

# 生成式人工智能素材利用之侵权风险与应对

孙菁遥

澳门科技大学法学院, 澳门

收稿日期: 2025年10月11日; 录用日期: 2025年11月3日; 发布日期: 2025年11月13日

## 摘要

生成式人工智能(Generative Artificial Intelligence, 以下简称GenAI)技术的广泛滥用, 司法实践中所涉嫌的案例越来越多, 鼓励技术创新与保护权利人利益之间产生冲突。GenAI基于对海量作品的学习和模仿, 生成用户指令内容。然而, 这一智能过程深陷著作权侵权风险与规制困境。首先, 分析在素材获取阶段和算法训练阶段, 可能涉及复制权、汇编权、翻译权等著作权权利的侵犯, 详细阐述GenAI素材训练的关键环节, 如数据抓取、数据清洗与标注等, 将其逐一与复制权、汇编权、改编权进行对应分析; 其次, 分析现行著作权法在应对此类问题时的困境, 即我国“合理使用”以及“法定许可”制度难以适配回应GenAI素材利用之侵权风险; 最后, 提出确立“GenAI素材训练合理使用适用之例外”, 同时提出以“技术赋能”建立数据披露机制。本文旨在分析GenAI在素材获取和算法训练阶段中所面临的著作权侵权风险, 审视现行著作权法体系规制框架的局限性, 探寻在人工智能时代背景下, 兼顾技术创新与权利保护的优化路径。

## 关键词

生成式人工智能, 著作权侵权, 合理使用, 法定许可

# Infringement Risks and Responses in the Utilization of Generative AI Materials

Jingyao Sun

Faculty of Law, Macao University of Science and Technology, Macao

Received: October 11, 2025; accepted: November 3, 2025; published: November 13, 2025

## Abstract

The widespread abuse of Generative Artificial Intelligence (GenAI) technology has led to an increasing number of suspected cases in judicial practice, creating a conflict between encouraging technological

文章引用: 孙菁遥. 生成式人工智能素材利用之侵权风险与应对[J]. 争议解决, 2025, 11(11): 130-135.

DOI: 10.12677/ds.2025.1111353

innovation and protecting the interests of rights holders. GenAI generates user-instructed content based on learning and imitating a vast amount of works. However, this intelligent process is deeply entangled in the risk of copyright infringement and regulatory dilemmas. Firstly, this paper analyzes the potential infringement of copyright rights such as the right of reproduction, the right of compilation, and the right of translation during the stages of material acquisition and algorithm training. Secondly, it examines the predicament of the current copyright law in addressing such issues, namely, the difficulty of the “fair use” system in China to adapt to and respond to the infringement risks of GenAI’s material utilization. Finally, it proposes the establishment of an “exception for the application of fair use in GenAI material training” and suggests the establishment of a data disclosure mechanism through “technology empowerment”. This paper aims to analyze the copyright infringement risks faced by GenAI during the stages of material acquisition and algorithm training, review the limitations of the current copyright law regulatory framework, and explore an optimized path that balances technological innovation and rights protection in the context of the AI era.

## Keywords

GenAI, Copyright Infringement, Fair Use, Statutory License

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

生成式人工智能技术爆发式发展，应用领域越来越广泛。其通过学习或模仿海量数据以生成新的内容，对包括受著作权保护的作品在内的海量素材进行获取、复制与算法解析，这一过程存在潜在的著作权侵权风险，引发技术迭代与法律滞后之间的矛盾。当前，司法实践中已然出现多起针对 GenAI 素材训练行为之诉讼案例，凸显了问题的紧迫性，如何平衡好鼓励技术创新与保护权利人合法权益之间关系，是立法与司法都要面对严峻挑战。若将素材训练行为简单认定为侵权，无疑是对技术创新的扼杀；若对其完全放任，则可能会破坏创作人的积极性，破坏市场秩序。对此，厘清 GenAI 对素材利用环节的侵权风险，并有针对性提出建设性意见，显得尤为重要。

## 2. GenAI 素材利用之侵权风险分析

生成式人工智能(Generative Artificial Intelligence，以下简称 GenAI)是指利用机器学习算法、通过对海量素材训练模型，能够生成新文本、图像、音频等内容的人工智能技术[1]。生成式人工智能对素材之利用，本文着重对素材获取、算法训练阶段的著作权侵权风险进行分析。素材获取与算法训练是 GenAI 输出内容的基础和前置程序。

