C

SCCR/49/6
原文：英文
日期：2026 年 10 月 23 日

版权及相关权常设委员会

第四十九届会议
2026 年 11 月 30 日至 12 月 4 日，日内瓦

人工智能训练与版权问题范围界定研究

丹尼尔·J·热尔韦、崔成宰（音译）、梅丽莎·奥米诺、罗德里戈·科普克·萨利纳斯和王迁
编拟

人工智能训练与版权问题
范围界定研究

免责声明：
本研究由世界知识产权组织（产权组织）秘书处委托完成。本文件中表达的观点为作者的观点，并不反映产权组织成员国或秘书处的意见。

致　谢

丹尼尔·J·热尔韦（DANIEL J. GERVAIS）是本研究报告的牵头合著者。他是范德比尔特大学法学院米尔顿·R·安德伍德法学教授（名誉），阿姆斯特丹自由大学高级副教授，也是美国法学会当选会员。可通过以下网址查阅他的研究成果：www.collectivemanagement.net、www.tripsagreement.com和www.thehumancause,.org。

崔成宰（SUNG JAI CHOI）是世宗大学法学教授。他拥有哥伦比亚大学法学院的法学硕士学位。他的研究聚焦知识产权法和竞争法。他是《版权法》（2024 年）的合著者，并作为“新一代领军人物”荣获 2021 年韩国仲裁大奖。

梅丽莎·奥米诺（MELISSA OMINO）是内罗毕斯特拉思莫尔大学知识产权与信息技术法中心（CIPIT）主任。她拥有知识产权与贸易法博士学位，专攻全球南方的知识产权、人工智能、数据保护及数字权利。

罗德里戈·科普克·萨利纳斯（RODRIGO KOPKE SALINAS）是巴西热图利奥·瓦加斯基金会（FGV）圣保罗法学院（GVlaw）教授，拥有哥伦比亚大学法学硕士学位。他是巴西律师协会版权特别委员会成员，专攻媒体法与娱乐法。

王迁现任华东政法大学法律学院教授、学术委员会副主任委员，兼任中国版权协会副理事长，曾担任中国《著作权法》第三次修改专家委员会成员。

注：本研究报告使用语言模型对句法、语法和内容进行了检查并作了编辑建议，以使这篇合著文章的表述协调一致。

目 录

导 言 5
第一部分 受保护内容在生成式人工智能模型开发中的使用 6
A. 获取数据 6
1. 按系统类型划分的数据取用模式 6
2. 网页抓取 6
3. 检索增强生成 7
B. 在训练中使用数据 8
1. 训练模型 8
2. 优化与强化学习 10
C. 训练完成后 10
1. 训练语料库的其他用途 10
2. 复制件在模型中的留存 10
3. 输入-输出关系 13
第二部分 授权、许可和补偿 14
A. 开发者获取可受版权保护客体用于训练的现有方案 14
1. 五种获取途径 14
2. 合作安排 16
B. 主要政策要素 16
1. 受生成式人工智能模型训练影响的权利 16
(a) 复制权 16
(b) 改编权和演绎作品 17
(c) 向公众传播和提供 17
(d) 精神权利 17
(e) 权利管理信息（RMI）和反规避 18
2. 自愿（集体或个人）许可 18
(a) 人工智能使用的个人和集体许可 18
(b) 主要的集体许可模式（自愿许可、延伸性许可和强制许可） 22
3. 原始语言数据的社区许可与管理许可 25
4. 法定文本与数据挖掘例外与一般性条款 26
5. 报酬制度 29
(a) 报酬制度的类型 29
(b) 报酬方式和依据 30
第三部分 权利执行 31
A. 现行判例法的启示 31

1.“模型即复制件”问题 31
2.文本与数据挖掘例外 33
3.对域外训练的管辖权 34
4.替代问题 36
B.针对人工智能开发者执行权利 39
1.权利执行格局的结构性特征 39
2.当事方之间的不对等 39
C.执行权利面临的挑战 39
1.透明度 40
2.数据攻击、过滤及选择退出标准激增 41
3.判例法概况 43
结　论 44

导　言

本研究报告是应版权及相关权常设委员会（SCCR）的请求编写的。在第四十七届会议（2025年12月1日至5日，日内瓦）上，常设委员会请秘书处就人工智能训练与版权之间的关系及相关规则的政策或监管方法，以及授权、权利执行和使用补偿方面的适用做法，拟定一份研究报告。该请求记录于该届会议的《主席总结》（2025年12月5日文件SCCR/47/SUMMARY[1]，第21段）。

本研究报告本质上属于描述性研究和比较研究，审视了多个司法管辖区采用的方法及正在形成的做法，并探讨了已提出的各种方案，但并未向成员国推荐任何具体做法，也未就成员国尚未解决的争议问题表明立场。

本研究报告关注生成式人工智能模型的训练问题，即利用内容来构建和改进此类模型，以及当相关内容受版权或相关权保护时，此类使用所引发的授权、权利执行和补偿问题。人工智能与版权之间更广泛的交集远不止于训练。本研究报告特别探讨了已部署模型产出的法律地位，既涉及任何特定产出是否侵犯了该模型训练所基于的作品，也涉及产出本身是否以及在何种条件下可受版权保护。与产出相关的问题引发了其他一些问题，而常设委员会并未要求本研究报告对此进行探讨。仅在需要全面阐述训练问题或提供必要背景信息时才会提及这些问题。

本研究报告分为三个部分。第一部分从技术和描述层面阐述了在生成式人工智能模型开发过程中使用受保护内容的操作，并指出了这些操作可能涉及各国版权立法和国际版权条约所承认的专有权的具体情形。第二部分探讨了所涉及权利的范围和程度，以及行使这些权利的机制：自愿许可（无论是个人许可还是集体许可）；以及适用的例外和限制。第三部分涉及权利执行问题，并探讨了从多个司法管辖区开始形成的判例法体系中可能汲取的经验教训。该研究报告还涉及训练的跨境性质及相关的权利执行问题，包括透明度以及在适用情况下表达“选择退出”的标准。本研究报告中援引的司法管辖区仅供说明之用，并非详尽无遗。其中包括巴西、加拿大、中国、欧洲联盟及其众多成员国、肯尼亚、大韩民国、联合王国和美利坚合众国。

[1] 见 https://www.wipo.int/meetings/zh/doc_details.jsp?doc_id=653196

第一部分　受保护内容在生成式人工智能模型开发中的使用

生成式人工智能模型是分阶段逐步开发的，通常在每个阶段都会使用第三方内容。部分内容可能受版权及相关权保护；部分内容则可能属于公有领域，或是因为保护期限已届满，或是因为从一开始就未受保护。无论这类内容的法律地位如何，对其所采用的技术操作通常都是相同的。包括爬取、复制、分词和统计分析在内的相同处理流程（在下文阐释），可以同样应用于近期小说和公有领域的八世纪文本。本部分对这些操作进行了说明，并从描述性角度指出了这些操作可能涉及各国版权立法和国际版权及相关权条约所承认的权利的具体情形。

下文将依次介绍训练的三个阶段。A 节讨论数据采集，即人工智能模型开发者如何构建语料库，该语料库通常很大部分包含受保护的客体，用于训练模型。B 节介绍这些数据在训练中的实际应用，包括主要训练运行以及随后进行的各种微调。C 节讨论训练完成后的情况，包括对训练语料库的处理，以及训练后模型本身是否包含其训练所基于的作品[2]的复制件这一问题。

A.获取数据

大语言模型的开发始于语料库的构建，语料库通常规模非常庞大，且来源多样。开发者通过若干方式获取数据。某些数据可能通过许可获取，在这种情况下，内容可连同其使用许可一并提供。在这种情况下，开发者可以购买由第三方汇编的精选数据集的访问权限，或者与出版商、新闻机构或集体管理组织（CMO）协商签订直接许可协议。如果未就受保护作品的使用授予许可，则其使用可能基于以下假设：该客体适用相关专有权的例外或限制。就数量及其引发的版权问题而言，一条主要训练途径是从开放网络（互联网）自动检索资料。训练也可能使用合成数据，即由现有人工智能模型生成的数据，而非由人类创作者收集的数据，尤其是在开发的后期阶段。

1.按系统类型划分的数据取用模式

系统对数据的需求通常不仅限于最初的训练运行：数据还用于重新训练和微调，用于对维护中的搜索库进行的文本和数据分析，用于检索增强生成（RAG）[3]，以及用于生成预览和摘要片段。这些用途在所需材料量、单份作品的权重，以及来源可追溯的难易程度上各不相同。系统的类型往往会影响取用模式：媒体生成系统和基于大语言模型的聊天机器人所用数据主要来自训练阶段的摄取；企业基于语言（亦称基础）模型构建的可定制系统会加入特定领域的语料库，并且往往会加入在使用时查询的搜索库；而对话式搜索引擎则需要在生成内容时访问在线资源。一个系统在这一范围中的位置，会影响其数据使用所涉及的专有权组合，以及任何许可的相应结构。这一问题在第二部分 B.2 节中详述。

2.网页抓取

网页抓取是基础模型开发者构建其训练语料库时常用的方法。不过，这一术语并未得到普遍采用。有些资料区分了“网络爬取”（程序顺着页面上的链接逐页在网站间自动移动）与“网络抓取”（从已检索到的页面中选择性地提取确定的内容）。实际上，这两项活动可能是结合在一起进行的：开发者部署一个爬虫，由其在网络上顺着链接爬行，并在每个页面上识别并提取需要保留的文本、图像或其他内容。输出结果通常是该内容的副本，一般以结构化的方式存储在开发者自己的基础设施上。

---

[2] 为简化行文，“作品”一词应酌情理解为包括相关权的客体。
[3] 检索增强生成（通常简称 RAG）将在本部分第 3 节中进一步讨论。

有四个特点与本讨论相关。首先，该技术涉及制作复制件。复制件可能会保存在服务器或缓存中，也可能在进一步转换之前在内存中处理。这些操作可能涉及《伯尔尼公约》第九条第 1 款、《世界知识产权组织版权条约》（WCT）第 1 条第（4）款以及各国立法相应条款中规定的复制权。关于 WCT 第 1 条第（4）款的议定声明确认，以数字形式存储受保护的作品属于复制，且复制权及其例外完全适用于数字环境。[4]某些复制件的临时性，或者附带复制和暂时复制例外等的适用，是否会改变分析，这个问题将在本研究报告的后文中予以探讨。《世界知识产权组织表演和录音制品条约》（WPPT）第 7 条和《视听表演北京条约》第 7 条均规定了针对“以任何方式或形式”直接和间接地进行复制的权利。

其次，检索可能是非选择性的。爬虫并不总能区分属于公有领域的内容、在开放许可下提供的内容、在没有任何许可声明的情况下提供的内容，以及受到保护且需遵守服务条款或通过技术保护措施才能访问的内容。例如，有些（但并非所有）爬虫会遵守 robots.txt 指令。[5]此类区分有时是在事后通过语料库过滤或去重完成的，但初始复制行为并非总是选择性的。

第三，权利人对该活动的知情程度有限。权利人可能并不知道自己的作品已被抓取，也不知道是哪款爬虫何时进行的抓取，以及后续被用于何种用途。一些国家法律中开始出现的披露义务——特别是在欧洲联盟《2024 年人工智能法案》中——要求通用人工智能模型的提供商公布用于训练的“足够详细的”内容摘要。[6]现在评估这些义务的实际效果还为时过早。如何将披露与尊重开发者的合法商业利益相结合，这一问题本身正是当前讨论的主题。[7]

第四，为训练目的制作和使用受保护作品的复制件，可能涉及删除或更改权利管理信息，以及规避权利人根据 WCT、WPPT 和《北京条约》行使权利时所采用的有效技术措施。

3.检索增强生成

另一种独立但相关的技术是检索增强生成（RAG）。[8]它是人工智能系统回答查询的问题时采用的一种方法，即首先查阅一个单独预先汇编好的文档集合，而不是仅依赖其在训练过程中吸收的内容。一个有用的对比是：一个人凭记忆回答问题，与同一个人停下来查阅桌旁参考书架上的资料，这两者之间的区别。当收到查询指令时，系统搜索该集合，检索与问题最相关的段落，将其与查询内容并列显示，并利用这些段落来构建答案。该集合通过制作并存储文档副本来预先构建，且在生成答案的瞬间，相关段落往往会被再次复制到系统中，这些要点对于下文所述的版权分析至关重要。

---

[4] 《世界知识产权组织版权条约》（1996 年 12 月 20 日）关于第 1 条第（4）款的议定声明：“《伯尔尼公约》第 9 条所规定的复制权及其所允许的例外，完全适用于数字环境，尤其是以数字形式使用作品的情况。不言而喻，在电子媒体中以数字形式存储受保护的作品，构成《伯尔尼公约》第 9 条意义下的复制。”

[5] 《机器人排除协议》（Robots Exclusion Protocol）作为互联网工程任务组 RFC 9309（2022）标准，是一项约定：网站运营商在网站根目录下放置一个纯文本文件（robots.txt），以表明哪些自动化爬虫可以访问网站的哪些部分。对该协议的遵守是自愿性的；该协议未设强制执行机制。

[6] 2024 年 6 月 13 日欧洲议会和理事会关于制定人工智能统一规则的第(EU)2024/1689 号条例（《人工智能法案》），第 53 条第 1 款 d 项（通用人工智能模型提供商有义务“根据人工智能办公室提供的模板，编制并公开提供关于用于训练通用人工智能模型的内容的充分详细的摘要”）。

[7] 然而，被抓取素材的来源偶尔会从生成模型的行为中显现出来：某款知名图像生成模型的早期版本曾复现了可识别的某图库机构水印图案，这表明该机构的大量图片未经授权便被摄取。这一现象被广泛报道，并在第三部分 A.1 节所讨论的诉讼中被重点提及（Getty Images（美国）公司诉 Stability AI 有限公司）。这一现象表明，尽管权利人通常无法察觉某部作品是否被抓取，但训练后模型本身可能会通过其输出揭示出受保护材料曾被用于训练。

[8] 见上文脚注 2。

在 RAG 系统中，会维护一个外部语料库，通常在将每份文档或文档片段转换为所谓的嵌入[9]后，将其编入数据库的索引中。当系统收到用户查询时，会将其嵌入并用于从语料库中检索最相关的片段。随后，系统会将用户的查询与检索到的片段结合起来，汇编成要输入给语言模型的提示词，该模型则会根据其训练数据和检索到的内容来生成输出结果。

