的数据复制行为并不构成对个人作品的实质性使用。因为该学说强调的对作品的利用，通常指是否直接涉及作品本身的表达性内容或艺术价值，而模型训练的本质是非表达性地处理和学习文本，其目的在于提取统计规律而非再现作品的独创性表达。因此即便存在技术上的复制，只要没有直接呈现或者替代原作品的表达性内容，便难以构成著作权法意义上的“使用”。

1.2 表达性使用论

“表达性使用论”为理解和界定大语言模型训练过程中的数据复制行为提供一个正面框架。该观点主张，模型从训练到输出，其核心都与受版权保护的“表达”紧密相连，而非仅仅提取“思想”。具体而言，大语言模型对作品数据的训练行为，是在重新解构人类作者已完成的“表达”。模型无法从零开始创造思想，于是学习海量作品中的表达模式，构建出能够生成类似表达的能力[5]。这种学习更像是在“理解表达的结构”，或“再现表达的形态”。这使得其复制与使用的对象超越了纯粹抽象的思想，触及了“表达”的领域[6]。最终模型输出的是小说、诗歌、代码、图画等具有表达性目的内容，即传递信息、交流思想、生成独创性内容，并非单纯地“再现”或“消费”产品本身。整个过程始于表达，也终于表达，因此其使用行为本质上是表达性的。

有学者认为，大语言模型在训练阶段对受版权保护作品的抓取、存储与学习，实质性复制并内化了作品中受版权保护的表达性内容。大量实证研究也显示，大模型对训练数据具有高度记忆能力，能够在生成内容时逐字复刻原文或输出原作核心表达。在这种情况下，模型的生成内容无疑是表达性的，并可能构成侵权[7]。从思想/表达二分法出发，模型训练也并未止步于提取不受保护的思想，还不可避免地使用了具有独创性的表达，因此该复制行为也具有明显的表达性质。

2 大语言模型训练作品数据复制行为的“复制”定性

在明确作品数据复制行为在著作权法上的“使用”属性之后，无论将其界定为

“非表达性使用”抑或“表达性使用”，一个无法回避的前提是，该行为在技术上必然涉及对受版权保护作品的复制。

2.1 侵犯复制权论

有观点认为，严格遵循著作权法的基本原理，应将作品数据复制行为纳入复制权的控制范围。理由在于，复制权的范畴不因新技术形态的发展而弱化，只要未经许可对作品予以固定并能被感知，那么这就属于著作权法规制的复制行为。

此立场与欧盟法院在Infopaq International v. Danske Dagblades Forening案的裁判形成呼应，体现对传统版权保护的坚守。欧盟法院指出，“复制行为包括将受保护作品存储在另一种介质上的行为，无论该存储是永久性还是临时性的，也无论该存储是完整的还是部分的。”[①]人工智能训练的数字化抓取、存储和预处理，无论是将作品永久存储在服务器中，还是在训练过程中产生的临时性缓存，只要构成对作品的固定，就属于复制权的规制范围。虽然训练过程中产生的临时复制行为存在技术必要性争议，但在缺乏明确法律例外规定的情况下，多数司法实践仍倾向于认定其需要受到复制权的约束。

侵犯著作权论为作品数据复制行为的侵权判断提供了分析框架：首先判断是否存在复制或演绎行为，其次审查是否获得授权，最后考察能否适用合理使用等例外规定。依此理论，未经许可的商业性AI训练活动很可能被认定为侵权。然而，这种严格立场正面临着技术创新带来的挑战，如何平衡版权保护与技术发展已成为全球著作权法演进的重要课题。

2.2 临时复制论

我国《信息网络传播权保护条例》第21条规定，网络服务提供者在提供网络接入服务、自动缓存等场景下，实施的技术性、短暂而非持久的临时复制行为可获豁免。依据英国《版权、外观设计与专利法案》（Copyright, Designs and Patents Act）第28A条规定，如果技术性、短暂的临时复制行为是结束过程中不可或缺的一部分，且其唯一目的是使作品能够被合法使用，则临时复制行为应当被排除在复制权的控制范围之外。多数法域在不直接影响著作权人经济利益的司法实践中，都对临时复制持较为宽容的态度。相反，欧盟《信息社会版权指令》（Directive on Copyright and Information Society）第5条第1款将技术性、短暂而非持久的临时复制行为，视作与永久性或稳定复制相同的受复制权控制的复制行为。美国联邦第二巡回上诉法院在Cartoon Network v. CSC Holdings案中判定，只要复制状态能够被感知或传播的时间超过短暂瞬间，即认为符合版权法对“固定”的要求，构成版权法意义上的复制行为。[②]

有学者主张，大语言模型在训练过程中对作品数据的摄取与处理，与网络服务中的临时复制具有相似的技术特征和法律属性，都是数字技术运作过程中不可避免的中间环节，不直接面向终端用户传播作品内容，因此可以参照适用临时复制的豁免规则[8]。具体而言，大语言模型的“临

①Infopaq International *v.* Danske Dagblades Forening (Case C—5/08, ECLI:EU:C:2009:465).

②See Cartoon Network *v.* CSC Holdings, 536 F.3d 121 (2d Cir. 2008).