### (一) 素材获取阶段之侵权风险分析

素材的获取是整个过程的开端。获取阶段包括对数据的抓取、数据清洗与标注等环节，这每一环节均可能涉及对权利人复制权、改编权等著作权权项的侵犯，以下对此逐一分析。

#### (1) 数据抓取之复制权侵权风险

数据的抓取是整个阶段的首要环节。数据的来源非常广泛，包括公共领域的的数据，也包括受著作权保护的作品。通过网络爬虫的方式，即存在非法获取手段或未经授权，对包括文字、图像、音频视频在内的作品进行大规模采集，可能属于对材料的非法使用，即将作品从原始载体复制到智能数据库，该行

为构成《著作权法》第十条第一款第五项<sup>1</sup>所规定的“复制权”所控制的行为方式，即非法爬取网页受著作权法保护的作品，且未获得授权，不属于侵权豁免事由，构成对复制权的直接侵犯。

部分学者认为，人工智能为了便于“机器阅读”，将已有作品数字化并转换数据格式，这个过程出现的复制行为属于临时复制，不属于著作权法意义上的复制行为，不构成对复制权的侵犯[2]。但是，在GenAI 素材训练过程中，数据的抓取并非仅为“临时”使用，而是为了服务于后续人工智能的学习，模型构建，内容生成，并且后续过程的进行完全建立在已有作品，不同于人类学习后展现个性的过程，难以排除在复制权规制之外。

## (2) 数据清洗与标注之汇编权、改编权侵权风险

数据清洗是指对抓取的素材进行格式统一等处理；数据标注是为了后续的数据训练，添加分类标签等处理。数据的清洗与标注往往具有连贯性，便于人工智能对模型的识别与学习。在数据的清洗与标注过程中，若对作品内容进行选择或重新组合，会构成《著作权法》第十条第一款第十六项<sup>2</sup>所规定的“汇编权”控制的行为。

在数据的清洗与标注过程中，涉及对原作品进行实质性修改或重新表达，若生成具有独创性的新作品，则会构成对《著作权法》第十条第一款第十四项所规定的“改编权”的侵犯。对此，判断是否构成对改编权的侵犯，就需要具体情况具体分析，即以是否具有独创性反推行为性质。思想属于主观范畴，必须以一定形式表达出来，在思想向表达转化过程中，对情节的取舍、安排、设计等创作系独立完成，包含一定的智力创作，就具有独创性。对此，只有在作品的创作过程中存在自然人的干预贡献，人工智能生成的内容才有资格构成作品，如果没有自然人的干预，人工智能生成的内容不应受到版权的保护<sup>3</sup>。对原作品表达形式改变若具备一定的独创性，构成对改编权的侵犯。

## (二) 算法训练阶段之侵权风险分析

人工智能在进行素材学习时，会对大量原始文本进行分析，转换为机器可读取且可理解的机器语言。在此机器学习过程中涉及对训练素材的翻译、改编等行为，这些行为是否侵犯著作权需要进一步分析。首先需要明确的是，机器学习阶段的各项行为都是在获得训练素材的基础上进行的，若以非法手段或未经授权获得素材，那么机器学习过程中的翻译、改编等行为当然构成侵权。而重点在于合法方式获取数据后，对机器学习阶段各项行为的法律属性进行分析。

分析机器学习过程中的翻译、改编等行为是否构成对著作权的侵权，需要以结果反推行为性质。《著作权法》意义上的作品是以智力表达为构成要素，同时也应描述为作品的可版权性，即实质上的独创性和形式上的可再现性[3]。翻译权<sup>4</sup>，即将作品从一种语言文字转换成另一种语言文字的权利。人工智能学习时，需要将原始文本转换为机器可读取的形式并建立训练数据库，但人工智能对作品的使用并非翻译[4]。