就本文而言，RAG 有时被认为能够改变语言模型训练中产生的版权问题。其理由在于，模型本身不从语料库中学习，因为语料库是单独存储的，仅在推理时才被调用。构建索引语料库需要对源文档或其组成片段进行预处理，并在传统的 RAG 架构中对其进行存储。[10]将每份文档转换为嵌入向量会生成源自该文档的数值表示，不过，与存储的源文档副本不同，嵌入向量通常无法逆向还原为原始文本。在运行（推理）时，相关片段会被复制到模型的上下文窗口中，并可能在输出中被全部或大部分复现。这些步骤涉及复制权，也可能涉及改编权（《伯尔尼公约》第 12 条），取决于适用法律及表现形式的特征。下文将讨论这些权利的可能适用情况。

B. 在训练中使用数据

训练生成式人工智能模型是一系列操作，通常被划分为四个阶段：预训练、监督微调、奖励建模和强化学习。[11]其中第一阶段通常要求最高，无论是在计算成本还是数据消耗量方面；后三个阶段往往合称“微调”，有时也被称为“后训练”。模型评估与优化是一项独立但相关的活动，它贯穿于训练过程，并延续至模型的部署阶段。

就版权而言，核心要点在于，上述每一阶段通常均涉及对基础语料库所有部分的存储与处理。训练流程并非只是短暂查阅数据后便将其丢弃；数据会被存储、加载、转换，并用于更新模型。如果数据中包含受保护的作品，则这些操作涉及与第二部分中进行的复制分析相关的复制或转换。

1. 训练模型

主要训练运行（通常称为预训练）是模型获取其大部分语言知识（对于语言模型而言），或图像、音频和视频统计知识（对于多模态模型而言）的阶段。预训练分为若干技术步骤，每个步骤均在语料库的存储副本上进行。首先对文本内容进行分词处理，即将单词和其他序列转换为从

---

[9] 词嵌入将每个词元表示为一个由实数值组成的稠密高维向量。当赋予某个词元的表示形式取决于其周围的词元时，这类嵌入被称为“上下文相关的”。见下文注 10《问题的核心》第二部分 C 节。

[10] 从更具技术性的角度来看，应当区分用于向量搜索的段落嵌入与生成此类嵌入的源文本。嵌入是由编码器生成的稠密数值表示；它并非文本的精确复制件，通常也无法通过简单地解码该向量来还原。尽管如此，专门的“嵌入反转”技术仍能从其嵌入中重建相对较短的源段落的部分内容，或在特定实验条件下，重建其全部内容。在传统的 RAG 系统中，源文档或片段也会以文本形式保留下来，储存在向量存储库中，或在与向量关联的独立文档存储库中。嵌入用于识别相关段落；相关文本（而非嵌入向量）随后被检索，并被作为输入上下文的一部分提供给语言模型。见 John X. Morris 等，“Text Embeddings Reveal (Almost) As Much As Text”，Proc. 2023 Conf. on Empirical Methods in Natural Language Processing (EMNLP) (2023)。有关检索流程的行业说明，见：https://infohub.delltechnologies.com/en-us/p/demystifying-on-device-intelligent-search-using-rag-architecture/

[11] 本文所采用的四阶段描述源自技术文献，详见 The Heart of the Matter: Copyright, AI Training, and LLMs71 J Cop Soc’y of the USA 482 (2024)（以下简称“Heart of the Matter”），可查阅：https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4963711。韩国文化体育观光部与韩国著作权委员会发布的《生成式人工智能模型训练适用合理使用原则指南》（2026 年 2 月）（以下简称《韩国合理使用指南》）提出了类似的四阶段流程（数据收集、数据预处理、模型训练以及模型评估与优化），见第 10 页。

固定词表中提取的一串数字词元。[12]现代分词器通常在子词而非完整词层面进行处理。[13]这一步骤对于本分析至关重要，因为就主要当代模型所使用的字节级分词器而言，该步骤是无损且可逆的：经此类分词器编码后再解码的文本可还原为原始文本，因此分词后的语料库并非作品的摘要或抽象表达，而是以另一种形式对作品的重新编码。

随后，这些词元被表示为嵌入。[14]这种方式自 2017 年以来主导该领域的模型架构，即 Transformer（缩写 GPT 中的“T”，GPT 全称为 Generative Pre-trained Transformer，即生成式预训练变换器），通过一种称为“注意力”的机制对这些表征进行处理。[15]关于嵌入以及训练后模型本身的特征描述（其中涉及更高程度的变换），将在下文 C.2 节中讨论。

模型初始化完成后，通过反复向其输入从语料库中抽取的一系列词元来训练。该模型基于前面的词元依次预测每个词元，并通过多次小幅增量进行调整，使预测结果更接近观察到的文本；因此，在整个训练运行过程中，语料库并非仅被查询一次，而是在整个训练运行过程中逐词元地被使用。在处理了从语料库中提取的海量词元后，该模型已内化了一个在庞大的可能词元序列空间内的巨大概率分布。正是这种编码在模型参数中的分布，使得模型能够根据提示词生成连贯的文本。

训练数据与训练后模型之间的关系一直是备受广泛讨论的话题。有观点认为，由于训练后模型仅包含参数而不包含原始训练数据，因此训练完成后，受保护作品的复制件已不复存在。[16]这些参数描述的是一个在极其庞大的向量空间上的概率分布。虽然这种分布使模型能够进行泛化并生成完全新颖的序列，但它可能会给训练过程中反复出现的特定精确序列分配过高且集中的概率。由于这种分布在这些有限的训练路径上具有很高的权重，因此在许多情况下，可以诱导模型逐字或近乎逐字地重现这些路径。该现象在技术文献中有大量记载，下文 C 节将对此进行进一步讨论。如下文所述，该现象（通常被称为“记忆”）与从版权法角度来看，在用于训练的所有作品的模型内部是否存在“复制”这一问题有所区别。

对于生成图像、音频或视频的模型，最常见的模型被称为“扩散”。它们的学习方式为：逐步向训练样本中添加噪声，然后学习逆转这一过程，从而从噪声中生成新的样本[17]。这不同于语言（文本）模型，但两者的训练过程都会生成对输入数据的复杂表征。

---

[12] 关于分词，见上文脚注 10 所引 Heart of the Matter，第二部分 B 节。《韩国合理使用指南》将分词列为数据预处理的技术步骤之一，与清洗和归一化并列。见《韩国合理使用指南》，上文脚注 10，第 10 页。

[13] 词汇表的构建从单个字符开始，然后逐步将出现频率最高的相邻字符对合并为更大的单元，使得常见词由单个词元表示，而较少见的词则被拆分为若干个词元。OpenAI 的生成式预训练转换器（GPT）模型以及其他多家主要开发者均采用字节对编码。见上文脚注 9 所引 Heart of the Matter，第二部分 B 节。

[14] 更具体而言，这些词元被映射到一个高维向量空间，在该空间中，每个词元都被表示为一个嵌入。

[15] 注意力机制使模型在计算某一词元的表征时，能够权衡序列中每个其他词元的相关性，从而捕捉较长文本段落中的依赖关系。Ashish Vaswani 等，“Attention Is All You Need”，31 Conf.Neural Inf.Processing Syst.(NeurIPS) (2017)。Transformer 架构是当今主要大语言模型的基础。

[16] 见上文脚注 10 所引《问题的核心》，第二部分 C 节。

[17] 关于扩散模型的起源，见 Jascha Sohl-Dickstein 等，“Deep Unsupervised Learning using Nonequilibrium Thermodynamics”，Proc.32nd Int’l Conf. on Machine Learning (ICML) (2015)；Jonathan Ho、Ajay Jain 和 Pieter Abbeel，“Denoising Diffusion Probabilistic Models”，34 Conf.Neural Inf.Processing Syst.(NeurIPS) (2020)；关于 Stable Diffusion 采用的且在第三部分讨论的 Getty 诉讼中成为争议焦点的潜在扩散架构，见 Robin Rombach 等，“High-Resolution Image Synthesis with Latent Diffusion Models”，IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)。如需通俗易懂的入门介绍，另见 Yang Song and Stefano Ermon，“Generative Modeling by Estimating Gradients of the Data Distribution”，33 Conf.Neural Inf.Processing Syst.(NeurIPS) (2019)，可查阅：http://papers.neurips.cc/paper/9361-generative-modeling-

2.优化与强化学习

预训练之后的各个阶段负责将经过预训练的模型适配至特定任务、特定用例以及特定的行为标准。虽然不同开发者使用的术语各不相同，但流程都包括监督微调、奖励建模和强化学习。[18]

监督微调是指在规模较小且经过精心筛选的数据集上进行进一步训练，该数据集包含由模型预期生成的输入-输出对组成。该数据集的规模远小于预训练语料库，但其内容均经过质量筛选，且通常由人工构建或人工编辑。每个微调示例，就像每个预训练示例一样，都会被加载并用于更新模型的参数。

奖励建模涉及构建一个“奖励”模型，该模型根据其理想程度对主模型的候选输出进行评分。该奖励模型本身是基于人类生成的候选输出的对比进行训练的。基于人类反馈的强化学习（RLHF）是一种广泛用于使生成式人工智能模型与人类偏好保持一致的技术，它利用奖励模型来引导主模型的进一步更新。

上述每个阶段都需要对训练材料进行存储和处理。这些训练材料的规模虽小于预训练所需的规模，但训练材料本身可能包含受保护的作品，特别是当人工标注者提供的对比基于此类作品或包含此类作品时。从版权角度来看，关注材料的质量而非数量并不会改变分析。

模型评估与优化在训练阶段同步进行，并在部署后持续进行，其过程涉及利用评估数据集来评估模型的性能，并识别需要进一步训练的方面。评估数据集本身可能包含受保护的作品，特别是在基准数据集是基于公开可用的测试语料库构建的情况下。

C.训练完成后

训练完成后，会产生三个不同的问题。第一个问题是语料库是否会被保留、由谁来保留，以及后续将用于何种目的。第二个问题是该模型是否以及在何种意义上包含其训练所依据的作品的复制件。第三个问题是，输出进行了复制、改编，还是以其他方式涉及训练中使用的作品。

1.训练语料库的其他用途

训练运行完成后，训练语料库仍然具有相关性。开发者可能会出于若干目的保留训练语料库：评估模型的未来版本、支持对特定子集进行进一步微调、推动持续研究，或者通过许可或其他方式提供给第三方。训练语料库可以与新汇编的材料相结合，形成一个更大的语料库，用于构建下一代模型。它也可能会被出售或共享（例如，与关联组织共享）。在某些情况下，语料库的演绎作品（包括经过过滤或去重处理的子集）会根据宽松的许可协议发布到公有领域。

此类后续使用中的每一项都可能涉及相关国家及国际规则所界定的复制、传播和提供行为。换言之，一份已经完成训练用途的语料库在之后数年里仍可继续被使用，而其后续使用的规则与最初汇编该语料库时所遵循的规则是一致的。

2.复制件在模型中的留存

训练后模型本身是否包含其训练所依据的作品的“复制件”，是本部分讨论的技术难度最大的问题之一，也是生成式人工智能与早期技术之间的区别所在。直到最近，人工智能开发者一方

---

by-estimating-gradients-of-the-data-distribution.pdf。本文此处讨论的一般性观点，并不取决于任何特定类别模型的数学细节。

[18] 见上文脚注 10 所引 Heart of the Matter，第二部分 D 节。“预训练”、“监督微调”、“奖励建模”和“强化学习”这些术语在人工智能开发者中被广泛使用。

与权利人一方所持的立场一直截然对立。[19]人工智能开发者通常认为，训练后模型仅包含参数，而不包含作品本身，且就版权法而言，并未保留任何复制件。权利人认为，该模型实际上是对整个语料库的复制。过去两年积累的实证证据表明，这两种观点都倾向于过度简化记忆与模型之间的关系。而且，无论该争议如何解决，都不影响先前的行为：语料库的汇编及其在训练中的使用均涉及复制行为，其法律定性与已完成模型本身是否包含复制件无关。

在分析中可以采用两类技术证据。记忆是训练后模型的一种属性，即该模型能够以足够高的保真度从其参数中还原特定的训练序列。由于难以直接观察模型的内部结构，因此通过“提取”来研究记忆，即模型在适当的提示下所表现出的逐字或近乎逐字地再现训练数据中序列的能力。因此，提取被用作记忆的证据。某项提取程序的失败并不能证明某部作品未被记忆。近期由库珀（Cooper）等人开展的一项系统性研究，将一种概率提取技术应用于 14 个开放权重语言模型中的 200 本书中，结果发现，记忆程度随模型、图书的不同而存在显著差异。[20]大多数最大的模型并没有记住大部分图书的全部或部分内容。然而，一旦发生记忆，其程度可能相当显著：该研究称，一种被广泛应用的模型几乎完全记住了《哈利·波特》系列的第一部小说和乔治·奥威尔的《1984》；就前者而言，研究证明该模型几乎可以逐字生成整本书的内容。

然而，这些关于记忆的实证论证并非没有争议。人工智能开发者对此回应称，诱导系统逐字复制所需的提示技术具有“对抗性”，不能代表系统在实际中的使用方式，且此类行为本身可能构成对产品的操纵或滥用，超出了正常使用范围（在某些情况下，甚至违反了《服务条款》）。在《纽约时报》诉 OpenAI 一案的诉讼中，被告提出动议，请求驳回直接版权侵权主张，其部分理由是，原告使用被告所称的“精心设计且经过操控的提示词”，人为制造出原告所指控的复述。[21]2025 年 4 月，纽约南区地方法院拒绝以该理由驳回上述主张。此后，该案件的争议从证据开示阶段的“复述”是否可人为制造，转向其在实际中自然发生的频率，最终法院作出裁决，强制要求提交 2,000 万条真实用户交互日志作为代表性样本。[22]关于该诉讼以及该命令的意义，将在本研究报告第三部分 A.4 节中进一步讨论。

---

[19] 这种分歧不仅体现在诉讼立场方面，还延伸至设计优先事项。一项针对人工智能伦理学家、法律专家、音乐人工智能从业者及艺术家开展的多阶段德尔菲研究，旨在验证一套适用于生成式音乐系统的负责任设计框架。该研究发现，在数据主权问题上，艺术家与法律专家均倾向于“保护主义”， 这与从业者侧重可行性的立场形成鲜明对比。研究结论指出，此类框架最适合作为各利益攸关方不同优先事项之间的“转换层”——包括在模型透明度与知识产权利益之间的权衡——而非作为强制达成共识的工具。J. de Berardinis 等，“RAIM: Operationalising Trustworthy AI in Music Generation”，IEEE Transactions on Technology and Society，Special Issue on Ethical Innovation with/in Music Technology（2026 年即将出版），框架网站：https://casbahlab.github.io/raim-framework/，于 2026 年 8 月 25 日访问。