时复制”，是指在训练期间对海量文本数据的读取、缓存和处理，继而加载到计算机内存或临时存储介质中[9]。这种临时复制具有明显的功能性特征，纯粹是以实现机器学习的技术为目的，作品数据在内存中的驻留时间通常很短，而且训练完成后不会以原始形式保留在模型中。模型最终学习到的是文本数据的统计规律和语言模式。这种更接近于计算机运行过程中技术性操作的临时复制，与复制权控制的实质性复制存在本质区别。

临时复制论试图在现有法律框架下为人工智能训练行为寻找合法性基础，但其合理性仍有待进一步检验[10]，面临如何界定“临时性”的挑战。在大语言模型训练的场景下，虽然单个训练样本在内存中的停留时间是短暂的，但整个训练过程可能持续数周甚至数月，这使得“临时性”在概念上名存实亡。更重要的是，模型参数在训练完成后确实以某种形式，将从训练数据中学习到的知识固定在有形载体中[11]。此外，临时复制论在应用于大语言模型训练时，还需要考虑复制行为的必要性和不可避免性。现代机器学习算法本质上需要将作品数据加载到内存进行处理，这种临时复制是技术实现中不可分割的组成部分[12]。若将此类技术必需的复制行为都纳入复制权的控制范围，可能会不合理地阻碍技术创新。

3 大语言模型训练作品数据复制的“合理使用”定性

即便认定作品数据复制行为在形式上落入“复制权”的控制范围，也不意味着该行为必然构成侵权。若可认定此行为属于法律规定的版权侵权例外情形，或者构成合理使用，则此行为合法。

3.1 数据限制与例外论

数据限制和例外论是大陆法系的典型模式，指所有在未经著作权人许可的情况下，可以合法使用其作品的特殊情况。该理论承认著作权人对其作品享有排他性权利，但同时也规定在某些特定情形下，他人可以不经许可、无偿或以特定方式使用受保护的作品，而不构成侵权。

欧盟《数字单一市场版权指令》（Directive on Copyright in the Digital Single Market）严格区分科学研究与商业应用的文本与数据挖掘（Text and Data Mining, TDM）。指令第3条允许研究机构和文化遗产组织在非商业性科学研究中，无需取得权利人授权，便可复制和提取合法获取的作品或数据；第4条将例外范围扩展至商业或非科学领域的文本与数据挖掘，但采取“选择－退出”机制（opt-out），即权利人可通过明确声明禁止其作品被用于文本与数据挖掘，若未作声明反对，则被视为默许。使用者必须通过授权订阅、开放许可等合法方式获取数据，不得规避技术保护措施。前者适用强制性例外，后者则允许权利人保留选择退出权，体现出欧盟试图在商业创新与版权保护间寻找平衡。“德国图片社诉Stability AI案”体现出欧盟对文本与数据挖掘的立场：与美国的合理使用原则不同，欧盟显然更注重权利保护与风险防控[13]，著作权人对其作品是否被用于AI训练，拥有更主动的控制权。企业负有更高的注意义务，需确认使用数据是否被权利人排除在文本与数据挖掘之外。

较之欧盟的“有条件例外”模式，日本的例外范围更广。日本《著作权法》第30-4条允许人工智能以“非享受使用目的”对作品进行必要复制，但明确排除娱乐性使用，前提是这不会“不当损害著作

权人利益”。判断是否构成“不当损害”，则需考察行为是否会与作品形成市场竞争，或者影响作品的潜在市场收益[14]。

我国有关部门对模型训练数据使用的立场则更侧重对输入端的管控。《网络数据安全管理条例》第18条要求大语言模型在收集网络数据时，应当考虑到对网络服务者的影响，即不得突破网站的技术保护措施、不干扰服务器正常运行；《生成式人工智能服务管理暂行办法》第7条要求训练数据来源可追溯、使用范围合规可控，体现出保护知识产权的政策立场。此外，国家数据局在2025年发布的有关自动化程序收集数据的合法途径的政策解读更遵循了“严进宽出”的特点，强调数据来源合法的合法边界，同时肯定了数据处理者对公开数据的使用权和持有权，在“不实质性替代被收集方产品和服务等的前提下对外提供数据产品”[15]。相较而言，我国既不照搬欧盟“选择-退出”的例外机制，也不采纳美国对商业训练的弹性包容，而是以知识产权尊重、数据安全、隐私保护为核心，牢牢守住数据获取、使用的合法边界。

3.2 合理使用论

合理使用是指在特定条件下，著作权人以外的主体可以不经著作权人许可、无偿使用作品的制度[16]。在我国，《著作权法》采取源自《伯尔尼公约》的“三步检测法”（限于特殊情形、不与作品的正常利用冲突、没有不合理损害权利人合法权益），将合理使用限定于该法第24条规定的十三种情形，而其中并不包括文本与数据挖掘或者人工智能领域的作品数据复制行为。学界有观点认为，AI训练使用行为不会替代原作品的市场价值，反而可能通过技术分析增加作品的传播与利用机会，故符合三步检验法的要求[17]。但反对观点认为，AI生成内容可能与原作品形成市场竞争，且大规模未经许可的使用必然损害著作权人的许可利益，已经构成对著作权人合法权益的不合理损害[18]。

在美国，基于美国《版权法》与判例法确立的合理使用“四要素测试法”（使用的目的和性质、被使用作品的性质、使用部分的数量和实质性、对作品潜在市场或价值的影响），足以认定人工智能场景下机器训练行为是否合法，该行为根据具体情形可能构成侵权或合理使用[13]。美国法院在2015年审理Authors Guild v. Google案时判定，Google公司扫描图书用于提供检索功能的行为属于合理使用，因其行为具有高度转化性，不会替代原作品市场，且能带来显著的公共利益，有利于促进文化保存与教育普及①。这一判例常被援引支持AI训练中作品数据使用的合法性[19]。但是，美国法院在2023年审理New York Times Company v. Microsoft Corporation案时，持相反立场。在该案中，原告New York Times公司起诉被告OpenAI和Microsoft公司，指控后者未经许可使用数百万篇新闻文章训练AI模型，侵犯其复制权和演绎权，并可能通过提供替代性信息产品损害《纽约时报》的市场。被告主张，其训练数据均来自公开网络，属于转换性使用，且《纽约时报》内容对模型性能影响有限。法院在阶段性裁决中虽未直接认定诉争行为是否构成合理使用，但指出人工智能输出“复现原文”可能不符合转换性标准，需进一步审查市场替代

①See Authors Guild *v.* Google, 804 F.3d 202 (2d Cir. 2015).