改编权，即改变作品，创作出具有独创性的新作品的权利。获取数据后进行训练阶段，AI 会对作品中的所有元素进行分析，包括行为逻辑、内在的思想情感、作者的表达风格等。将原作品中的用词、排版等进行拆解，重新组合，如果此过程未经著作权人授权许可，存在侵犯改编权的法律风险。通说观点认为，改编权的构成要件有以下两点：改变原作品的表现形式或用途、创作出具有独创性的新作品[3]。可以看出，改编权要求产生一个具有独创性的新作品。人工智能在获取作品后，对其中的用词等转换为机器可以理解的“机器语言”，这个过程改变了原作品的表现形式，但是，此过程输出的产物不具备新

<sup>1</sup> 《中华人民共和国著作权法》第十条第一款第五项：复制权，即以印刷、复印、拓印、录音、录像、翻录、翻拍、数字化等方式将作品制作一份或者多份的权利。

<sup>2</sup> 《中华人民共和国著作权法》第十条第一款第十六项：汇编权，即将作品或者作品的片段通过选择或者编排，汇集成新作品的权利。

<sup>3</sup> 国际保护知识产权协会 2019 年 9 月 18 日发布的《关于人工智能生成作品版权问题的决议》规定。

<sup>4</sup> 《中华人民共和国著作权法》第十条第一款第十五项：翻译权，即将作品从一种语言文字转换成另一种语言文字的权利。

作品的特点。在数据训练阶段作品形式的改变，只是形成数据模型的必要步骤，属于技术手段的应用，不涉及独创性的表达，并没有形成固定完成的作品，即数据模型缺乏再现性，所以，训练阶段的数据应用不构成对改编权的侵犯。

### 3. GenAI 素材训练法律规制现状

#### (一) 合理使用适用局限

立法层面，目前我国《著作权法》对 GenAI 素材训练行为的合理使用缺乏准确适用的制度设计，对 GenAI 素材训练的规制仍框于传统著作权制度体系。我国对合理使用制度采取“列举式”立法模式，即现行《著作权法》第二十四条列举适用合理使用制度的十二种情形。其中，“个人使用”、“科学研究”具有较高适用可能性，学界讨论最多，下面将对其详细分析。另外，若将 GenAI 对海量素材之利用行为泛泛纳入“合理使用”制度，将对原创内容市场造成难以估量的市场替代效应和经济损害。

##### (1) 无法适用“个人使用”情形

我国现行《著作权法》第 24 条第 1 款第 1 项规定“个人使用”情形<sup>5</sup>。GenAI 素材训练为构建大数据模型涉及对海量作品的学习模仿，似乎符合“个人使用”情形的规定范畴，但该条无法对素材训练提供合理的抗辩理由。原因在于该条限定的主体为“个人”，通常仅限于“自然人”，但素材训练的主体条件不符合。首先，目前人工智能仍具有工具属性，无法独立作为法律关系的主体。其次，人工智能开发者或研究者也不满足主体条件，人工智能开发者通常是企业，即法人，与自然人是并列的法律关系主体类型；主导人工智能“机器学习”过程的通常是科研人员，但这些科研人员的研究是基于企业安排的工作任务，属于职务行为，并非是为了个人学习或研究而使用作品。由此，合理使用制度下的“个人使用”无法成为 GenAI 素材训练合法化的抗辩理由。

##### (2) 无法适用“科学研究”情形

我国现行《著作权法》第 24 条第 1 款第 6 项规定“科学研究”情形<sup>6</sup>。根据法律规定可以分析得知，“科学研究”与“课堂教学”并列规定，其使用目的也应当局限在“学术性”、“非商业性”。首先，GenAI 素材训练的使用目的不符合该条规定，人工智能开发者大多是互联网企业，具有明确的营利目的；其次，该条规定对作品的使用数量限制在“少量”，而 GenAI 为保证输出内容的准确性以及场景的贴合度，防止生成的内容出现观点偏向等问题，对作品的使用必然是海量的，并且通常是全文复制作品。可见，GenAI 对素材的使用不符合该条的规定。

#### (二) 法定许可适用局限

法定许可制度与合理使用制度同为著作权限制制度。法定许可制度允许使用者在无著作权人授权的情况下使用其作品。以法定许可制度适用 GenAI 对素材的使用行为，法定许可使用制度相关法律规范的构建、技术支撑以及实施保证路径难度很大[5]。