[20] Feder Cooper、Aaron Gokaslan、Ahmed Ahmed、Amy B. Cyphert、Christopher De Sa、Mark A. Lemley、Daniel E. Ho 和 Percy Liang，“Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models”，arXiv:2505.12546 (2025)，2025 年 11 月 30 日版本（亦作为斯坦福大学公共法律工作论文及西弗吉尼亚大学法学院 2025-005 号研究论文发表）。作者对该问题的表述（见第 1 页）值得全文引用：“在涉及生成式人工智能的版权诉讼中，原告和被告往往就大语言模型（LLM）在多大程度上记住了原告受保护的表达，提出笼统且相互对立的主张。通过结合机器学习和版权法，我们证明了这些对立的观点极大地简化了记忆与版权之间的关系。”关于记忆内容必须编码在模型参数中（因为“别无其他存储位置”）这一观点，另见 A. Feder Cooper 和 James Grimmelmann（该文第 25 页所引）。

[21] 纽约时报公司诉微软公司，案号：1:23-cv-11195，《支持被告驳回动议的备忘录》（纽约南区联邦地区法院，2024 年 2 月 26 日）；意见与命令（纽约南区联邦地区法院，2025 年 4 月 4 日）（斯坦法官）（就相关部分驳回针对直接侵权与帮助侵权主张的驳回动议）。关于 OpenAI 自身对该论点的公开表述，见 OpenAI，“Reporting the Facts About the New York Times’ Lawsuit”，可查阅：openai.com/new-york-times。

[22] 关于 OpenAI 公司版权侵权诉讼案，案号：1:25-md-03143-SHS-OTW（纽约南区联邦地区法院），意见与命令（斯坦法官，2026 年 1 月 5 日），维持意见与命令（Wang 治安法官，2025 年 12 月 2 日）（驳回 OpenAI 的复议动议）以及 2025 年 11 月的基础证据开示命令。法院下令制作从已保存的数据集中抽取的 2,000 万条 ChatGPT 用户交互日志的随机、去

总而言之，实证研究结果使这两种极端立场都变得复杂起来：训练后模型通常并非其训练语料库的“复制件”，但也不完全“脱离”该语料库。确切范围因模型而异。

第二类证据是关于数据提取攻击的更广泛文献，这些文献记录了通过定向提示可从训练后模型中还原个别训练样本的情形。[23]所还原的样本包括篇幅较长的文本段落、可识别的图像以及个人身份信息。上述研究成果被引用为证据，表明可提取性可以是大型生成式模型的一种结构属性，而不仅仅是用户操作的结果。一些人工智能公司称，此类提示属于不公平的操纵，或违反了服务条款。[24]然而，过去一年中，此类评估变得更加复杂：已部署的系统日益在模型与用户之间设置过滤机制和其他防护措施，以便减少提取已记忆内容的能力；近期研究对某些系统不得不采用对抗性技术，而另一些系统则无需此类技术即可输出受保护内容。[25]

有一个类比可能有助于理解这一情况，即与编译型或解释型计算机程序的类比[26]。当某一程序的源代码（通常适用类似自然语言的语言编写）被编译成可执行形式时，其结果是不同于源代码的，且（不易）为人类读取。然而，编译过程并非与源代码毫无关联：可执行文件保留了源程序的信息，通过逆向工程，可以在很大程度上重建源代码。从这个意义上说，可执行代码仍与源程序相似。如果按照这个类比，由训练语料库生成训练后模型的步骤，可以说与上述编译过程类似。在这两种情况下，可理解的源对象（源代码形式的程序；作品语料库）均被转换为一个功能性数值对象，该对象虽然不再与其源对象相似，但仍然承载着由该源对象所衍生的信息。然而，这种类比具有局限性：编译是一个确定性过程，会将整个源程序带入可执行文件中，而模型中对训练数据的编码则是选择性的，且在一定程度上具有概率性，因此该类比至多仅适用于已证实发生记忆的作品子集；近期学术研究认为，这种确定性可还原性与单纯概率性可还原性之间的差异，恰恰可能是对模型进行法律定性时的关键所在。[27]

本文审查的技术证据表明，即使经过了大幅变换，且对任何特定作品的访问可能是部分的且具有概率性的，训练后模型仍会保留其训练数据中作品所衍生的信息。这项技术层面的观察结论如何转化为对模型的法律定性，以及从版权法角度来看，如何回应模型本身是否是其训练所依据作品的复制件或改编/演绎作品这一问题，是法院与学术界文献中正在持续分析的议题，正如本研究后文所述。

标识化样本，理由是，面向消费者的广泛输出证据与合理使用分析相关，与模型输出是否发挥了原告作品市场替代作用这一问题尤其相关。

[23] 这类的早期研究包括 Antonia Karamolegkou 等，“Copyright Violations and Large Language Models”，arXiv:2310.13771 (2023)（研究表明，在多种语言模型中，通过数据提取攻击可以还原个别训练数据样本）。

[24] 见 Nicholas Carlini 等，Extracting Training Data from Large Language Models，30th USENIX Security Symposium (2021) 2633； Milad Nasr 等，Scalable Extraction of Training Data from Aligned, Production Language Models，ICLR 2025（扩展自 Nasr 等，arXiv:2311.17035（2023））。

[25] Ahmed Ahmed 等，“Extracting Books from Production Language Models”，arXiv:2601.02671 (2026)（研究发现，从某些已部署的系统中提取出记忆的书籍文本需要使用对抗性提示，但从其他系统中提取则无需如此）。

[26] 见上文脚注 10 所引 Heart of the Matter，第二部分 C 节（该部分将语言模型的训练与计算机程序编译成可执行形式进行了类比，并指出，在这两种情况下， “原受版权保护作品”均被“转换为人类无法直接理解的另一种形式”，但原作品的内容“[并非]不存在于那个机器可读版本中”）。

[27] M.A.Lemley 和 A.F.Cooper，“Probabilistic ‘Copies’ in Generative AI Models”，Berkeley Technology Law Journal（2026），可查阅：https://ssrn.com/abstract=7067878 （基于本部分所讨论的实证提取文献，该文称，只有当某一特定作品能够以直接且在最极端的情况下具有确定性的方式被提取出来时，才应将模型视为包含该作品的“复制件”；若提取仅在多次尝试后在概率意义上成功，则所存储的并非该作品的复制件，而是一组“构成要素”；因此，该问题属于证据问题，需针对每部作品和每个模型分别予以解答）。

在此可以得出两个观点。首先，记忆这一概念既被用作训练阶段的论据，即作为模型训练过程中涉及对副本进行编码的证据；也被用作输出阶段的论据，即作为部署后的模型能够生成以替代原作形式存在的受保护作品的证据。其次，原告所依据的示例涉及特定模型中的特定作品。这些示例虽在一定程度上证实，对于特定、个别且频繁出现的作品而言，记忆可能具有实质性影响，但并不能证明这是所有模型的统一特征，也无法证明模型的参数在一般情况下可归结为存储的副本。关于这些问题的法律定性将在第三部分中进行探讨。[28]

最后，有必要注意的是，模型在训练完成后可能更改，方式包括微调、对齐训练，或针对特定内容尝试机器遗忘。一个模型在经过此类干预后不再复现某一特定作品，并不能据此证明模型自始就未曾对该作品进行编码。技术文献表明，要从训练后模型中可靠地移除特定作品仍十分困难，而且此类干预可能会以意想不到的方式导致模型性能下降。[29]

3.输入-输出关系

模型部署完成后，其运行方式为接收用户输入或提示词，并生成输出内容。输出内容可以是文本、代码、图片、音频、视频或其他形式的内容。输入与输出之间的关系在某些方面类似于搜索引擎中查询与响应之间的关系，但这种类比并不完全成立。模型并非查询已建立索引的语料库并从中返回段落；而是通过对训练期间所内化的概率分布进行采样，来生成新内容。在某些情况下，模型生成的内容与训练数据中的特定作品密切相关；而在其他情况下，输出的是全新的组合内容，与任何单个训练样本都不太相似。

有三种看法尤其相关。首先，输出内容可能是对训练数据中某部作品的逐字或近乎逐字的复制；在这种情况下，该输出即构成对原作品的复制。[30]这一点已在关于语言模型复现受版权保护书籍中的段落，以及图像生成模型复现受保护照片和艺术作品的研究中得到证实[31]。此类复制是否构成侵权，将取决于一系列其他问题，包括例外和限制的适用性、开发者部署的任何输出过滤器的作用，以及模型开发者与发出提示词的用户之间的责任归属。本研究后文将简要提及这些问题。[32]

其次，输出内容可能与训练数据中的某项作品构成实质性相似，但并非逐字复制。根据某些国家的版权法中常见的“实质性相似”原则等规定，这种行为可能构成侵权。

第三，输出内容可能是全新的，即无法将训练数据中的任何特定作品识别为其来源。生成式模型能够生成与任何单个训练样本都不一致的内容。事实上，在大多数情况下，这正是模型的运作方式。在这种情况下，输出与任何特定受保护作品之间的关系会减弱。输出是否在版权法所认

---

[28] 关于合并诉讼，见“关于 OpenAI 公司版权侵权诉讼案”，案号：1:25-md-03143-SHS-OTW（纽约南区联邦地区法院），该案合并审理了《纽约时报》及多个作者团体针对 OpenAI 和微软提起的诉讼。将训练后模型定性为一种存储和检索机制，将参数定性为所摄入作品的压缩形式，反映了原告的诉状内容；而将参数定性为编码语言模式的统计概括，则反映了被告的立场。关于警告不应采用这两种定性的实证文献，见本节前文引用的库珀等人的研究。

[29] A. Feder Cooper 等，“Machine Unlearning Doesn’t Do What You Think: Lessons for Generative AI Policy, Research, and Practice”，39 Conf.Neural Inf.Processing Syst.(NeurIPS) (2025)。

[30] 当大语言模型输出与某部现有的受保护作品相似，而该模型在训练过程中并未使用该作品，也未以其他方式接触过该作品时，是否构成“复制”就不太明确了。

[31] 语言模型进行逐字复现的情况已在学术研究和诉讼文件中得到证实；图像生成模型对可识别照片和艺术作品的复现，已在扩散模型的相关情形中得到记录。已记录的案例数量并不直接影响对任何特定输出的法律定性，该定性取决于适用的国内法和案件事实。

[32] 关于输入提示词、模型与输出之间的关系，以及开发者与用户之间应如何划分责任这一开放性问题，见下文第三部分。

可的任何意义上仍“基于”训练数据，这一问题已在若干司法管辖区提交法院审理，但目前尚未形成共识。

因此，第一部分指出了在生成式人工智能模型开发和运行的以下多个阶段中受保护作品的使用情况，以及涉及这些作品留存或复制的问题：最初汇编的语料库、训练期间对语料库实施的存储和处理操作、训练后模型本身，以及部署后生成的输出。在每个阶段，国家版权立法和国际条约所承认的主要专有权都可能涉及其中。关于每次复制是否可能构成初步侵权（视例外或限制而定），以及现有或可开发何种机制用于对相关使用授予许可，将在下文各部分中予以探讨。

第二部分 授权、许可和补偿

本研究第一部分从技术层面指出了在生成式人工智能模型的开发和运行过程中，各国版权立法及国际版权条约所承认的主要专有权可能涉及其中的环节。本部分将从这些权利的讨论转向在人工智能背景下行使、放弃或替代这些权利的机制。关于相关使用在未经授权的情况下是否合法的问题，以及若不合法，开发者如何使自身运营符合法律规定、权利人又应如何在此过程中获得补偿的问题，最终归结为一套法理体系：版权许可法律，包括个人许可与集体许可，自愿许可与法定许可，以及决定是否首先需要获得许可的例外与限制。

本部分的结构如下。A 节探讨了人工智能开发者目前获取用于训练其模型的受版权保护客体的主要渠道，以及各渠道的法律地位。B 节探讨了当前影响许可做法的主要政策问题：受训练影响的权利、自愿许可、三种主要形式的集体许可、欧盟文本与数据挖掘例外条款的现状及其与人工智能训练相关性的学术讨论，以及报酬方案。提及这些国家只是为了举例说明世界各地采取的不同做法；它们并非分析的对象，且本调查并非详尽无遗。

A.开发者获取可受版权保护客体用于训练的现有方案

1.五种获取途径

实际上，人工智能开发者通过以下五种方式之一，或这些方式的某种组合来获取训练数据。其中一种是使用由第三方创建的开放或精选数据集，这些数据集有时免费提供，有时则需付费。一个突出的例子是 Common Crawl。这是一个由非营利基金会爬取并提供的、持续更新的网页存储库，它直接或间接地支撑着大多数主要语言模型的文本训练语料库。LAION-5B 及其前身在图像生成模型领域发挥同等作用。[33]此类数据集本身就是大规模自动化检索操作的结果，可能延续与底层检索相关的版权问题。法律立场将取决于原始资料的获取依据，以及适用于该数据集后续使用的法律。这些经过精心整理的数据集在公共部门也开始出现对应版本：在大韩民国，公共数据许可系统（Gonggong Nuri）已得到更新，新增了一类许可，根据这类许可，公共数据集会明确标示其已获得授权，可用于人工智能模型训练。[34]

---

[33] 关于 Common Crawl 的介绍见 https://commoncrawl.org。LAION-5B 数据集（用于构建 Stable Diffusion 及其他图像生成模型）及其前身 LAION-400M 是由大规模人工智能开放网络编纂的、向公众开放的图像-文本数据集。该数据集本身并不包含原始图像，而是包含 URL 与文本描述的配对；图像是在训练时从其原始位置调取的。关于将数据集汇编操作在法律上定性为文本与数据挖掘的问题，见 Kneschke 诉 LAION e.V.案，汉堡地区法院，2024 年 9 月 27 日（310 O 227/23）；以及二审，Kneschke 诉 LAION e.V.，汉堡高等法院，2025 年 12 月 10 日（5 U 104/24）。

[34] 此处指的是“Gonggong Nuri”（韩国开放政府许可）体系，根据该体系，公共机构会按照标准化许可类别公开数据以供再利用；更新后的框架新增了一类许可，明确允许将数据用于人工智能模型训练。

另一个来源是所谓的“影子图书馆”，即未经授权的大型书籍、文章、音乐及其他作品存储库，据报道或据称，许多广泛使用的模型曾直接或通过包含这些内容的训练数据集使用过这些内容。下文第三部分 A 节中讨论的美国盗版相关认定，正是依赖此类存储库所引发的法律风险的例证。