效应[①]。Authors Guild v. Google 案与 New York Times Company v. Microsoft Corporation 案确立一个重要标准，即使用行为对原作品市场的替代效应是判断合理使用的关键因素。

在我国，司法实践对合理使用的认定则呈现出以结果为导向的立场。以“王莘与北京谷翔信息技术有限公司、谷歌公司侵犯著作权纠纷案”为例，谷歌公司对原告的小说进行全文复制，谷翔公司则在其运营的网站上提供该作品片段，终审判决对这两种关联行为作出不同的法律认定[②]。法院作出不同判断的关键在于，对“使用行为的性质”与“对原作品市场的影响程度”进行了精细的区分。与美国的谷歌案不同，我国法院在处理此类“全文复制、片段化展示”行为时，并未全盘采纳美国以转换性使用为核心的分析框架，而是结合《著作权法》第 22 条及“三步检验法”标准，重点审查使用行为是否损害了权利人合法利益。谷翔公司提供作品片段的行为之所以不构成侵权，就是因为其展示作品的片段不会满足读者完整的阅读需求，所以对原作品市场影响有限，而且其目的在于提供图书信息检索服务，具有转换性。与此同时，谷歌公司的复制行为不利于王莘对作品的利用，且对涉案作品的市场利益造成潜在危险，被认定构成侵权。因此，我国法院在判断合理使用是否成立时，更关注该行为是否实质性替代原作品市场，对“转换性”因素的认定更为克制，尤其在商业性使用色彩明显的情况下，更倾向于否定合理使用。

这说明，人工智能训练行为在合理使用认定中正面临复杂局面，因其既有可能带来显著的公共利益，也可能对原创内容市场造成难以预测的影响。这要求法院在裁判时更加谨慎地权衡各方利益，既要避免过度保护阻碍技术创新，又要防止权利弱化损害创作激励。

以上诸理论表明，关于大模型训练中作品复制行为的法律定性仍存在显著分歧。这种分歧既源于对技术理解的不同，也反映在版权法基本理念和价值取向上的差异。解决这一困境，需要超越简单的“侵权或不侵权”的争议，构建更加精细化的分析框架，充分考虑 AI 技术的特殊性、产业发展的现实需求以及版权保护的基本宗旨。

4 现行诸理论的适用困境剖析

自《伯尔尼公约》确立复制权的基本框架以来，法律始终以“固定性”和“可感知性”作为判定复制的标准，即作品必须被“以某种物质形式固定”并可供人类直接或间接感知[20]。然而，AI 技术的迅猛发展使得这一标准变得模糊。机器学习过程中的数据摄取、临时存储、向量化处理等行为，既不完全符合传统复制的定义，又无法被现行法律体系完全豁免。非表达性使用理论、数据例外论、合理使用论均存在或多或少的缺陷，难以真正调和技术创新与著作权保护之间的矛盾。面对这一困境，必须根本性地重塑侵权判定标准，才能适应技术变革带来的新需求。

4.1 非表达性使用论简化 AI 训练的根

①See New York Times Company *v.* Microsoft Corporation, No. 23—cv—11195 (S.D.N.Y. 2023).

②参见北京市第一中级人民法院(2011)一中民初字第1321号民事判决书,北京市高级人民法院(2013)高民终字第1221号民事判决书。

本矛盾

非表达性使用论体现的是一种被动适应的改良主义思路，其局限性不仅体现在其理论上的薄弱，更反映在其与现行法律体系、技术实践和产业生态的冲突中[21]。该理论成立的前提，是作品使用行为完全剥离作品的表达性价值,例如在 Authors Guild v. Google 案中，Google 公司提供的是比较机械、生涩，功利性的搜索服务，用户无法在其中获取完整的表达性元素。但是，大语言模型的训练过程并非真正脱离原始作品的表达形式，而是通过复杂的算法保留作品的核心特征。人工智能复制数据时，看似将文字、图片转换为数字语言，但只是改变作品的呈现形式，仍旧保留着作家的文风、画家的笔触等表达性元素。人工智能系统对作品的利用方式极其特殊：一方面，机器学习的算法确实不涉及对人类可读表达的直接使用；另一方面，训练过程又必须依赖完整的作品表达作为学习素材。如果仅因技术处理使作品的呈现形态发生改变，就否认其属于法律意义上的复制行为，那么任何数字化转换都可能成为规避责任的借口，这将架空著作权的表达保护功能。

此外，非表达使用论在价值取向上存在失衡风险。该理论单方面强调技术创新的自由空间，却未能给予著作权人相应的利益补偿。若人工智能公司可以凭借“非表达性使用”的理论无限制地利用现有作品而不给予合理补偿，内容创作者要承担个人作品被无偿利用的成本，这必然会影响创作者产出的积极性。这种不合理的权利义务安排，长远来看不利于文化产业的可持续发展。

4.2 数据限制与例外论对原创作者价值分配的失衡

著作权法的核心目的是通过给予创作者专有权利以激励创作，数据例外论很大程度上会削弱这种激励机制。该理论允许科技公司无偿或低成本地使用海量由创作者投入心血和资源完成的作品，可以视为一种“搭便车”行为，剥夺权利人在其作品被用于商业训练时获得的合理报酬。长此以往，必将打击原创作者的积极性，不利于文化的创新与可持续发展。

更进一步说，这种新型的“公地悲剧”在实践中更加剧数据来源不透明与权利归属模糊的问题。训练方作为该理论的受益者，而缺乏建立数据溯源机制的意识和动力。当海量作品被不加区分地抓取混用，权利人的维权将变得异常艰难，这反过来又放纵训练方形成恶性循环。

在具体适用时，数据限制与例外论虽然展现出更强的针对性，但其适用范围与界限却十分模糊。欧盟《数字单一市场版权指令》创设的文本与数据例外，本质上是一种政策上的妥协而非法律层面的革新。这种例外模式将人工智能数据利用问题简化为权利限制问题，却回避更为根本的侵权判定标准重构的需求。结果，立法者不得不在例外条款中设置复杂的适用条件，如合法获取来源要求、opt-out 机制等。这种路径依赖既有的法律框架，使法律改革始终停留在修修补补的层面，难以实现制度性的突破。一切试图在传统复制权的逻辑下“打补丁”的行为，都无法直面技术现实，也无法为人工智能产业奠定一个合法、稳定且能够获得社会广泛支持的坚实基础。