首先，我国现行《著作权法》中法定许可制度的适用条件较为严格，生成式人工智能的创作过程和使用场景难以满足<sup>7</sup>。我国《著作权法》第 25 条规定“法定许可”制度<sup>8</sup>。根据法律规定可知，法定许可制度适用的场景限于编写出版教科书，但生成式人工智能对素材的创作场景多元，使用范围过于宽泛，

<sup>5</sup> 《中华人民共和国著作权法》第二十四条第一款第一项：为个人学习、研究或欣赏，使用他人已经发表的作品，可以不经著作权人许可，不向其支付报酬。

<sup>6</sup> 《中华人民共和国著作权法》第二十四条第一款第六项：为学校课堂教学或者科学研究，翻译、改变、汇编、播放或者少量复制已经发表的作品，供教学或者科研人员使用，但不得出版发行，可以不经著作权人许可，不向其支付报酬。

<sup>7</sup> 王胜捷. 生成式人工智能文本与数据挖掘的合理边界与侵权规制[C]//上海市法学会.《智慧法治》集刊 2024 年第 1 卷——2024 年世界人工智能大会法治论坛文集. 上海市法学会: 上海市法学会, 2024: 13.

<sup>8</sup> 《中华人民共和国著作权法》第二十五条：为实施义务教育和国家教育规划而编写出版教科书，可以不经著作权人许可，在教科书中汇编已经发表的作品片段或者短小的文字作品、音乐作品或者单幅的美术作品、摄影作品、图形作品，但应当按照规定向著作权人支付报酬，指明作者姓名或者名称、作品名称，并且不得侵犯著作权人依照本法享有的其他权利。

无法简单归入特定场景。

其次，法定许可的司法适用层面也面临极大困难。对报酬的确定标准面临困难。报酬标准需要法律预先规定，但生成式人工智能技术更新迭代迅速，报酬标准的确定基础随之不断更新，法律的稳定性无法适应著作权经济价值动态性。同时，目前我国生成式人工智能产业处于发展初期，立法机关和司法机关缺乏有效的数据参考，著作权主管部门甚至需要根据相关企业的成本利润等数据，建立经济数据模型进行复杂计算，才能确定报酬支付报酬<sup>[6]</sup>，可见，更加难以适应人工智能技术的迭代。

然后，法定许可适用面临的最大挑战之一是交易成本极高，费用支付操作难度大，巨额的交易成本可能远超过社会收益。GenAI 素材训练涉及海量作品，如果将其适用法定许可，开发者需要向著作权人支付报酬，海量作品的费用支付往往需要著作权集体管理组织配合，集体管理组织需要定位著作权人，但这海量作品中必然存在“孤儿作品”，即难以确定真实的著作权人，那么报酬支付对象的确定又是一高成本需要解决的问题；再者，集体管理组织代为诉讼维权和未分配费用的披露监管，都需要收取法定许可费弥补运行成本；可见，执行法定许可的成本，甚至可能超出合理使用判定后著作权人的损失<sup>[7]</sup>。

最后，作品追踪识别难度大。GenAI 训练过程对作品的使用是隐蔽性的，没有直接产生受外界接触的内容，这就需要相应技术手段进行识别与追踪，这对权利人来说是高成本且高难度的技术要求，对此，有学者提出，适用区块链技术确保及时获取训练中的作品使用记录并且降低被篡改或删除的可能<sup>[2]</sup>。但是，区块链技术目前发展尚不成熟，难以在短时间内普及使用，并且存在数据泄露的商业风险，增加营利者对数据维护的成本。

#### 4. GenAI 素材利用之规制路径优化

##### （一）确立相应合理使用适用例外

为有效应对生成式人工智能素材利用之侵权风险，我国应考虑在《著作权法》中增设人工智能训练作品合理使用规则，并且该规则的增设需要综合考量技术创新需求与权利人权利的保护<sup>[4]</sup>。本文认为对 GenAI 素材利用行为的规制可以通过合理使用制度的“兜底条款”进行适用，确立合理使用制度的例外。

首先，对 GenAI 素材训练合理使用例外之主体而言，不应局限在个人或科学研究机构。随着人工智能技术的广泛应用，商业机构才是生成式人工智能训练的关键主体，越来越多的科技企业成为推动人工智能技术发展的核心力量，若现行《著作权法》体系仍然回避商业机构是 GenAI 素材训练主体的法律问题，立法滞后性和司法紧迫性将会更加严重。