为了进行比较，一些法院将非法获取作品与另一个问题——在训练中使用作品是否属于例外或限制的范围——区分开来。在下文讨论的一项美国裁决中，下载和留存盗版复制件被认定不属于合理使用，而基于合法获得的同批作品复制件对模型进行训练则被认定具有转换性且属于合理使用。从这个意义上讲，获取的合法性与训练的合法性在分析层面上是相互独立的。这一区别至关重要，因为同一个模型内的训练数据可能既来自合法来源，也来自非法来源，从而增加了许可和权利执行的复杂性。因此，开发者可能既需要合法获取数据集，也需要为训练本身确立合法依据，同时须遵守适用的国家法律。

第三种方法是个人许可和集体许可。开发者可以购买精选语料库的使用权，或与出版商、新闻机构、广播机构、图片库、学术数据库运营商或任何机构权利人进行协商，以获取使用权。在过去三年里，许可已成为一种常见的做法，各大人工智能开发商纷纷与欧洲、美洲和亚洲的主要新闻机构及图库签署协议。[35]许可的结构性特征在于，开发者至少就许可方自身的权利获得了被许可语料库的权威权利链。下文 B.2 节讨论了该链条的局限性。在某些情况下，合同中包含内容的交付。

第四种途径，也是从数据量来看的一条主要途径，是由开发者自身（而非第三方）通过网络爬取和抓取，直接从开放网络中收集材料。本研究第一部分 A.2 节对技术程序作了说明。自动化检索的规模相当可观：据部分估算，自动化代理目前已占全球网络流量的绝大多数。[36]爬取通常未征得被检索材料权利人的事先同意，其依据为以下三种法律理论之一：该材料属于公有领域，或者已基于宽松条款有效许可给公众；该操作属于可适用版权法中的例外或限制范围（在欧洲联盟，适用《数字化单一市场版权指令》（《CDSM 指令》）第 4 条中的文本与数据挖掘例外；在美利坚合众国和大韩民国（依据《版权法》第 35-5 条），适用一般性合理使用原则；在新加坡和日本，则适用具体的法定条款）；或者该操作根本不涉及复制权，因为所生成的是临时复制件或暂时复制件，或者在某些司法管辖区，此类复制件属于临时复制行为的明确例外（如欧洲联盟《信息社会指令》第 5 条第(1)款所规定）。上述理论并不互斥，开发者可能会针对语料库的不同部分使用不同的理论。

第五种途径是合成数据：用现有人工智能模型生成的数据训练或微调另一个模型。合成数据是一种应用日益广泛的工具，最常用于模型开发的后期阶段，尤其是监督微调和奖励建模。[37]从版权的角度来看，合成数据带来另一个问题。如果生成合成数据的模型是基于受保护的作品进行训

---

[35] 关于人工智能开发者与出版商之间的双边许可协议，见（众多协议中的）以下实例：OpenAI 与美联社、新闻集团、《世界报》、普利沙传媒、阿克塞尔·施普林格、沃克斯传媒、《大西洋》和赫斯特集团达成的协议；谷歌与 Reddit 和美联社达成的协议；以及 Anthropic 与多家出版商达成的协议。这些交易的披露细节程度各不相同，其财务条款属于商业机密。

[36] 见 Imperva，2025 年《恶意机器人报告》（第十二版，2025 年）指出，2024 年自动化流量占互联网总流量的 51%，十年来首次超过人类流量，其中所谓的“恶意机器人”占比达 37%。

[37] 合成数据是指由现有人工智能模型生成的、用于训练或微调另一个模型的数据。在生成模型本身利用受版权保护作品进行训练的情况下，合成数据便回引发不同的版权问题。关于其在模型开发后期阶段应用日益广泛的情况，见 R. Liu 等，“Best Practices and Lessons Learned on Synthetic Data”，arXiv:2404.07503 (2024)，该论文在首届语言建模会议上发表。

练的，那么合成输出是否也复制或改编了受保护的表达？该问题尚未以能够产生具有约束力判决的方式提交法院。

2.合作安排

最后，除了上述五大主要途径外，在某些司法管辖区，我们还观察到人工智能开发者与机构权利人或国有权利人之间，存在与特定行业应用最为相关的有组织合作形式。中国就是一个例子。在垂直领域大模型的开发过程中，出现了以下合作：通用人工智能模型开发者与香港广播电视台（针对粤语语音数据）之间的合作；开发者与拥有大量签约作家作品库的在线出版商（针对数字小说）之间的合作；开发者与通过转让许可或专有许可获得数十亿张图像的大型图片库（针对视觉训练数据）之间的合作，以及开发者与学术数据库运营商联合搭建专业模型的合作。据称该合作模式应用广泛，不过相关协议本身往往并不以版权许可协议命名，且其确切范围也并非总是予以披露。如下文 B.2 节所述，该方法的结构性局限在于：出版商、广播机构或存储库通常仅持有原始作者授予的特定权利，而其自身也无法确定所持有的这些权利是否涵盖人工智能训练。

B.主要政策要素

1.受生成式人工智能模型训练影响的权利

(a)复制权

初步来看，第一部分所述的技术操作涉及国家版权立法和国际版权条约所承认的若干专有权。复制权是本分析的核心。[38]训练语料库的构建涉及制作复制件；在训练的各个阶段对该语料库进行存储和处理，会产生更多复制件，其中部分是持续复制件，部分是暂时复制件；根据第一部分 C.2 节所载技术分析，经训练形成的模型的某些方面本身也可被视为该模型训练所用作品的复制件。初步来看，上述每项操作均属于《伯尔尼公约》第 9 条第(1)款、《产权组织版权条约》第 1 条第(4)款、WPPT 第 7 条和《北京条约》第 7 条，以及各国相关立法所认定的复制权范围。

对于临时复制件和暂时复制件的处理需要单独考虑，因为两种不同的法律技术会产生相似的实际结果：在某些司法管辖区，此类复制件不被认定为为属于复制权的范围；而在其他司法管辖区，则通过明确的例外规定予以容许。中国是第一种技术的主要例证：临时复制一贯被认定不属于该权利的范围。在中国司法实践中，被排除在外的临时复制是在随机访问内存或缓存中生成的、对数据处理过程严格必要的暂时复制件；关于副本存在多长时间仍可被视为临时复制件，目前尚无定论。中国《著作权法》第十条第（五）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制作一份或者多份的权利，[39]中国代表团及其他一些代表团曾在外交会议上对一项议定声明投反对票，该声明本拟把临时数字存储明确纳入复制权的范围。[40]《信息网络传播权保护条例》（2006 年）中未包含将临时复制认定为复制行为的条款，中国法院在涉及转码及类似暂时数字操作的案件中均采纳了这一理解。从这一角度来看，就训练行为本身而言，只要其仅涉及对存储在内存或缓存中的作品进行暂时访问，且该访问在训练运行完成后即消失，则其依据中国法可能不涉及复制权，但构建底层作品数据库无疑会涉及复制权。在其他司法管辖区，则是通过明确的例外来实现相同的实际效果：在欧洲联盟，《信息社会指令》

[38] 见第一部分 A.1 和 A.2 节。
[39] 《中华人民共和国著作权法》第十条第（五）款。
[40] 见产权组织文件 CRNR/DC/101，全体会议简要记录，第 497 至 502 段。尽管如此，关于 WCT 第 1 条第（4）款的议定声明仍在外交会议上通过；中国投了反对票。

第 5 条第(1)款通过一项界定严格的例外来容许暂时复制件和附带复制件，而非自始就将其排除在复制权之外。两种制度所涉及的复制件在实质上是相同的；区别在于所采用的法律技术。[41]在本报告调查的大多数其他司法管辖区，训练行为通常被视为涉及复制权，而分析的重点转向是否可适用例外或限制规定。

(b)改编权和演绎作品

除了复制权之外，还涉及或可能涉及另外三项权利。《伯尔尼公约》第 12 条及相应的国家条款所承认的改编权，是在输出阶段被援引的，因为输出的生成被视为对原作品的改编或修改。例如，巴西法律明确将改编权列为需要事先明确授权的使用形式之一，[42]而文化部版权与知识产权秘书处则认为，生成输出既涉及改编权，也涉及复制权。加拿大、法国和德国法律中的立场大体一致。

(c)向公众传播和提供

《世界知识产权组织版权条约》（WCT）第 8 条以及《世界知识产权组织表演和录音制品条约》（WPPT）第 10 条和第 14 条所规定的向公众提供权，主要在输出阶段触发，即部署的模型向用户返回内容时，如果该内容含有受保护表达的副本。在巴西（该国不是 WCT 或 WPPT 的缔约方），最高司法法院于 2017 年在 REsp 1.559.264-RJ 案中裁定，通过互联网进行音乐流媒体播放涉及向公众传播权，该权利涵盖了其他司法管辖区可能认定的提供权。[43]这与 1996 年外交会议通过的“总体解决方案”一致，根据该方案，提供权可通过向公众传播权、发行权或两者的结合来实施。将类比推理应用于已部署的人工智能模型的输出，这一做法在巴西法院尚未经过检验。[44]表演者和录音制品制作者根据前文所述 WPPT 第 10 条和第 14 条享有对应的提供权，因此本分析的范围超出了作者权利的范畴。

(d)精神权利

精神权利也可能受到影响。保护作品完整权是被援引更为频繁的一项权利。在巴西，训练过程中对作品的拆分处理被视为可能侵犯保护作品完整权，而未对所用素材作者署名则被视为可能侵犯署名权。[45]《加拿大版权法》将保护作品完整权和署名权均认定为作者的精神权利。[46]在若干实行作者权（droit d’auteur）的司法管辖区，学术评论中曾提出如下问题：在训练语料库中使

---

[41] 2001 年 5 月 22 日欧洲议会和理事会关于协调信息社会中版权及相关权某些方面的第 2001/29/EC 号指令第 5 条第 1 款，OJ L 167/10[往往被称为“信息社会指令”]。例外适用于短暂或附带的临时复制行为，这些行为是技术工艺不可或缺的一部分，其唯一用途是促使通过中介与第三方进行网络传输或合法使用，并且这些行为不具有独立的经济意义。

[42] 《巴西版权法》（第 9,610/1998 号法）第 29 条第 III 和 IV 项（改编、音乐编排、翻译及其他改编形式）。巴西文化部在 2024 年 10 月致国家数据保护局的一份技术说明中，就某大型开发商模型的训练问题阐述了其立场，其中援引了第 29 条第 III 项（为制作演绎作品而进行的改编）；见报告：https://nucleo.jor.br/curtas/2024/10/10/ministerio-cultura-prejuizos-direitos-ia-meta/，于 2026 年 7 月 7 日访问。

[43] 高等法院，REsp 1.559.264-RJ (2017)（认定，通过互联网进行音乐流媒体传输构成一种公开表演形式，涉及向公众传播权，且 ECAD 有权据此收取费用）。

[44] 《巴西版权法》（第 9,610/1998 号法）第 5 条第 V 项（“向公众传播”的定义）和第 29 条第 VII 项（通过允许用户选择作品的任何系统进行传播）。

[45] 《巴西版权法》（第 9,610/1998 号法）第 24 条第 I 至 IV 项。根据第 92 条，表演者还对其表演享有保护作品完整权和作者身份权这两项精神权利。

[46] 《加拿大版权法》，1985 年《加拿大修订法规》，第 C-42 章，第 14.1 和 28.1 条。

用某部作品时，若未署名，且其使用方式可能产生被视为贬损或歪曲原作的输出，是否构成对精神权利的侵害？但这一问题尚未得到权威解答。

最后，在欧洲联盟以及已采纳类似保护措施的某些司法管辖区，如果训练语料库是通过提取和再利用受保护数据库的实质性部分构建的，则可能涉及第 96/9/EC 号指令保护的数据库特殊权利。[47]该权利独立于版权，并有其自身的实质性投资和“实质性部分”标准（针对“提取权”）。该权利在人工智能训练语料库构建方面的适用范围尚无定论，包括涉及法律信息数据库、科学期刊聚合平台以及大型图库的情形。

因此，生成式人工智能模型的训练可能涉及的权利可能超出复制权。因此，双边许可或集体许可能否准许相关用途，取决于其所涵盖的权利；仅限于复制的许可可能无法应对基于改编、提供、精神权利或（如适用）数据库特殊权利提出的主张。第三部分在跨境权利执行的背景下还提到，一系列权利以及每项权利的具体范围因司法管辖区而异，因此，一项旨在在某一司法管辖区内批准权利的许可安排，可能会使开发者在另一司法管辖区内面临风险。[48]

(e)权利管理信息（RMI）和反规避

根据 WCT 第 11 条和第 12 条、WPPT 第 18 条和第 19 条以及《北京条约》第 15 条和第 16 条，权利人对其为行使上述条约及《伯尔尼公约》所规定的权利而采取的有效技术措施遭到规避的行为，有权获得充分的法律保护和有效的法律救济， 同时，对于任何明知或（就民事救济而言，有合理理由知道）其行为将诱导、使之成为可能、便利或隐瞒对上述条约或《伯尔尼公约》所涵盖的任何权利的侵权，却仍实施下列行为的人，权利人也有权获得充分且有效的法律救济：

(i)未经授权删除或更改任何电子权利管理信息；

(ii)未经授权，发行、为发行目的进口、广播或向公众传播作品或作品的复制件、录音制品或表演，且（依据相关文书）明知电子权利管理信息已被未经授权地删除或更改。

据我们所知，截至本文撰写之时，尚未有针对企业因训练人工智能模型而提起的成功诉讼。

2.自愿（集体或个人）许可

(a)人工智能使用的个人和集体许可

自愿许可是指人工智能开发者与众多权利人或权利人代表（包括集体管理组织或其他中介机构）在无法定任务授权的情况下，协商授权事宜。在所有研究过的司法管辖区中，这都是默认规定。

两个结构性特征决定了自愿许可在实践中采取的形式。第一个特征是少数大型机构权利人与数量庞大的个人作者和小型权利人之间的不对等，前者能够与开发者就具有商业意义的条件进行谈判，而后者由于交易成本、缺乏专门知识或其他因素，往往难以进行此类谈判。过去三年间达成的主要双边协议，均发生在人工智能开发者与大型出版商、新闻机构、广播公司及图库之间；

[47] 1996 年 3 月 11 日欧洲议会和理事会关于数据库法律保护的第 96/9/EC 号指令，OJ L 77/20，第 7 条（特殊权利）。另见欧盟法院，C-203/02，英国赛马委员会诉威廉·希尔公司案，ECLI:EU:C:2004:695，关于实质性投资标准的条件。

[48] 见多伦多星报有限公司诉 OpenAI 公司案，2025 ONSC 6217（基梅尔法官，2025 年 11 月 7 日）（以与安大略省存在真实且实质性的联系为由，驳回了 OpenAI 关于管辖权的动议）。