4.3 合理使用论的适用困境

在我国现行《著作权法》框架下，将

作品数据复制行为认定为合理使用的可行路径，是将此行为纳入“个人学习”或“科学研究”范畴，将主体严格限定为自然人，目的限定为“学习、研究或者欣赏”等非商业用途。而人工智能训练数据通常由企业或研究机构开展，最终目标多与商业应用相关，明显超出该条款的适用范围。人工智能训练需要海量数据支持，远非“少量”概念所能涵盖。

这种路径忽视商业性人工智能开发的营利本质，且与多数成文法国家封闭式合理使用条款相冲突。即便在美国，如Authors Guild v. Google案中确立的片段使用的豁免，也无法直接适用于大规模、系统性复制完整作品的行为。美国合理使用原则中，虽然充分考虑了模型作品数据训练的转化性使用，但对作品使用的数量、实质性和对原作品市场的潜在影响，在各个司法实践中几乎都存在事实性争议，带来极大的不确定性。

这种不确定性在美国近期典型判例中暴露得尤为充分。2025年6月的Bartz v. Anthropic案的判决表明，合理使用原则的边界在规模化、工业化复制面前变得模糊而脆弱。该案源于Anthropic被指控从Library Genesis和Pirate Library Mirror等盗版图书网站下载超过700万本书，用于训练Claude模型，即其“蒸馏”人类知识库过程中存在非法获取数据的行为。尽管美国加利福尼亚北区联邦地区法院裁定其AI训练行为本质上具有转化性(inherently transformative)，因此属于合理使用，延续了早年Google案的逻辑[①]。但当模型训练的目的从“辅助人类检索”转向“直接生成竞争性内容”，其行为是否超出了合理使用的初衷？最终，该案以Anthropic赔偿15亿美元达成和解为结果，这笔赔偿折算下来，每本书的侵权成本远低于法定赔偿上限。这一现象传递出一个危险的信号，即对于资本雄厚的科技巨头来说，只要侵权带来的技术收益和商业回报远超潜在的赔偿金，它们就有经济动机去“先使用，后和解”，从而将法律风险内化为一种可以计算的运营成本。这实质上削弱了版权法对原创者的保护力度，迫使创作者接受不对等的条件，长远来看会大大损害现有的创作生态。

当技术的演进在速度和规模上远超法律制度的更新迭代时，我们亟须重新审视和定义数字时代知识创造、所有权、使用权和收益权的边界。否则，所谓“技术创新”可能演变为一场对既有知识公地的攫取，最终损害的是人类文化创造活动的可持续健康发展。

5 大语言模型训练阶段复制作品行为的侵权认定标准的改进

诸种理论困境共同揭示出，现有侵权判定路径已难以有效平衡技术创新与原创保护，不仅使产业陷入法律风险不可预见的灰色地带，亦使创作者权益暴露于法律漏洞之下。要确保人工智能产业在法治轨道上稳健发展、重振创作生态，就必须超越作品数据复制行为在传统法律框架下的二元定性，构建一种前瞻性、系统化且兼具国际兼容性的侵权判定新标准。

5.1 训练作品数据复制构成著作权法意义上的“复制”

我国《著作权法》第10条第1款将复制权定义为“以印刷、复印、拓印、录音、录像、翻录、翻拍、数字化等方式将作品

①See Bartz *v.* Anthropic PBC, 787 F.Supp.3d 1007 (N.D. Cal. 2025).

制作一份或者多份的权利”。构成著作权法意义上复制行为的关键，在于对作品内容的固定和再现，而不论采用何种技术手段。在人工智能训练过程中，无论是前期数据收集阶段对作品的全文保存，还是后期机器学习阶段将作品转换为机器可读格式，都涉及对作品的完整复制和系统存储。这种技术性的行为虽然可能不为普通用户所感知，但实质上已经构成对作品的“再现”，符合《著作权法》对复制行为的界定。虽然有观点认为，训练过程中的复制属于“临时复制”，不应被视为侵权[22]，但我国《著作权法》并未明确将临时复制排除在复制权控制范围之外。实践中，人工智能训练数据的存储时间往往根据技术需要而定，可能长期甚至永久保存，这与传统理解的“临时性”存在显著差异。因此，简单地将所有训练数据复制行为归为临时复制而主张免责的观点，难以获得法律支持。

从技术视角看，AI训练数据的获取通常通过多种方式实现：网络爬虫抓取、技术手段破解复制、非电子资料的数字化以及用户协议中的强制许可条款等。这些获取方式中，除合法授权的渠道外，多数都可能涉及对原作品的复制权侵权风险。尤其是出于训练目的而进行的复制，往往需要制作多份副本并长期保存，这与仅为传输、浏览等技术需要的临时存储存在本质区别。

据此，在我国现行《著作权法》框架下，人工智能训练过程中未经许可复制他人作品并用于模型训练的行为，很大可能构成对复制权的侵犯。除非存在特定的免责事由，否则此类行为将面临较高的法律风险。这一判断也与国内外近期出现的一系列司法判例所体现出的倾向基本一致。

5.2 从形式判断到功能判断的标准转换

依据《伯尔尼公约》第9条规定，复制权是以任何形式完整再现作品表达的权利，当然包括印刷、数字化、临时存储等形式。在人工智能时代，若对此定义作扩张解释，则根据技术中立原则，数据训练中的临时存储、格式转换均属“复制”；而缩小解释主张复制权应回归本质，即控制作品的传播性复制，非传播性技术过程不侵权，如单纯的训练缓存。但需要注意的是，《伯尔尼公约》第9条第1款规定，“受本公约保护的文学艺术作品的作者，享有授权以任何方式和采取任何形式复制这些作品的权利。”这里的“任何方式和采取任何形式”说明复制权并没有限定特定的复制形式，立法技术上都采取前瞻性的态度。