与之相对应，GenAI 素材训练合理使用例外之目的界定，就不应仅局限在公益性用途，对商业目的下的企业研发活动也应涵盖在内。从促进文化和科技创新的角度来讲，对 GenAI 素材训练合理使用例外应当扩大解释，将其界定在公益性研究与商业性研发层面。欧洲在规制计算机相关著作权例外规定时对公益性目的的过分关注，导致其在全球互联网产业竞争中处于劣势，可以为我国立法提供借鉴经验。

##### （二）区块链技术赋能训练素材披露机制

建立 GenAI 训练素材披露机制有助于提升技术透明度，形成公众监督，在一定程度上能确保在素材训练过程中所使用的数据来源合法。另一方面，对训练使用的素材进行披露符合当前国际立法趋势。比如，欧盟《人工智能法案》将披露训练素材来源和版权信息列为提高人工智能技术透明度的重要措施<sup>9</sup>。可见，建立生成式人工智能素材训练披露机制有助于保障原权利人的合法权益，还会显著提升我国企业的国际竞争力。

训练素材披露机制的首要环节是对版权信息的获取，对此，学界有观点认为，生成式人工智能素材训练对作品的使用是海量性的，加之海量作品中亦存在“孤儿作品”等，难以对真实的版权信息进行精确定

<sup>9</sup>See EU Artificial Intelligence Act, Recital 107.

位, 对其进行数据溯源和公开披露不切实际。但是这一观点忽视了数字时代人工智能技术的发展迭代, 生成式人工智能技术自身就可以通过算法自动定位到相关版权信息, 使披露机制的构建成为可行。素材披露机制的披露实施还需要借助先进的数字技术帮助实现公开, 其中区块链技术可以提供解决方案[8]。

区块链采取分布式存储结构, 对训练素材的来源、使用时间、授权主体等关键且隐蔽的信息做到精准记录, 为构建透明、可追溯的素材披露机制提供了技术支撑, 为监管机构等相关方审查数据来源的合法性提供可行方法, 也有助于防范数据泄露风险, 是对人工智能时代素材利用之应对的最好技术回应。

对任何合法的文本与数据挖掘行为, 包括以商业性使用为目的, 若权利人未以适当方式明确保留其权利, 则推定其允许该使用; 但若权利人已通过技术措施或声明方式明确表示反对, 则使用者需获得授权。既保障了非商业研究的自由空间, 也为商业应用提供了法律确定性, 同时尊重了权利人的自主选择权。

## 5. 结论

生成式人工智能素材利用之相关问题, 深刻揭示了新兴技术与传统规则之间的矛盾。本文分析 GenAI 对素材的利用, 因其黑箱操作、临时复制等特性, 使其在法律定性上不能简单等同于传统著作权侵权行为。对此, 我国现行《著作权法》中合理使用制度因主体、目的和作品使用数量等的限制而难以适用; 法定许可制度也因场景不符, 司法适用交易成本过高而缺乏可行性。对此, 本文提出确立“GenAI 素材训练之合理使用例外”、以及技术赋能建立素材数据披露机制, 有效提升训练过程的透明度和可追溯性。

## 参考文献

- [1] 徐恩平. 生成式人工智能司法应用的风险与出路[J]. 中阿科技论坛(中英文), 2024(1): 149-155.
- [2] 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, 22(2): 68-79.
- [3] 吴汉东. 知识产权总论[M]. 北京: 中国人民大学出版社, 2020.
- [4] Li, Z. (2025) Copyright Infringement Risks and Responses of Generative Artificial Intelligence. *Science of Law Journal*, 4, 23-32.
- [5] 吴汉东. 人工智能生成作品的著作权法之问[J]. 中外法学, 2020, 32(3): 653-673.
- [6] 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.
- [7] 赵旭. 生成式人工智能在机器学习中的合理使用问题[J]. 暨南学报(哲学社会科学版), 2024, 46(3): 79-95.
- [8] 黄保勇, 施一正. 区块链技术在版权登记中的创新应用[J]. 重庆大学学报(社会科学版), 2020, 26(6): 117-126.