而对于个人作者、摄影师、插画师或音乐人而言，则缺乏相应的机制，他/她们往往依赖出版商及其他被许可方达成的协议，才能分得一部分收益。

其中一些安排的规模如今已相当可观。据报道，某大型新闻出版商与一家人工智能开发商签订的一份为期五年的协议金额超过 2.5 亿美元，[49]用于计算用途的内容许可已成为大型图库的一项重要收入来源：Shutterstock 公司 2025 全年财报显示，其数据、分发与服务收入达 2.033 亿美元，约占 9.899 亿美元总收入的 20.5%；Getty Images 另行报告其 2025 全年收入为 9.813 亿美元。这两家公司于 2025 年 1 月 7 日宣布进行对等合并；Getty Images 于 2026 年 6 月 30 日宣布，鉴于英国竞争与市场管理局将剥离 Shutterstock 的编辑业务作为批准条件后，其将终止该合并协议。据报道，该协议已于 2026 年 7 月 7 日终止。[50]法国新闻界可体现该市场的选择性：2024 年 3 月，《世界报》（Le Monde）与 OpenAI 达成了一项协议，该报称这是法国媒体与大型生成式人工智能开发商之间的首份协议。[51]同年晚些时候，APIG 寻求与 OpenAI 进行集体讨论，但现有报道显示，这些讨论当时并未达成更广泛的许可安排。[52]针对这种不对等现象，已经开始出现一种应对形式，即以中介机构聚合小型出版商的权利：一家商业信息提供商称，其网络中的数千家出版商（从规模更大的出版商池选出）已选择许可其内容供人工智能使用；[53]而一家获得许可的问答引擎开发商报告称，其同样聚合了数量相当的参与出版物。[54]

第二个特征是权利链问题。在许多司法管辖区，出版商、广播机构或存储库已通过生成式人工智能出现之前签订的许可或转让合同，从原始作者处获得某些权利，这些权利是否延伸至将作品用于训练人工智能模型，既在法律上不确定，在合同上也未作规定。这种不确定性并非人工智能所独有：在作者与音乐出版商之间、表演者与录音制品制作者之间，以及视听作品制片者与为某项制作聘请的演艺人员之间的协议中，只要在订立合同时未预见到某种利用方式，就会出现类

---

[49] 据《华尔街日报》报道，新闻集团与 OpenAI 达成的这项为期五年的协议金额超过 2.5 亿美元。此类协议的财务条款通常属于机密，该金额系媒体报道的数字，而非由各方公开确认的金额。见“OpenAI, WSJ Owner News Corp Strike Content Deal Valued at Over $250 Million”（《华尔街日报》，2024 年 5 月 22 日）<https://www.wsj.com/business/media/openai-news-corp-strike-deal-23f186ba> 于 2026 年 7 月 3 日访问；“Sam Altman’s OpenAI signs content agreement with News Corp”（路透社，2024 年 5 月 22 日）<https://finance.yahoo.com/news/sam-altmans-openai-signs-content-201808782.html> 于 2026 年 7 月 3 日访问。

[50] Shutterstock，“Shutterstock Reports Full Year 2025 and Fourth Quarter Financial Results”（2026 年 2 月 17 日）<https://investor.shutterstock.com/node/14551/pdf>，于 2026 年 7 月 3 日访问；Getty Images 控股公司，“Getty Images Reports Fourth Quarter and Full Year 2025 Results”（2026 年 3 月 16 日）<https://investors.gettyimages.com/node/15726/pdf>，于 2026 年 7 月 3 日访问；Getty Images 控股公司和 Shutterstock 公司，“Getty Images and Shutterstock to Merge, Creating a Premier Visual Content Company”（2025 年 1 月 7 日）<https://www.sec.gov/Archives/edgar/data/1898496/000110465925001732/tm252186d1_ex99-1.htm> 于 2026 年 7 月 3 日访问。关于合并终止，见英国竞争与市场管理局，“Getty abandons Shutterstock merger plans”（2026 年 7 月）<https://www.gov.uk/government/news/getty-abandons-shutterstock-merger-plans>，于 2026 年 7 月 31 日访问。

[51] 《世界报》称，这项于 2024 年 3 月达成的协议是法国媒体机构与生成式人工智能模型主要开发商之间签署的首份协议。见“Le Monde signs artificial intelligence partnership agreement with OpenAI”（《世界报》，2024 年 3 月 13 日）<https://www.lemonde.fr/en/about-us/article/2024/03/13/le-monde-signs-artificial-intelligence-partnership-agreement-with-open-ai_6615418_115.html>，于 2026 年 7 月 3 日访问。

[52] 此处所指机构为“综合新闻媒体联盟”（APIG），该联盟与“杂志出版商协会”（SEPM）于 2024 年 6 月正式提议与 OpenAI 展开谈判；据报道，该开发商回应称，其认为现阶段尚无必要与法国出版商签订新的许可协议。见“Intelligence artificielle : un accord de partenariat entre Le Monde et OpenAI”（《世界报》，2024 年 3 月 13 日）<https://www.lemonde.fr/le-monde-et-vous/article/2024/03/13/intelligence-artificielle-un-accord-de-partenariat-entre-le-monde-et-openai_6221836_6065879.html>，于 2026 年 7 月 3 日访问。

[53] 道琼斯旗下的 Factiva 报道称，截至 2025 年底，在约 30,000 家出版商组成的网络中，约有 5,000 家选择许可其内容用于生成式人工智能系统。见道琼斯，“Factiva Sees Growing Publisher Participation in AI Licensing”（道琼斯，2025 年）<https://www.dowjones.com>，于 2026 年 7 月 3 日访问。

[54] ProRata.ai 自称其网络由“1000 多家出版物和内容创作者”组成。

似的问题；人工智能领域所独有的特点在于，该问题所涉及的规模之大，以及重新谈判的实际困难。例如在中国，多数作者将其作品的印刷出版专有权，以及日益增多的数字出版专有权授予其出版商。这些权利是指以纸质形式复制作品、发行纸质复制件，以及复制和发行数字复制件的权利。上述任何一种使用方式，均不能自然地解释为包括将作品用于人工智能模型的训练。因此，从出版商处获得版权许可的语料库提供商和人工智能开发者的立场是，就原始作者而言，这些获许可活动的合法性仍存疑。即使开发者已从机构中的权利人处获得了许可，仍存在原始作者提起诉讼的剩余风险。根据不同司法管辖区的规定，这种风险并不限于许可方：许可方可能因违反合同或法律而承担责任；而开发者作为使用的最终受益方，尽管已获得许可，仍可能面临原始作者提出的侵权主张。据报道，在视听领域，模板许可协议通过合同来应对这一风险：许可方声明并保证其持有训练使用所需的所有权利。对于记者和编辑在 2021 年 6 月 1 日《中华人民共和国著作权法》相关修正案生效之前撰写的报刊文章，这些文章的版权通常归属于记者或编辑，而非新闻机构。这意味着，即使获得了新闻机构的许可，也不能使文章用于人工智能训练获得批准。此类限制本身可能会阻碍自愿许可市场的形成，因为若中介方自身的权利不明确，其授予的许可会在权利链中增加一层潜在的侵权风险。这类限制还存在收益分配层面的问题：当许可由出版商、广播机构和存储库作为中介授予时，个人作者或表演者与这些机构之间的议价能力不对等，意味着许可带来的收益可能无法流向作品被使用的个人，这一考量与 B 节中讨论的各种许可模式的选择相关。

然而，中国的实践表明，即使存在上述结构性制约，在各行业及垂直领域，大规模自愿许可仍在发生。香港的广播和电视台与人工智能开发者合作，提供高质量的粤语语音数据。数字小说通过开发者与拥有大量签约作家的在线出版商之间的合作提供。图片通过开发者与大型图片库之间的合作提供，这些图片库基于从摄影师处获得的转让或专有许可而持有这些图片。科学论文通过开发者与学术数据库运营商之间的合作提供，在某些情况下，这些运营商会成为由此产生的领域专用模型的共同开发者。一些语料库提供商还会自行制作训练材料，通过与摄影师、模特和学校教师签约，由后者转让或许可产出的作品，供训练数据集使用。

其他司法管辖区也呈现出类似的模式，不过制度安排有所不同。在巴西，自愿许可的法律制度遵循《版权法》的一般规定。第 29 条关于事先明确授权的要求，以及第 4、31 和 49 条对版权交易的限制性解释意味着，即使在生成式人工智能出现之前经济权利已被完全转让，将作品用于人工智能训练仍将存在争议，因为合同不能具体预见到一种当时尚不存在的技术。在法国、意大利及其他实行作者权的司法管辖区，也存在类似的严格解释原则。[55]

若法律对舞台演出、表演或其他明确界定的客体明确规定了作者权利，自愿许可的局面可能会更清晰，因为基础权利的归属更为明确。对于由众多贡献者共同创作的作品类别，例如视听作品和录音制品，情况可能会更为复杂：法律上的不确定性或制作方与各个贡献者之间的合同的不确定性，可能会阻碍自愿许可，或者导致相关收益无法流向这些贡献者。塞内加尔的法律便是一例，其明确承认舞台演出为受保护作品，[56]但这一承认的实际应用尚未在人工智能训练领域得到检验。在肯尼亚，《版权法》（第 130 章）承认了一套标准的专有权，并规定通过集体管理组织对版权进行集体管理，但该制度架构并非针对计算用途而设计的，且针对人工智能训练的自愿许可

[55] 关于 2021 年前中国著作权法中有关记者、编辑创作的报刊文章的权属规则，见《中华人民共和国著作权法》（2020 年修订，2021 年 6 月 1 日起生效）第十八条（职务作品）。

[56] 《塞内加尔版权法》（2008 年 1 月 25 日第 2008-09 号法）第 2 条。《非洲知识产权组织班吉协定》（2015 年修订版）附件 VII 中也有类似的规定。

目前尚无有文献记载的做法。在加拿大，自愿许可是默认机制，而合同谈判则是核心活动。当前主要的诉讼案件是多伦多星报有限公司诉 OpenAI 公司案，该案是在未经授权使用存在争议的背景下产生的，而非涉及许可本身，本研究第三部分将对此进行讨论。

无论是个人自愿许可还是集体自愿许可，一个常见特征是，根据寻求数据所用系统的数据使用特征定制许可条款，而这一特征在目前使用的各类生成式人工智能系统之间差异显著。在数据仅使用一次的情形下——用于模型的初始训练或微调，而模型权重随后即固定——一种常见的模式是，针对既定语料库收取一次性费用或按收益分成。在系统生成输出需要持续访问被许可材料的情况下（例如 RAG 以及对素材库进行持续文本和数据分析的系统），这种访问并非一次性操作，而是一种持续性关系。实践中逐渐形成的补偿模式反映了这一点，包括持续付费以及按爬取或推理收费。除价款外，双方需要商定的主要事项还包括：许可的使用范围、许可期限、许可地域范围、许可终止时的处理方式，以及模型是否可以转让给第三方。人工智能背景下产生了一个问题：终止许可是否意味着开发者必须停止使用已基于许可材料完成训练的模型、对其进行重新训练，或是销毁存储的副本；上述任何一种处理方式均无法完全套用为此前的许可关系制定的救济措施。许可终止凸显了这一困难：在传统许可中，终止许可会使权利人重新获得对未来使用的控制权，但实际上，终止人工智能训练许可的权利人无法强制开发者停止使用通过训练已固定在模型中的内容，特别是在相关目录相对于语料库而言规模较小时；救济措施即便可行，也只能通过重新训练或尝试机器遗忘来实现，而机器遗忘的局限性在第一部分已有说明。即使此类干预措施被证明可行，也会产生运营和处理成本，双方可能需要在协议中明确这些成本的分摊方式，这进一步增加了此类许可的谈判难度。当此类模型权重被发布或出售给下游用户时，便会引发可转让性问题：许可所附带的合同义务是否随模型权重一并转移？一个相关但不同的问题是，下游的报酬是否可以根据模型本身计算，例如参考许可材料对权重的贡献计算。在已公开的合同中，这些问题尚未得到解决，且与下文 B.6 节中讨论的报酬模式之间的可能选择有关。

在高价值领域，音乐和视听行业的自愿许可主要采取个性化合作形式，将许可与远超出受保护作品使用范围的商业条款绑定。2025 年间，各大唱片公司针对主要人工智能音乐工具开发商，从诉讼转向许可：其中两家开发商就针对其提出的侵权指控达成和解，并同意构建新的、获许可的平台；其中一项和解协议附带的安排是，开发商从唱片公司收购一项音乐会信息服务；另有一家流媒体服务平台则分别与三大唱片公司及另外两家权利人达成协议，以选择加入为基础开发人工智能音乐产品。[57]在达成和解之前，该领域已发表了若干集体原则声明：行业机构、艺术家组织和科技公司组成的联盟阐明了相关原则，即在人工智能开发中使用受保护的作品库必须获得授权和许可，并呼吁在训练实践中保持透明度。[58]在视听领域，2025 年 12 月公布了一项许可安排，根据该安排，一家制片公司将向一个视频生成平台提供其两百余个角色，并取得该平台开发商 10 亿

[57] 2025 年底，环球音乐集团和华纳音乐集团分别与 Udio 达成了诉讼和解，并就新平台达成了许可协议；此外，华纳音乐集团还与 Suno 达成了和解，该协议包括 Suno 收购音乐会信息服务平台 Songkick。另一家开发商 KLAY Vision 与各大唱片公司及其旗下的出版公司达成了协议。2025 年 10 月，Spotify 与索尼音乐集团、环球音乐集团、华纳音乐集团、Merlin 和 Believe 共同宣布开展流媒体合作。

[58] “人类艺术运动”由包括主要音乐行业机构和艺术家代表团体在内的 150 多个组织于 2023 年组成，该组织阐述了七项核心原则 <https://www.humanartistrycampaign.com>，于 2026 年 8 月 25 日访问；罗兰公司和环球音乐集团提出的《人工智能音乐创作原则》已获得 50 多个组织的认可 <https://aiformusic.info>，于 2026 年 8 月 25 日访问。这两项倡议均被讨论，并置于更广泛的负责任设计框架之下，见 de Berardinis 等人，“RAIM: Operationalising Trustworthy AI in Music Generation”，本研究第一部分脚注 17。

美元的股权，据报道，因开发商关停该平台，该安排在 2026 年 3 月终止。[59]上述安排部分涉及输出的地位问题，该问题不在本研究讨论范围之内；在此提及仅用于说明当前出现的定制化、高价值许可结构及其脆弱性。