在技术层面，大语言模型训练过程可划分为三个关键环节，每个环节都可能涉及侵权问题。一是在作品数据收集与输入阶段，即在人工智能进行深度自主学习前，需先将个人作品数字化，转换成机器可读的标准数据格式。此过程是对已有作品的全文复制和再现，并形成永久性存储，属于著作权法上的“复制”行为，存在侵犯复制权的风险。二是在模型训练阶段，通俗地说，就是大模型提取训练数据中的信息并总结规律，此过程也叫作机器学习。虽然此过程本身不直接涉及对原始数据的复制，但训练数据来源若是未经授权抓取的个人作品，那么整个训练行为的合法性就存疑，因为其基础数据获取可能侵犯复制权。三是在作品输出阶段，根据“接触+实质性相似”的著作权侵权判定规则，若人工智能的输出内容与之前所使用的训练数据，存在表达层面的实质性相似，则可能会构成侵权。当然，这一判断还需结

合“思想/表达二分法”原则。只有当AI复制的是受保护的具体表达而非思想时，才可能产生侵权责任；如果构成实质性相似的是思想而非表达，则难以构成著作权法意义上的复制。社会各界目前一般不认为在这个阶段具体或存在侵权风险。

据此，可对复制权的形式和特性予以补充：复制权包括数字化复制，即将作品转化为数字格式存储和传播；复制权不应当控制改变作品整体表达形式的“再创作”。这一思路能被《法国知识产权法典》第L.122-5（6）条所印证，该条规定作者不得禁止作品发表后的下列复制：“过渡性或附属性的临时复制，该复制必须是某个技术方案完整和基本的组成部分，该复制仅在于允许作品的合法使用或借助中介网络在第三人之间的传播；但该临时复制仅适用于软件和数据库以外的作品，且自身不得具有经济价值。”该规定精准地把握技术必要性与权利保护的平衡点，其将“不具有独立经济价值”作为判断标准的方法值得借鉴[23]。这一思路同样适用于人工智能训练场景：当数据复制仅作为模型训练的必经环节，且未导致原作品表达进入流通领域时，将其排除在复制权控制范围之外，既符合技术中立原则，又能避免版权法成为阻碍创新的壁垒。但这种例外应当设置严格的前提条件，包括确保训练数据来源合法、防止输出内容与原作形成市场竞争等。

侵权判定标准的现代化改造，需完成从形式到功能判断的转换。未来的制度不应拘泥于数据输入阶段复制行为的技术特征，而应聚焦其生成物的市场影响：只有当数据处理行为可能导致原作品表达进入流通领域，或者人工智能输出的内容实质替代或者超过原作品的市场价值时，才可能构成侵权。这种功能主义的判断路径，既能保有著作权法激励创作的核心价值，又能为人工智能技术创新提供清晰的指引，最终实现文化繁荣与技术进步的良性互动。

5.3 免责事由的体系化构建：以转换性使用规则的适用为核心

当现行合理使用制度难以适用于人工智能训练数据场景时，法定许可制度作为一种折中方案被提出。我国《著作权法》规定的法定许可情形有限，主要适用于教育、新闻传播等具有较强公共利益的领域，而AI训练的商业属性显然更为突出。面对AI训练对于多类型作品的需求，必须协调美术、图像、声音等各领域的管理，其复杂程度远超现有法定许可制度的场景。我国著作权集体管理体系在覆盖范围、运作效率等方面仍存在不足，也难以胜任大模型多样化训练数据的大规模管理工作[24]。这表明，基于我国现行《著作权法》框架，难以主张纯粹且无争议的免责事由，更可行的途径是构建一个“安全港”规则，只要AI训练方或是使用者遵循了特定的合法流程，就可以被推定为合理使用，从而免除或减轻侵权责任。

鉴于此，转换性使用规则（transformative use）为合理使用制度的适用提供新的法理依据。转换性使用指对原作品的使用已脱离其原始表达目的，以不同方式或基于不同目的对作品进行使用[25]。其本质在于使用行为是否为原作品增添了新的表达、目的、功能或者意义，从而区分单纯的复制性使用与具有创新性的转换性使用。具体到大语言模型的场景，转换性使用的评估则体现为一种完整的链条。大语言模型在训练阶段对海量作品进行非直接复制形式的学习，并将其转化为机器可学习的数据信号；在生成阶段，模型调用这些参数，创造出了全新的、独立

的表达内容，从内容、形式等各方面都超越了既有的表达，产生具有新功能、新目的的新表达。全过程并非以传播或再现原作品表达为目的[26]，这与传统合理使用中评论引用作品片段或戏仿（parody）创作存在本质差异。由于著作权法仅保护表达而不保护思想，大语言模型的作品数据复制行为属于著作权法意义上的使用，但若该使用行为具有高度转换性，生成了具有创新价值的新内容，且不会对原作品潜在市场或价值产生负面影响，应被认定为转换性使用，在版权法上具有正当性。