(b)主要的集体许可模式（自愿许可、延伸性许可和强制许可）

最常见的集体许可形式是自愿集体管理。集体管理组织通过合同获得其成员授予的某些权利的管理权；它在任务授权范围内进行谈判并签订许可协议；所得收益分配给其成员。多数具备成熟版权制度的司法管辖区都设有自愿性集体管理组织，但这些组织所管理的权利和作品存在显著差异。音乐公开表演协会是最具普遍性的例子。在中国，有五家集体管理组织已投入运营或即将全面投入运营：中国音乐著作权协会、中国音像著作权集体管理协会、中国文字著作权协会、中国摄影著作权协会和中国电影著作权协会。[60]这些集体管理组织已公开表示愿意为人工智能训练授予许可，但截至本研究撰写之日，中国尚未就人工智能训练达成任何重大的集体许可协议。[61]在中国的具体情况下，其结构性原因在于，集体管理组织的任务授权仅限于其成员的权利，而对于原始作者并非集体管理组织成员的作品，开发者仍面临剩余风险。

在大韩民国，也可以观察到类似的特点，例如韩国音乐著作权协会及类似机构。集体作品库的有限性似乎只是未能达成许可的部分原因。在现行监管制度下，开发者主动申请许可的动力可能不足。还有两个因素似乎能够解释为什么在美国仍能达成交易。首先，市场结构使得少数几家主要权利人和内容提供商能够与开发者直接谈判。其次，诉讼风险，包括法定赔偿，促使达成协议。B.2 节所述的与和解相关的许可协议便是此类情况的例子。因此，要解释为何许可在中国和大韩民国尚未形成势头，必须在上述市场和法律因素与集体任务授权的局限性之间取得平衡。截至本研究撰写之日，大韩民国文化体育观光部以及韩国著作权委员会已着手推进相关工作，旨在完善监管框架。

在制度层面，美国拥有密集的行业性集体管理机构网络。表演权组织（美国作曲家、作家和出版商协会（ASCAP）、广播音乐公司（BMI）、欧洲舞台作者和作曲者协会（SESAC）、全球音乐版权组织和AllTrack）管理音乐作品的公开表演权；根据2018年《音乐现代化法》成立的机械复制许可集体组织，管理音乐作品数字化使用的一揽子机械许可；[62]根据《版权法》第 114 条，SoundExchange 管理录音制品数字演出的法定使用费；美国版权结算中心（CCC）管理已出版文本材料的复制权；而艺术家权利协会（ARS）管理视觉艺术领域的某些复制权。各组织均在其所属领域内运作，并按照与其成员群体协商确定的条款运作。鉴于人工智能训练已成为受保护作品的一种使用方式，CCC 已将其年度企业许可范围扩展至涵盖某些人工智能用途，并宣布计划再推出若干专门针对人工智能的许可。[63]

过去两年间，除中国和美国外，还有集体管理组织和其他中介机构采取了若干项具体措施，将人工智能训练纳入其许可任务授权范围，其中部分措施直到2025年底才公布。在瑞典，瑞典表

[59] 华特迪士尼公司与 OpenAI 于 2025 年 12 月 11 日宣布的安排涉及 200 多个角色，并计划进行 10 亿美元的股权投资；据称，在 OpenAI 于 2026 年 3 月 24 日撤下其视频生成产品 Sora 之后，该安排已被放弃，且未完成任何投资。

[60] 前三个通常简称为 MCSC、CAVCA 和 CWWCS；中国摄影著作权协会和中国电影著作权协会尚未全面开展业务。

[61] 中国主要的集体管理组织已公开表示愿意为人工智能训练发放许可，但迄今为止尚未达成任何重大的许可协议。

[62] 《奥林·G·哈奇—鲍勃·古德拉特音乐现代化法》，第 115-264 号公法，《美国法律汇编》第 132 卷第 3676 页（2018 年），编入《美国法典》第 17 编第 115 条及后续条款。机械复制权许可集体管理组织依据《美国法典》第 17 编第 115 条 d 款第 3 项指定成立。

[63] 见 www.copyright.com/ai。

演权协会（STIM）于2025年9月推出了一项集体许可安排，涉及人工智能音乐公司Songfox以及溯源服务提供商 Sureel——后者的技术能够从人工智能输出追溯对其产生影响的作品。[64]在挪威，Kopinor于2025年12月达成一项协议，根据该协议，挪威国家图书馆可使用报纸内容来训练挪威语语言模型，相关费用由挪威政府资助。[65]在丹麦，同样在2025年12月，丹麦新闻出版集体管理组织与ProRata AI达成了一项“选择加入”许可合作，允许后者在检索式问答引擎中使用其成员的内容；与此同时，另一家丹麦组织Tekst&Node正在根据丹麦延伸性集体许可框架制定一项许可模式。[66]在荷兰，自2025年7月起，荷兰私营新闻出版商协会和ANP通讯社与公共组织合作，开发了一款本土语言模型 GPT-NL。[67]在西班牙，西班牙复制权中心（CEDRO）已加入“真正简单许可”（Really Simple Licensing）倡议，旨在为西班牙市场制定一种人工智能许可模式。[68]在英国，版权许可代理公司（CLA）联合出版商许可服务机构（PLS）及作者许可与集体管理组织（ALCS），一直在共同制定一项专门针对生成式人工智能的集体许可，涵盖训练、微调和 RAG。[69]在其他地区，复制权组织已将“人工智能再利用权”纳入其现有的数字版权许可中，或正在制定类似的解决方案：在日本，日本学术著作权许可协会于2025年4月在其“数字版权许可”中增加

---

[64] STIM，“STIM launches world’s first collective AI music licence”（2025年9月9日），CISAC报道，https://www.cisac.org/Newsroom/society-news/stim-launches-worlds-first-collective-ai-music-licence 。“选择加入”许可同时涵盖模型训练和人工智能输出的下游使用；补偿由预付款和开发者收入分成两部分组成，而归属技术提供商Sureel则对输出追溯至对其产生影响的作品。

[65] “Kopinor signs historic agreement on newspaper content for AI in Norway”（2025年12月19日），IFRRO报道，https://ifrro.org/page/article-detail/ifrro-member-kopinor-signs-historic-agreement-on-newspaper-content-for-ai-in-norway/。该协议由 Kopinor 代表挪威媒体企业协会签署，允许挪威国家图书馆访问出版超过一年的报纸内容，用于训练挪威语语言模型，并由挪威政府每年提供4,500万挪威克朗的资助。它与此前关于图书数字传播的“Bookshelf”协议有所不同。

[66] ProRata AI和丹麦新闻出版集体管理组织（DPCMO），“ProRata Partners with Danish Publishers Group DPCMO”（2025 年 12 月 10 日），https://www.prnewswire.com/news-releases/prorata-partners-with-danish-publishers-group-dpcmo-to-launch-the-first-decentralized-sovereign-ai-answer-engine-302637182.html 。DPCMO（该组织声称其代表丹麦新闻业 99%的从业者）签署了一项“选择加入”集体许可合作协议，允许成员的内容基于收入分成机制，用于ProRata的检索式问答引擎。另一家丹麦组织（Tekst&Node）正在根据丹麦延伸性集体许可框架开发一种人工智能许可模式，该模式允许权利人就文学作品的某些文本与数据挖掘以及人工智能应用授予许可，但须遵守“选择退出”条款并获得文化部的批准。见Terese Foged，“[DK] First extended collective licence approval for text and data mining and the use of works by AI” （欧洲视听观察站 IRIS 法律观察，2026 年）<https://merlin.obs.coe.int/article/10505>，于2026年7月3日访问。

[67] NDP Nieuwsmedia和TNO，“Large dataset news organizations for Dutch AI language model GPT-NL”（2025年7 月 17 日），https://www.ndpnieuwsmedia.nl/2025/07/17/large-dataset-news-organizations-for-dutch-ai-language-model-gpt-nl/。荷兰私营新闻出版商协会NDP Nieuwsmedia的成员单位，连同ANP通讯社，正将其部分档案资料提供给GPT-NL用于训练。该模型是一款由TNO、NFI 和 SURF 共同开发的非营利性荷兰语语言模型；提供内容的出版商将在该模型发布时获得报酬。

[68] CEDRO，关于与 RSL Collective 达成合作协议的新闻稿（2026 年 1 月），https://www.cedro.org/sala-de-prensa/noticias。西班牙复制权组织CEDRO已加入RSL Collective，成为其在西班牙的合作伙伴，并参与RSL技术委员会的工作，根据西班牙市场调整机器可读许可标准，以便其成员能够就其作品在人工智能训练中的使用设定相关条款。CEDRO于2025年12月16日另行发布了关于人工智能训练中权利转让的建议。

[69] 版权许可机构，“CLA announces development of Generative AI Training Licence”（2025 年），https://cla.co.uk/development-of-cla-generative-ai-licence/；出版商许可服务，“PLS and ALCS agree to development of pioneering CLA generative AI licence”（2025年4月），https://www.pls.org.uk/news-events-policy/news/pls-and-alcs-agree-to-development-of-pioneering-cla-generative-ai-licence/。CLA 正与 PLS 和 ALCS 共同开发一项针对生成式人工智能中文本使用的选择加入式集体许可，涵盖训练、微调以及检索增强生成；权利人通过登记其作品参与。

了此类权利；[70]澳大利亚版权代理公司一直在制定类似的解决方案[71]；在德国，VG Wort 已将其“数字版权许可”扩展至人工智能再利用，而集体管理组织 GEMA 则另行提出了一项专门的模式，但截至本研究撰写之日，该模式尚未达成最终协议。[72]这些进展仍不均衡，且在某些情况下还处于实验阶段，但这表明集体许可途径正在被积极检验，而不仅仅是停留在讨论层面。

延伸性集体许可（ECL）是一种许可模式，起源于 1960 年代的北欧司法管辖区，目前已被若干其他国家采用。它通过将集体管理组织的许可效力扩展至特定作品库内的非成员权利人，从而解决了剩余风险问题，并提供了退出权。在欧洲关于人工智能训练的政策辩论中，该模式曾被提及，作为获得大型文本和视觉语料库权利许可的一种手段——开发者需要使用这些语料库，但逐一获取许可并不现实。2025 年欧洲议会关于生成式人工智能与版权的研究报告中明确提到了这一点。[73]在美国，版权局在其《版权与人工智能报告》第三部分中，将延伸性集体许可认定为一种可能用于人工智能训练的、有针对性的折中方案，与强制许可不同。[74]在关于报酬的学术文献中，这一提议也有所体现，只是存在一些差异：森夫特莱本（Senftleben）教授在《生成式人工智能与作者报酬》一文中提出了一种制度，即集体管理组织将收取使用生成式人工智能服务（其输出替代可识别作者的作品或与之形成竞争）所产生的报酬，并将所得收益分配给其成员。[75]

除了已经提到的例子，还有其他两个例子说明了现有的自愿集体管理机制如何在现有制度框架内调整集体管理机制以适应人工智能训练，而非等待新的立法，即印证了这种可能性。在巴西，音乐的公开表演制度由一个机构——ECAD 集中负责，该机构拥有法律和司法支持，负责针对作者权和相关权，集中收取和分配音乐作品和录音制品的公开表演使用费。[76]在现行《版权法》框架内，且在立法未变更的情况下，对于涉及将作品和录音制品用于训练人工智能模型的一项特定权利，在法律上有可能按照类似目前音乐公开表演权所采用的模式，以集体方式行使。这将要求权利人加入集体管理组织，且该组织须在其章程中规定代表权利人行使其他版权，而不仅仅是音乐的公开表演权；此外，还需获得文化部（负责监督集体管理组织的联邦政府机构）的事先授权。巴西《版权法》第 97 条允许集体管理组织将其许可范围扩展至训练及该模式实施所涉及的权

---

[70] 日本学术著作权许可协会（JAC）和 RightsDirect Japan，“JAC and RightsDirect Japan Announce the Availability of AI Re-Use Rights for Digital Copyright License”（2025 年 4 月 9 日），https://www.businesswire.com/news/home/20250409032666/en/。JAC 在其《数字著作权许可》中增加了人工智能再使用权，这是一种面向总部位于日本的企业提供的全公司范围内部使用许可，作品库涵盖世界各地出版商和学术协会的作品。

[71] 澳大利亚版权局，“AI and Copyright in Australia”及随附问答，https://www.copyright.com.au/membership/ai-and-copyright-in-australia/。澳大利亚版权局是澳大利亚政府版权与人工智能参考小组”的成员，其表示正在为澳大利亚境内受保护作品用于生成式人工智能场景制定许可方案。

[72] 关于 VG Wort，通过 RightsDirect（CCC 的子公司）管理的数字版权许可范围已扩展至内部人工智能再利用；见 https://www.rightsdirect.com/license/vg-wort-digital-license-2/。集体管理组织 GEMA 于 2024 年提出了一项针对生成式人工智能的许可模式；截至本研究撰写之日，尚未达成协议，与此同时，GEMA 已对人工智能开发者提起侵权诉讼（见上文 *GEMA 诉 OpenAI* 案）。见 GEMA，“AI and music: GEMA calls for music creators to share in AI providers’ revenues and presents first licensing model”（2024 年 9 月 26 日）<https://www.cisac.org/Newsroom/society-news/ai-and-music-gema-calls-music-creators-share-ai-providers-revenues-and> 于 2026 年 7 月 3 日访问；GEMA，“Suno AI and Open AI: GEMA sues for fair compensation”（2025 年 1 月 20 日）<https://www.gema.de/en/news/ai-and-music/ai-lawsuit>，于 2026 年 7 月 3 日访问。

[73] 欧洲议会内部政策总司，Study on “Generative AI and Copyright: Training, Creation, Regulation”，IUST_STU(2025)774095_EN (2025)，建议将法定许可作为欧盟人工智能训练的最佳框架。

[74] 美国版权局，*Copyright and Artificial Intelligence, Part 3: Generative AI Training*（2025 年）。

[75] M.R.F.Senftleben, “Generative AI and Author Remuneration”, 54 IIC 1535 (2023).

[76] 关于巴西集体管理制度的结构以及公共表演使用费征收向 ECAD 集中化的情况，见《巴西版权法》（第 9,610/1998 号法）第 97 至 100 条，以及第 9,574/2018 号法令。联邦最高法院在 ADI 2.054 (2003) 号案件以及关于 ADI 5062 (2016) 和 ADI 5065 (2016)号案件的联合裁决中，确认了该模式符合宪法。

利许可 （即复制权、向公众传播权或提供权，以及根据文化部的立场，还包括输出阶段的改编权），前提是已获得必要的授权。然而，统一收费的要求仅适用于音乐公开表演权，因此，即使 ECAD 或其附属机构内部集中代表人工智能训练权，也不会妨碍其他具有相同目的的集体管理组织成立。

在肯尼亚，《2020 年版权（集体管理）条例》对集体管理组织的注册、运营和监管作了规定。[77]该框架具有通用性，并未专门针对计算用途进行调整，但为人工智能训练的自愿集体许可制度的建立提供了制度基础。肯尼亚版权委员会于 2026 年 3 月发布《2026 年版权及相关权法案（草案）》以征求公众意见，意见征集期于当月月底截止。该草案包含一条条款（第 34 条），允许为非商业性研究进行文本和数据分析，但该法案并未对“非商业性”给出明确的法定定义，未设立权利人的“选择退出”机制，也未就人工智能训练中受版权保护作品的使用，以及人工智能生成作品的可版权性作出具体规定。肯尼亚的情况说明了那些正在分阶段调整版权框架的司法管辖区中的一种更普遍的模式：现有的集体管理组织架构为人工智能训练的集体许可提供了一种可能的载体，但这种调整需要采取积极的立法或合同措施，而这些措施目前尚未实施。

3.原始语言数据的社区许可与管理许可

上述许可机制适用于在开发者接触到时已经确定并发布的训练数据。另一类训练数据的生成方式则有所不同。由于此类文本很少，因此无法利用网络规模的文本语料库来构建大多数非洲语言的语音和语言模型。相反，这些数据是以原始数据的形式收集的：通过实地调研、社区录音活动以及与广播机构的合作，从讲话者处收集的音频录音和文字转录。版权产生于作品被“固定”之时，而非“爬取”之时。录音作为录音制品受保护，文字转录作为文学作品受保护，且讲话者可能享有表演者权。[78]因此，在数据收集之时，即在涉及文本与数据挖掘的任何问题出现之前，授权和补偿问题便已产生。

这类数据还涉及另一层知识产权。口头传统、民间故事和谚语的录音既是归属于其固定者的版权作品，也是社区根据特殊立法享有集体权利的文化表达形式。在肯尼亚，2016 年第 33 号《传统知识与传统文化表现形式保护法》旨在落实 2010 年《肯尼亚宪法》第 11 和 40 条以及第 69 条第(1)款 c 项，该法承认社区是其文化表现形式的权利人，要求利用其时须获得事先知情同意并签订授权使用协议，并规定应向社区支付补偿；《宪法》第 11 条第 3 款 a 项规定，议会有义务确保社区因其文化及文化遗产被使用而获得使用费。[79]当此类语言数据被纳入训练语料库时，版权权利链与社区权利并行存在，且录音固定者授予的许可本身并不能解决社区的法定地位问题。

为此，一种独特的许可实践应运而生。标准开放许可，尤其是公有领域贡献，将此类数据纳入训练流程，却未向贡献社区提供治理权或经济回报。以社区为中心的许可在保持开放性的同时，对这一立场进行了调整。Nwulite Obodo 开放数据许可根据用户的地理位置和发展背景，采用分级的使用条件，并纳入了惠益分享义务；该许可已应用于非洲语音数据集，包括与马塞诺大学合作开发的DhoNam Dholuo语音数据集。[80]为Vuk’uzenzele isiXhosa语音数据集制定的Esethu

---

[77] 《2020 年版权（集体管理）条例》（肯尼亚）。许可制度载于《版权法》（第 130 章）第 33 和 34 条以及第七部分。

[78] 《版权法》（第 130 章），《肯尼亚法律》第 2 条（除其他外，界定了“录音制品”和“文学作品”）以及关于表演者权利的条款。关于国际层面的表演者权利，见 WPPT 第 10 和 14 条，如上文 B.1 节所述。

[79] 《传统知识和文化表现形式保护法》（2016 年第 33 号法）（肯尼亚）；《肯尼亚宪法》（2010 年），第 11 条、第 40 条第 5 款和第 69 条第 1 款 c 项。

[80] 《Nwulite Obodo 开放数据许可协议》第 1.0 版（比勒陀利亚大学数据科学法律实验室，2025 年）<https://datasciencelawlab.africa/nwulite-obodo-open-data-license/>，于 2026 年 7 月 31 日访问；C.I.Okorie

许可允许非洲实体免费用于研究和商业用途，但要求其他商业用户支付许可费，所得收益将重新投资于创建更多语言数据。[81]Te Hiku Media 针对毛利语数据制定的 Kaitiakitanga 许可体现了另一地区采取的相同做法。[82]这些工具构成了除个人许可和集体许可之外的第三种机制，通过该机制，人工智能训练中使用原始语言数据的授权和补偿事宜得以统筹安排；它们还说明了针对集体持有的文化资产制定的国家法定框架如何为合同实践提供支撑。在肯尼亚，这一层面还将与前一小节中讨论的《2026 年版权及相关权法案（草案）》中提出的文本和数据分析例外产生关联。

4.法定文本与数据挖掘例外与一般性条款

欧洲联盟版权法中包含与人工智能训练所用操作相关的明确文本与数据挖掘例外，尽管这些例外在此背景下的适用性和充分性仍存在争议。关于数字化单一市场版权及相关权的(EU)2019/790 号指令（《CDSM 指令》）第 3 和 4 条，为文本与数据挖掘确立了两项强制性例外。第 3 条是针对研究组织和文化遗产机构的研究用途例外，不包含“选择退出”机制。第 4 条是一项适用于任何受益人的普遍例外，但权利人可“以适当方式（例如，对于在线公开的内容，采用机器可读方式）”选择退出。[83]《人工智能法案》通过援引方式纳入了该框架，要求通用人工智能模型提供商制定政策，以遵守欧盟版权法，识别和尊重根据《CDSM 指令》第 4 条第 3 款所表达的权利保留，并发布关于训练所用内容的足够详细的摘要。[84]

两项德国裁决探讨了这些例外对人工智能训练的适用问题。在 Kneschke 诉 LAION e.V.案中，汉堡地区法院于 2024 年 9 月裁定，在构建 LAION-5B 数据集（该数据集用于训练图像生成模型）时使用某张照片的行为，属于德国实施《CDSM 指令》第 3 条关于科学研究中文本与数据挖掘的例外所涵盖的范围。 法院将 LAION 视为该条款所指的科研机构，并认为涉案计算分析属于出于科学目的的文本与数据挖掘。法院还在附带意见中指出，第 4 条的一般文本与数据挖掘例外在任何情况下都不适用，因为在该法院看来，鉴于当代大语言模型的能力，来源平台使用条款中所载的权利保留是可机器读取的。[85]在 Kneschke 诉 LAION e.V.案的上诉审理中，汉堡高等法院维持了原判，但对第 4 条第 3 款规定的“机器可读性”要求进行了更详尽的阐述。[86]上诉法院认为，机器可读性不仅要求权利保留声明能够被机器捕获，还要求其能够被自动化程序解读，从而使得该保

和 M. Omino，“Addressing Inequitable Openness in Licences for Sharing African Data and Datasets Through the Nwulite Obodo Open Data Licence” （2025）7(3)Law, Technology and Humans 94。DhoNam Dholuo 语音数据集（马塞诺大学非洲人工智能中心）根据以下许可协议发布：<https://mozilladatacollective.com/datasets/cmjepxo6t08nmmk07iauvua6v>，于 2026 年 7 月 31 日访问。

[81] Lelapa AI, Way With Words, and Data Science for Social Impact (University of Pretoria), “The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages”, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (2025) <https://aclanthology.org/2025.acl-long.1487/>，于 2026 年 7 月 31 日访问。

[82] Te Hiku Media, Kaitiakitanga License <https://tehiku.nz/te-hiku-tech/te-hiku-dev-korero/25141/data-sovereignty-and-the-kaitiakitanga-license>，于 2026 年 7 月 31 日访问。

[83] 2019 年 4 月 17 日欧洲议会和理事会关于数字化单一市场中版权及相关权的第(EU) 2019/790 号指令，OJ L 130/92（《CDSM 指令》）， 第 3 条（针对研究机构和文化遗产机构为科学研究目的进行的文本与数据挖掘的强制性例外）和第 4 条（针对合法可获取作品的文本与数据挖掘的一般性选择退出例外）。

[84] 2024 年 6 月 13 日欧洲议会和理事会关于制定人工智能统一规则的第(EU) 2024/1689 号条例（《人工智能法案》），OJ L, 2024/1689, 12.7.2024，第 53 条第(1)款 c 项和 d 项，以及序言第 105 和 106 段。

[85] Kneschke 诉 LAION e.V.，汉堡地区法院，2024 年 9 月 27 日（310 O 227/23）（认定，为用于训练图像生成模型而构建的 LAION-5B 数据集属于德国为落实《CDSM 指令》第 3 条科学研究文本与数据挖掘例外所作的规定（《德国版权法》第 60d 条）的适用范围；还在附带意见中指出，第 4 条关于文本与数据挖掘的一般性例外并不适用，理由是源平台条款中包含的权利保留条款在相关意义上是机器可读的）。关于上诉，见下文讨论的 Kneschke 诉 LAION e.V.，汉堡高等法院，2025 年 12 月 10 日（5 U 104/24）。

[86] Kneschke 诉 LAION e.V.，汉堡高等法院，2025 年 12 月 10 日（5 U 104/24）。

留声明所涵盖的内容不被处理。简而言之，该保留声明必须同时具备“机器可读”与“机器可执行”两种属性。法院进一步指出，机器可读性并非一种静态属性，而是必须根据争议使用时普遍存在的技术能力进行评估；就本案事实而言，这意味着应以 2021 年下半年的现有技术为准。上诉法院的论证留下了一个未决问题：到 2025 年或之后，当能够从非结构化自然语言文本中提取语义的模型已广泛可用时，何谓“机器可读性”。[87]

大约一个月前，在 GEMA 诉 OpenAI 案中，慕尼黑第一地区法院针对一个相关但截然不同的问题，得出了不同的结论。[88]德国集体管理组织 GEMA 曾起诉 OpenAI，指控 ChatGPT“记忆”了九首知名德国歌曲的歌词，且可被诱导复制这些歌词。慕尼黑法院将歌词被摄入训练语料库以及随后在模型中被记忆的行为，均视为德国为实施《信息社会指令》所规定的复制行为； 该法院驳回了适用《德国版权法》第 44b 条中关于文本与数据挖掘的一般例外的申请，其理由是生成式模型会永久记忆歌词，而这超出了该例外旨在允许的对文本和基于模式的信息的分析范围；此外，法院将模型输出的结果视为证据，证明歌词以类似于数字图书馆的方式被固定在模型内。法院准予了禁令救济，责令提供信息，并判付损害赔偿金。[89]该判决目前正在上诉中。在一些人看来，该判决是欧洲首例在人工智能训练问题上支持权利人的重大判决；也有人批评该判决将技术上本是一系列不同的操作简化为单一的复制行为，并因将输出阶段的记忆证据与训练过程的技术现实混为一谈，从而误用了文本与数据挖掘例外。[90]

除判例法之外，学术界对于《CDSM 指令》第 4 条的例外究竟是否适用于人工智能训练，仍存在争议。文献中一种普遍观点认为，人工智能训练属于《CDSM 指令》第 3 和 4 条所涵盖的数据挖掘形式。这种观点往往援引《人工智能法案》中引用了这些条款这一事实。一些欧盟学者，包括蒂姆·多尔尼斯（Tim Dornis）教授，明确表达了相反的观点。他在发表于《欧洲知识产权评论》的一篇论文中，以及与计算机科学家施托贝尔（Stober）合著的一份更详尽的法律意见书中指出，仔细考察生成式人工智能训练的技术基础后发现，其执行的操作超出了文本与数据挖掘例外旨在涵盖的范围。其论点大致如下：文本与数据挖掘例外允许从作品中提取语义、事实和统计信息； 相比之下，生成式人工智能模型的训练还会提取并编码句法和表达特征，进而以此生成替代原作品的输出；由此对权利人合法权益造成的干预，超出了该例外所允许的范围。[91]

埃莱奥诺拉·罗萨蒂（Eleonora Rosati）教授也提出了类似的论点，但出发点有所不同。她认为，文本与数据挖掘虽可能是人工智能训练过程的一部分，但既不等同于人工智能训练，也不

---

[87] 汉堡上诉法院指出，机器可读性要求该保留条款“能够被机器以某种方式解读……使得在自动化处理过程中，[权利]保留所涵盖的内容不会被处理”（译文）。另见 P. Keller, “LAION Round 2: Machine-Readable but Still Not Actionable - The Lack of Progress on TDM Opt-Outs”, Kluwer Copyright Blog（2025 年 12 月 17 日和 18 日），第 1 部分。

[88] GEMA 诉 OpenAI，慕尼黑第一地区法院，2025 年 11 月 11 日（42 O 14139/24），终局判决。原告是德国集体管理组织 GEMA；被告是 OpenAI OpCo 有限责任公司和 OpenAI Ireland 有限公司。法院下令采取禁令救济措施、提供信息并赔偿损失。

[89] 同上。法院将歌词纳入训练语料库以及随后在模型中对其进行记忆的行为，均视为《德国版权法》第 16 条意义上的复制行为，该条系落实第 2001/29/EC 号指令第 2 条的规定。

[90] 关于这一点的批判性评论，见 G. Frosio, “Copyright in Formaldehyde: How GEMA v OpenAI Freezes Doctrine and Chills AI”, Kluwer Copyright Blog（2025 年 12 月 10 日和 11 日）， 第 1 和第 2 部分。文中指出慕尼黑法院将输出阶段的记忆证据与训练的技术现实混为一谈，并认为更合理的观点是将合法的中间技术复制件视为属于《信息社会法》第 5 条第 1 款或文本与数据挖掘例外范畴。

[91] T.W.Dornis, “The Training of Generative AI Is Not Text and Data Mining”, 47 European Intellectual Property Review 65 (2025).另见 T.W.Dornis and S. Stober, *Urheberrecht und Training generativer KI-Modelle - technologische und juristische Grundlagen*, Initiative Urheberrecht （Nomos，2024 年 9 月）。

等同于人工智能训练的全部内容；鉴于人工智能训练可能涉及多种受版权限制的行为，将欧洲的讨论仅限于《CDSM 指令》第 4 条的范围过于狭隘；而且，无论如何，《CDSM 指令》第 7 条第 2 款和《信息社会指令》第 5 条第 5 款中的“三步检验法”可能发挥作用，将人工智能训练排除在例外之外，理由是该使用与作品的正常利用相冲突，并不合理地损害了权利人的合法权益。如果这一观点成立，那么许可问题仍是分析中不可或缺的一部分。[92]马蒂亚斯·莱斯特纳（Matthias Leistner）教授曾另行指出，《CDSM 指令》第 4 条第 3 款中的选择退出机制很可能成为欧盟版权法中的核心争议点，且目前在实践中难以实施；这一观点已在欧洲的政策辩论中得到讨论。[93]

森夫特莱本（Senftleben）教授则认为，第 4 条不仅适用于人工智能训练的情境，而且在确定人工智能背景下何谓“作品的正常利用”时，也可能具有直接关联；这一点对于“三步检验法”的适用具有重要意义。[94]