在这套体系中，人工智能输入端的合规性是基石。人工智能产业应当优先使用公开可得的数据，如公共领域的知识、知识共享许可协议下的内容。借鉴美国合理使用“四要素测试法”，输入作品数据以训练人工智能的目的，应当是为了创造全新的、具有社会价值的内容或功能，无论从使用作品数据的数量、实质性相似或者对原作品市场的替代程度哪个维度，都能与原作品的“表达”层面有所区别。其次，对于人工智能训练端，应利用技术本身的特点论证其合理性。作品数据被加载到内存中进行处理的步骤是临时的、必需的。复制形成后的模型参数是一个高度抽象的、不可读的数学结构，不具有独立的经济价值，且必须在训练完成后被销毁。这种“临时复制”或许可因其必要性和临时性而受到法律豁免。最后，人工智能生成端是风险显现的重要环节，在此阶段应重点关注两点：一是输出内容与用户输入指令的关系；二是输出内容与人工智能输入端训练数据的关系。前者主要考量输出内容中用户的独创性参与程度，若仅为基于指令的回应，包含了用户的智力成果和独特选择，则侵权风险较低。后者则需过滤输出内容是否不当使用AI输入端训练数据中受版权保护的部分，确保输出内容不具有实质性复制或改编。“只要能‘管住生成端’，就可以‘豁免训练端’，即可以完全豁免模型训练中使用作品的著作权侵权责任，但允许著作权人对生成端可能出现的个案涉嫌侵权内容进行有效执法和维权。”[27]事实上，广州互联网法院对“上海新创华文化发展公司诉广州年光公司网络侵权责任纠纷案”（即“奥特曼”案）的裁判，也体现出这种思路。在该案中，被告某人工智能公司运营的Tab网站提供AI绘画服务，用户输入生成奥特曼的指令，即可生成与奥特曼形象相似的图片。法院判定，被告AI公司侵犯原告对奥特曼作品的复制权，但并未直接评判人工智能训练行为本身的合法性，而是主要聚焦于涉案图片与原告作品之间是否存在实质性相似，从而认定训练方侵犯复制权①。与此同时，在面临相似案由的情况下，杭州互联网法院审理的“上海新创华文化发展有限公司诉杭州水母智能科技有限公司侵害著作权及不正当竞争纠纷案”，并未直接认定模型训练与生成行为构成直接侵权，而是以被告未对明显具有侵权风险的生成物采取合理防控措施为由，认定其构成帮助侵害信息网络传播权，同时进一步认定构成不正当竞争②。两案裁判均认可涉案作品受著作权法保护，并要求AI公司在经营过程中承担与其技术能力、商业模式相匹配的法律注意义务。

据此，AI免责体系的核心框架由输入端、训练端与生成端的责任划分构成，其中生成端尤其需要从多维度加以规范。从版权法视角看，生成物是否构成侵权，则

①参见广州互联网法院(2024)粤0192民初113号民事判决书。
②参见杭州互联网法院(2024)浙0192民初1587号民事判决书。

主要取决于用户的指令是否具备足够的独创性。在复制行为使作品产生新的应用场景，且输出内容与原作品在市场定位上形成实质性区别时，可能构成合理使用。换言之，若输出内容是对训练数据的重组模仿,尤其是对原作品潜在市场或价值的存在负面影响，则可能构成复制权侵权，自然也不属于可豁免责任的合理使用制度情形；若输出内容为独立的新表达，则需要考查人类指令的独创性投入成分。

5.3.1 作品数据复制行为对作品潜在市场或价值的影响

作品数据复制行为对作品潜在市场或价值的影响，是判断该行为构成合理使用与否的第一项标准。从立法目的来看，为了更好地激励创作，著作权法给予创作者对自己作品的收益有限的垄断权。面对能够指数级地产出内容的新技术，生成式人工智能可能会引发“市场稀释”效应，即人工智能通过生成海量同质化内容，从而实质性挤占削弱原作市场。这种由技术更新带来的新型竞争关系，更凸显出在市场层面评估损害的必要性。判断是否构成转换性使用的关键，不仅在于使用行为是否“新”，更在于它是否动摇了原作品的经济基础。若某种作品使用行为不会对作品的市场价值构成实质性的替代或损害，则禁止这种使用反而会阻碍基于原作的创新。

法院在Kadrey v. Meta Platforms案的裁判，充分体现了这一点。在该案中，Meta公司提供了模型发布未影响销量的证据，说明其使用行为未对原作在书籍市场造成负面影响，法院据此判定Meta胜诉[①]。这表明，在人工智能领域，合理使用的边界是个动态的评估过程，而市场替代性影响是决定天平是否向己方倾倒的最后一根稻草。

5.3.2 作品数据复制行为所生成内容的独创性

作品数据复制行为所生成内容的独创性，是判断该行为构成合理使用与否的第二项标准。在著作权法律体系中，独创性概念是作品获得法律保护的核心要件，是连接创作行为与法律保护的纽带。只有当作品凝聚创作者独特的智力投入和个性表达，体现出创作者与作品之间不可分割的精神联系，才值得法律赋予排他性权利。独创性包含创作过程的独立性、成果的创造性这双重内涵[28]。

在人工智能创作场景，独创性的认定应着重于人类使用者的创造性参与程度，参与程度越高，人类对构思的贡献度越大，作品的独创性越高[29]。具体而言，使用者在提示词设计、参数调整、结果筛选等环节的个性化选择，往往能反映出其独特的审美判断和艺术构思。使用者实际上将自身的创意和构思融入创作过程，这种深度参与可能使最终作品具备充分的独创性。这就意味着，当用户对人工智能的指令足够具体时，AI在人机协同的创作过程中所起到的作用就如同摄像机之于摄像师，是人类创意、创造的载体[30]。反之，若仅仅是机械化地输入简单的指令，则难以构成版权法意义上的创造性劳动。

2023年，北京互联网法院就“李昀锴诉刘某侵害作品署名权、信息网络传播权”一案，作出了国内第一例生成式人工智能作品版权的判决。法院在该案中认定，原告通过AI工具生成的人物图片构成著作权法意义上的作品，依法享有该作品的作者身份。判决指出，AI模型本身不具备法律主体资格，而使用者在生成过程中通过构

①See Kadrey *v.* Meta Platforms, 788 F.Supp.3d 1026 (N.D. Cal. 2025).

思提示词、进行审美选择和多次修正，体现了其对最终作品呈现具有实质影响的个性化安排，因而其智力成果满足“独创性”这一要求①。这表明，在“人机协作”模式下，作品认定的核心在于人有多少独创性贡献，其中“独创性”可以体现在智力投入的过程中，只要该过程与最终成果之间存在直接的因果关系，且成果体现了使用者的个性化表达，即可认定该成果具备独创性。从我国实际出发，这无疑为未来著作权法在“作品”定义、权利归属等条款的修订提供了宝贵的司法实践基础。

无独有偶，美国法院在Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc.案中，对独创性的认定成为认定被控侵权行为构成合理使用的关键。在该案中，被告ROSS Intelligence公司在向原告Thomson Reuters公司申请Westlaw数据库被拒后，通过第三方LegalEase公司获取了基于Westlaw法律要点摘要生成的“批量备忘录”，用于训练其人工智能模型。法院指出，独创性标准要求极低，仅需作品具备最低限度的创造性标准，而原告的Westlaw法律要点摘要源于对司法意见进行提炼、筛选和编排的创造性劳动，当然构成受保护的独立作品，其关键编号系统的结构选择同样体现了独创性②。所以，被告未经许可复制摘要用于训练其竞争性人工智能工具的行为，构成直接侵权。该案判决表明，以竞争为目的对独创性内容进行实质性复制，不能获得合理使用原则的庇护，这成为AI产业使用版权数据的重要法律边界。

不仅如此，人工智能生成内容与现有作品之间的差异性，同样是判断其独创性的重要参考依据。这种差异性不仅体现在直观可感的风格特征上，还涵盖深层次的表达形式创新。法院在审理案件时，需综合考量作品的整体视觉效果、细节处理手法以及创作背景等多方面因素，方能作出精准的判断。对于那些独创性虽不足但依然具备商业价值的内容，权利人可借助外观设计专利或反不正当竞争等替代性途径来寻求保护。在人工智能领域，著作权法必须在激励创新与保护原创之间探索新的平衡点，既要防范技术滥用导致的创作同质化现象，也要为真正展现人类智慧的艺术探索留出充足的成长空间[31]。

6 结束语

人工智能技术的迅猛发展深刻重塑了著作权法的理论基础和制度框架。人工智能创作中的转换性使用理论与侵权判定标准的重塑，揭示了数字时代版权制度所面临的核心挑战。合理使用制度的现代化改造，不仅需防范技术滥用引发的权利失衡，还应为真正的技术创新预留制度空间。

AI免责体系的理论探讨为法律实践提供了分析框架，而我国有关部门正以具体行动回应大语言模型训练中的作品数据使用争议。《“十四五”数字经济发展规划》的出台和《生成式人工智能服务管理暂行办法》的实施，均体现出我国在数字治理领域的积极探索，为版权法视角下的侵权判定注入中国化的实践内涵。2026年全国两会解读访谈中，最高人民法院民三庭庭长李剑指出，法院会综合考量AI训练端的

①参见北京互联网法院(2023)京0491民初11279号民事判决书。

②See Thomson Reuters Enterprise Centre GmbH *v.* Ross Intelligence Inc., 765 F.Supp.3d 382 (D. Del. 2025).

注意义务与信息管理能力，避免抑制技术创新[32]。未来人工智能的司法治理，应当坚持守正创新的基本原则，在保护知识产权与促进技术创新之间寻求动态平衡：既要坚持保护原创的基本立场，又要适应技术发展的客观需求；既要维护著作权人的合法利益，又要促进产业创新，为AI技术创新提供合理的制度空间[33]。

AI技术与版权法的互动研究，仍有许多探索空间。在理论层面，需要进一步探讨人机协作背景下创作的本质，构建更具解释力的独创性理论。在制度层面，跨国比较研究借鉴国际经验将有助于提炼不同法域应对人工智能挑战的智慧，特别是欧盟《人工智能法案》与美国版权局最新政策指引的实践经验值得持续关注。在技术层面，区块链、数字水印等新兴技术为作品溯源与权利管理提供新的解决方案，其与法律制度的衔接机制有待深入研究。

参考文献：

[1] 张奇,桂韬,郑锐,等.大规模语言模型：从理论到实践[M].北京：电子工业出版社，2024:1.
Zhang Q, Gui T, Zheng R, et al. Large language models: from theory to practice [M]. Beijing: Electronic Industry Press, 2024: 1.

[2] 刘平,石勇,李何敏,等.“人工智能+”跨行业可持续融合与增长战略[J].大数据, 2026,12(01):4.
Liu P, Shi Y, Li H M, et al. Strategies for sustainable integration and growth of “ Artificial Intelligence +” across industries[J]. Big Data Research, 2026, 12(01):4.

[3] 舒文韬,李睿潇,孙天祥,等.大型语言模型:原理、实现与发展[J].计算机研究与发展,2024, 61(2):356.
Shu W, Li R X, Sun T X, et al. Large language models: principles, implementation and progress[J]. Journal of computer research and development, 2024, 61(2): 356.

[4] Sag M. Copyright and copy-reliant technology[J]. Northwestern University law review, 2009, 103(4): 1607-1682.

[5] 代江龙,何若楠.大模型训练数据版权侵权风险规制[J].数字图书馆论坛,2025,21(09):78.
Dai J L, He R N. Regulating Copyright Infringement Risks in Large Model Training Data[J]. Digital Library Forum, 2025,21(09):78.

[6] 卢海君.论思想表达两分法的法律地位[J].知识产权,2017,(09):22.
Lu H J. On the Legal Status of the Dichotomy of Thought and Expression [J]. Intellectual Property, 2017, (09): 22.

[7] 王诗童,杨利华.生成式人工智能机器学习的版权分层规制模式——以“表达性使用”为视角[J].编辑之友,2025,(02):82.
Wang S T, Yang L H. The hierarchical regulation model for copyright in generative ai machine learning: from the perspective of "expressive use"[J]. Editorial friend, 2025, (02): 82.

[8] 刘劭阳.论临时复制的法律性质[J].电子知识产权,2013,(Z1):110.
Liu S Y. On the legal nature of temporary reproduction[J]. Electronic Intellectual Property, 2013, (Z1): 110.

[9] 商建刚.数据训练的著作权法分析[J].法学论坛,2025,40(02):77.
Shang J G. Analysis of data training under copyright law[J]. Legal forum, 2025, 40(02): 77.

[10] 张今.版权法中私人复制问题研究——从印刷机到互联网[M].北京:中国政法大学出版社，2009:52.
Zhang J. Research on the issue of private copy in copyright law —— from the printing press to the internet[M]. Beijing: China University of Political Sci-

ence and Law Press, 2009: 52.

[11] 单莹,奈一雄.网络环境下临时复制的著作权法律规制[J].黑龙江省政法管理干部学院学报,2015,(06):57.
Shan Y, Nai Y X. Legal regulation of temporary reproduction in the network environment from the perspective of copyright law[J]. Journal of Heilongjiang Administrative Cadre College of Politics and Law, 2015, (06): 57.

[12] 罗胜华.网络临时复制问题法律研究[J].知识产权,2004,(04):20.
Luo S H. Legal research on the issue of temporary reproduction in the network environment[J]. Intellectual property, 2004, (04): 20.

[13] Lemley M A, Casey B. Fair learning[J]. Texas law review, 2021, 99: 743-760.

[14] 邱紫雁.人工智能时代机器学习版权合理使用制度的弹性分治设计——基于日本《著作权法》柔性合理使用条款的考察[J].中国出版,2025,(09):52.
Qiu Z Y. A flexible, multi-tiered design for the fair use doctrine of machine learning in the age of artificial intelligence: An examination based on the flexible fair use provision of Japan's Copyright Act [J]. China Publishing Journal, 2025, (09): 52.

[15] 国家数据局.政策解读:自动化程序收集公开数据的合法边界[Z].2026.
National Data Administration. Policy Interpretation: Legal boundaries for the collection of public data by automated programs[Z].2026.

[16] 李铭轩.论大模型训练数据的合理使用[J].法学家, 2025, (05): 32.
Li M X. On the fair use of training data for large models[J]. The jurist, 2025, (05): 32.

[17] Sag M. The new legal landscape for text data mining and machine learning[J]. Journal of the Copyright Society of the USA, 2019, 66(3): 291-328.

[18] 王迁.人工智能生成的内容是作品吗?——以学术规范和著作权法的关系为视角[J].中国法律评论,2025,(05):42-44.
Wang Q. Are AI-Generated contents works? ——From the perspective of the relationship between academic norms and copyright law[J]. China law review, 2025, (05): 42-44.

[19] 李顺德.馆藏文献数字化与复制权保护问题[J].国家图书馆学刊,2004,(04):54.
Li S D. Digitization of collection documents and protection of reproduction right[J]. Journal of the National Library of China, 2004, (04): 54.

[20] 冯晓青, 付继存.著作权法中的复制权研究[J].法学家, 2011, (03): 100.
Feng X Q, Fu J C. Studies on the right of copy in copyright law[J]. The jurist, 2011, (03): 100.

[21] Samuelson P. Justifications for copyright limitations & exceptions[M]// OKEDIJI R L. Copyright law in an age of limitations and exceptions. New York: Cambridge University Press, 2017:12-53.

[22] 万勇.人工智能时代著作权法合理使用制度的困境与出路[J].社会科学辑刊,2021,(05):95.
Wan Y. Dilemmas and solutions of the fair use system in copyright law in the age of artificial intelligence[J]. Social science journal, 2021, (05): 95.

[23] 施小雪.重塑复制权:生成式人工智能数据训练的合法化路径[J].东方法学, 2024, (06): 75.
Shi X X. Reshaping the right of reproduction: the legalization path of data training for generative artificial intelligence[J]. Oriental law, 2024, (06): 75.

[24] Samuelson P. Fair use defenses in disruptive technology cases[J]. UCLA law review, 2024, 71: 1567-1568.

[25] Leval P N. Toward a fair use standard [J]. Harvard law review, 1990, 103(5):

1111.

[26] 熊琦.著作权转换性使用的本土法释义[J]. 法学家, 2019, (02): 124.
Xiong Q. Transformative use interpretation in china copyright law[J]. The jurist, 2019, (02): 124.

[27] 张伟君.论大模型训练中使用数据的著作权规制路径[J]. 东方法学, 2025, (02): 88.
Zhang W J. On the copyright regulation path of data used in large model training [J]. Oriental law, 2025, (02): 88.

[28] 杨利华,王诗童.人工智能生成内容的著作权规制研究[J]. 法治研究, 2025, (03): 58.
Yang L H, Wang S T. Research on the copyright regulation of AI-Generated content[J]. Research on rule of law, 2025, (03): 58.

[29] 张新宝,卞龙.人工智能生成内容的著作权保护研究[J].比较法研究,2024,(02):86.
Zhang X B, Bian L. Research on the copyright of AI-Generated content[J]. Journal of comparative law, 2024, (02): 86.

[30] 朱阁."AI文生图"的法律属性与权利归属研究[J].知识产权,2024,(01):31.
Zhu G. Research on the legal attributes and rights attribution of AI-Generated images [J]. Intellectual property, 2024, (01): 31.

[31] 徐小奔.论算法创作物的可版权性与著作权归属[J]. 东方法学, 2021, (03): 51.
Xu X B. On the copyrightability and copyright ownership of algorithmic creations[J]. Oriental law, 2021, (03): 51.

[32] 最高人民法院新闻局.2026年全国两会《最高人民法院工作报告》解读系列全媒体直播访谈第二场[Z]. 2026.
Information Office of the Supreme People's Court of PRC. Second Session of the 2026 NPC and CPPCC Annual Sessions: Series of omnichannel live interviews on the Work Report of the Supreme People's Court of PRC [Z]. 2026.

[33] 郑宇.公共数据的产权运行机制与技术方案[J].大数据,2024,10(05):139.
Zheng Y. Mechanisms and techniques for operating public data right[J]. Big Data Research, 2024,10(05):139.

作者简介

赵璇（2006– ），女，北京工业大学经济与管理学院研究助理，主要研究方向为数据法、知识产权法等。

刘自钦（1989– ），男，博士，北京工业大学经济与管理学院副教授，主要研究方向为数据法、知识产权法等。

收稿日期: XXXX-XX-XX

通信作者:

基金项目: 国家社会科学基金重大项目(No.21&ZD164);北京市法学会市级法学研究课题(No.BLS(2025)B013-1)

Foundation Items: The Major Program of National Social Science Fund of China (No.21&ZD164); Municipal-Level Legal Research Project of Beijing Law Society (No.BLS(2025)B013-1)