欧盟委员会在对《人工智能法案》的解读中已将《CDSM 指令》第 4 条视为适用于人工智能训练——其做法侧重于落实选择退出机制，而非将人工智能训练排除在例外的范围之外。本研究第三部分在透明度和权利执行的背景下，探讨了选择退出机制。[95]

韩国立法机关已审议文本与数据挖掘例外，但尚未将其纳入立法；相关法案自 2021 年以来一直提交至国民议会审议。[96]在缺乏法定条款的情况下，人工智能训练问题应依据《版权法》第 35-5 条中的一般合理使用条款进行评估，该条款参照了《美国版权法》第 107 条，并在大韩民国与美国签署自由贸易协定后被引入。国内争论与本节所概述的考虑因素高度一致。支持制定法定例外的人士指出，明确的法定表述将带来确定性；反对制定法定例外的人则回应称，法定表述本身会产生解释上的歧义——正如上文讨论的德国关于机器可读性的诉讼所示；韩国大法院 2024 年对第 35-5 条适用标准的阐释，为该一般条款赋予了可操作的内容，而且，宽泛的豁免可能使外国开发者获得不成比例的利益，同时将价值从韩语内容的权利人手中转移出去。[97]

在涉及训练问题时，该一般性条款已通过行政指导被赋予实施内容。2026 年 2 月，文化体育观光部与韩国著作权委员会联合发布了《生成式人工智能模型训练适用合理使用原则指南》（本研究第一部分中提及），该指南将第 35-5 条规定的四项因素应用于本研究第一部分所述的训练流程各阶段。该指南说明，获取训练材料的方式具有相关性，因此未经授权的抓取、规避技术保护措施、绕过付费墙以及无视通过“机器人”排除协议表达的保留意见，均不利于作出“合理使用”的认定，指南还就是否属于“合理使用”分别提供了示例。[98]

---

[92] E. Rosati, “Is text and data mining synonymous with AI training?”, 19 Journal of Intellectual Property Law and Practice 851 (2024); E. Rosati, “The German LAION decision: A problematic understanding of the scope of the TDM copyright exceptions and the transition from TDM to AI training”, IPKat（2024 年 10 月 7 日）； E. Rosati, Copyright Exceptions and Fair Use Defences for AI Training Done for ‘Research’ and ‘Learning’, or the Inescapable Licensing Horizon, SSRN Working Paper No. 5331405 (2025)。

[93] M. Leistner, “TDM und KI-Training in der Europäischen Union”, GRUR 2024, 1665（文中认为，“选择退出”机制很可能成为欧盟版权法中的核心挑战，且目前在实践中难以实施）。

[94] Senftleben, M. Text and Data Mining, Generative AI, and the Copyright Three-Step Test. IIC 57, 67, at 79 (2026)。https://doi.org/10.1007/s40319-026-01680-2

[95] 《人工智能法案》第 53 条第 1 款 c 项（有义务制定政策以遵守欧盟版权法，并识别和尊重——包括通过采用最先进的技术——依据《CDSM 指令》第 4 条第 3 款提出的权利保留）。

[96] 从 2021 年 1 月 15 日提交国民议会的《版权法》全面修正案开始，该修正案中包含文本与数据挖掘例外；该修正案随第二十一届国民议会届满而失效，后续提案仍在审议之中。

[97] 韩国最高法院第 2021Da272001 (2024)号裁决，对《版权法》第 35-5 条所定标准的权威阐释。

[98] 见上文注 10 所引《韩国合理使用指南》。

5.报酬制度

(a)报酬制度的类型

在允许将受保护作品用于人工智能训练的情况下——无论是基于例外、许可，还是两者兼而有之——可能会涉及报酬问题。目前若干模式正在实施，这些模式借鉴了版权法在适应新的利用模式方面更广泛的经验。

第一种是许可费制度，可通过单独或集体协商确定，由开发者在对语料库进行汇编或训练时支付。该模式的优势在于，权利人只需将作品用于训练，即可获得报酬，而无需满足其他条件；就个人许可而言，其缺点在于，该模式遗漏了那些直接谈判不可行的个人权利人所构成的长尾群体。

第二种是基于征税的制度，即向人工智能开发者或生成式人工智能服务的特定类别用户征收费用，并将所得款项通过集体管理组织或类似机构分配给权利人。上文 B.3 节提到的欧洲议会研究报告建议，在欧盟对人工智能训练实行一种法定许可制度，该制度将兼具基于征税的方案与强制性集体管理的特点。欧盟以外的司法管辖区也提出了类似的提案，这些提案通常是参照许多国家现有的私人复制税制定的。关于此类方案与《伯尔尼公约》第九条第 2 款、WCT 第 10 条以及《TRIPS 协定》第 13 条中的三步检验法规定是否兼容；以及在涉及复制权以外的其他权利时，是否符合《伯尔尼公约》第十一条之二第 2 款和第十三条第 1 款的具体条件，均产生了疑问。

第三种模式，即最新且在概念上最为独特的模式，是输出层面的报酬。根据这种方法，报酬的重点不在于训练用途本身，而在于所部署的模型所生成的输出——这些输出替代可识别作者的作品或与之形成竞争。该模式将报酬与权利人作品在相关市场上可能被替代的情况直接挂钩。然而，该模式的实施似乎需要一些技术机制来识别替代性输出，并将其追溯到训练数据中的可识别作品。

此外还出现了两个问题。当已识别作品的数量非常庞大时，每部作品的报酬可能微乎其微，甚至少到仅为名义上的程度。此外，计算依据也需要予以澄清：在大多提案中，报酬的确定将不取决于证明某一特定输出复制或改编了特定作品，而是取决于开发者报告作品在系统中的使用方式，这一做法以本部分 C 节所讨论的透明度机制为前提。[99]此外，无论作者的作品是否被用于训练，作者都可能因替代而受到影响。如上所述，利用目前的过滤技术，输出在实践中不太可能侵犯任何受版权保护的特定作品，因此所考虑的损害属于系统性损害，而非典型的版权侵权。在训练阶段进行的复制，或存在于模型中的复制，并不必然意味着输出构成侵权：对任何特定输出的侵权分析取决于该输出与受保护作品的相似程度。

替代担忧的规模在音乐领域日益有据可查：2026 年 4 月，Deezer 报告称，完全由人工智能生成的曲目约占其每日上传量的 44%，即每天约 75,000 首；而 Spotify 则报告称，在过去 12 个月内已移除了超过 7,500 万首垃圾曲目；[100]一项行业研究预测，到 2028 年，音乐创作者的收入中很大

[99] 在音乐领域，训练材料所附作品层面和权利人层面的元数据完整性，已被确定为归属认定以及分享生成输出内容所得收入的设计前提：de Berardinis 等，“RAIM: Operationalising Trustworthy AI in Music Generation”，本研究第一部分脚注 17 所引。

[100] Deezer, “AI-generated tracks now represent 44% of all new uploaded music”（2026 年 4 月 19 日）<https://newsroom-deezer.com/2026/04/ai-generated-tracks-represent-44-of-new-uploaded-music/>于 2026 年 7 月 3 日访问； Spotify，“Spotify Strengthens AI Protections for Artists, Songwriters, and Producers”（2025

一部分可能会被人工智能生成的内容所取代。[101]对此，森夫特莱本教授认为，替代测试可作为一种概念工具，用于区分以下两种情况：一种是合法使用人工智能支持人类的创造力（此类使用无需支付报酬）；另一种是利用人工智能替代人类创造力（此类使用则需支付报酬）。可以说，类似的做法构成了巴西第 2338/2023 号法案参议院版本的基础，该版本将选择退出机制与报酬权相结合，权利人可单独或集体行使该权利。[102]

(b)报酬方式和依据

在巴西，在某些具体领域，合成创作对人类作品的替代效应已然可以观察到，其中包括环境音效设计——该领域的用户试图播放完全由人工智能系统生成的音乐，以此规避公开表演费。围绕该问题针对 ECAD 的未决案件，将适时提供司法指引，说明在商业场所公开播放人工智能生成的音乐是否会触发任何报酬义务。该实例说明了人工智能训练报酬与人工智能输出相关报酬之间的相互作用。

在当前的讨论中，关于可能的报酬制度出现了两个问题。第一个问题是，报酬应与训练挂钩，还是与替代效应挂钩，抑或两者兼而有之。欧洲议会的研究报告以及若干学术提案试图将这两者结合起来；而巴西第 2338/2023 号法案以其当前形式来看，更应被解读为将补偿与为训练进行的摄取挂钩：第 65 条规定了在数据挖掘以及人工智能系统的训练和开发中使用受保护内容所应支付的补偿，因此该法案确立了一种与训练行为挂钩的补偿机制，该机制不延伸至对输出的利用。[103]第二个问题是，报酬制度是否以已获许可或已被放弃的认可的专有权为前提，或者是否应被视为一种专门的公平报酬请求权，这或许可以参照许多司法管辖区现有的私人复制报酬。[104]然而，如上所述，即使第 2338/2023 号法案未获通过，巴西现行立法仍可通过复制权、改编权或向公众传播权，对输出阶段进行规制；在此基础上，需要获得权利人的事先明确授权。[105]私人复制是一种补偿形式，而非许可，因为出于各种原因（包括用户的隐私），私人复制行为本质上被视为不可获得许可。

在巴西，还有一项意见涉及报酬设计：合成创作对人类作品的替代效应已然在某些具体领域显现，其中包括环境音效设计领域，该领域的用户试图播放完全由人工智能系统生成的音乐，以此规避公开表演费。围绕该问题针对 ECAD 的未决案件，将适时提供司法指引，说明在商业场所公开播放人工智能生成的音乐是否会触发任何对 ECAD 会员的付酬义务。该实例说明了人工智能训练报酬与人工智能产出相关报酬之间的相互作用；一个领域的选择可能会影响另一个领域的设计。[106]

年 9 月 25 日）<https://newsroom.spotify.com/2025-09-25/spotify-strengthens-ai-protections/>于 2026 年 7 月 3 日访问。

[101] 国际作者和作曲家协会联合会（CISAC）委托 PMP Strategy 开展的研究（2024 年 12 月）预测，到 2028 年，约 24%的音乐创作者收入可能因生成式人工智能而面临风险，其中数字收入方面的比例更高。

[102] 2024 年 12 月 10 日，巴西联邦参议院批准并移交众议院的第 2338/2023 号参议院法案。截至本研究撰写之日，该法案正由众议院特别委员会审议；最终可能提交总统批准的版本尚未定稿。

[103] 本研究采用的是经联邦参议院批准的第 2338/2023 号法案第 65 条的解读。该法案还被解读为将“摄取”依据与对输出替代的考量相结合，这主要是因为与补偿相关的标准涉及市场影响。按照文本所采用的解读，补偿与数据挖掘、训练和开发相关。

[104] 对可能的专门制度及其涵盖的使用范围的全面讨论超出了本研究的范围。

[105] 本研究采用的是经联邦参议院批准的第 2338/2023 号法案第 65 条的解读。该法案还被解读为将“摄取”依据与对输出替代的考量相结合，这主要是因为与补偿相关的标准涉及市场影响。按照文本所采用的解读，补偿与数据挖掘、训练和开发相关。

[106] 第 2338/2023 号法案第 62 条及后续条款（提交至众议院的版本）。商业秘密例外对透明度义务设定了限制，并在立法过程中引发了广泛争论。

最后，在监管框架尚在形成中的司法管辖区，报酬问题目前尚无定论。第一部分中提及的、由文化体育观光部与韩国著作权委员会联合发布的《2026 年关于合理使用原则在生成式人工智能模型训练中适用性的指南》，主要讨论哪些训练操作可在无需许可的情况下开展这一问题；而关于报酬的相应问题，则留待目前在正进行的更广泛的改革进程中予以解决。[107]然而，国内的大部分讨论都是基于这样一个前提：对人工智能模型的商业化训练应向创作者支付报酬，因此，目前尚未解决的问题是报酬机制。肯尼亚、巴西和中国的状况虽各不相同，但具有可比性：这三个国家都已启动立法程序，政策选项尚待确定，最终将形成何种模式尚不确定。本研究的第三部分涉及权利执行问题，将重新探讨这种法律状态的不确定性对试图维护自身权利主张的权利人所产生的实际影响。

第三部分 权利执行

第一部分所述的技术操作往往涉及国家版权立法和国际版权条约所承认的专有权；第二部分探讨的许可安排则提供了在无例外或限制适用时，用于获得这些权利许可的机制。第三部分探讨当上述两种情况均未发生时的情形：即权利人认为其作品未经授权且不属于任何适用例外的范围，被用于生成式人工智能模型的训练，并试图通过法院或行政程序维护自身权利的情况。

直到不久前，可供参考的判例法还非常少。如今情况已不再如此。截至 2026 年 8 月 22 日，美国已有 137 起针对人工智能开发者、涉及生成式人工智能系统使用受保护作品的诉讼，其他司法管辖区另有约 30 起此类诉讼记录。[108]巴西、加拿大、中国、德国、英国、美国及其他地区的法院已作出多项裁决，涉及分析的多个方面：训练后模型是否构成或是否包含其训练所用作品的复制件；欧洲法律中的例外规定是否适用于人工智能训练；国家法院是否对在国外进行的训练具有管辖权；替代训练语料库中作品的输出是否涉及权利人的复制权或提供权；以及输出本身是否受版权保护（该问题超出本研究范围，此处仅作为背景说明）。

本部分围绕三个主要问题展开。A 节从现有的判例法中总结经验教训，并按照涉及的问题而非司法管辖区进行梳理。B 节转向探讨针对人工智能开发者权利执行这一结构性问题，包括增加权利执行复杂性的跨境因素。C 节指出权利人试图行使权利时面临的主要实际挑战，包括透明度、选择退出技术标准的激增，以及证据收集在操作层面的困难。

A.现行判例法的启示

过去十八个月积累的判例法较为零散，在各司法管辖区不尽相同，且尚未在任何单一法律体系中形成法理定论。尚没有任何问题得到解决。尽管如此，它揭示了至少四个截然不同的问题，针对其中每一个问题，不同司法管辖区的法院已开始趋同，或在某些情况下出现分歧。下文将依次讨论这些问题。

1.“模型即复制件”问题

第一个问题是：从版权法的角度来看，经过训练的人工智能模型本身是否包含其训练所依据的作品的复制件。这与之前关于在语料库构建和训练过程中是否发生复制行为的问题有所不同。

---

[107] 韩国文化体育观光部与韩国著作权委员会，《生成式人工智能模型训练适用合理使用原则指南》（2026 年 2 月）；见本研究报告第一部分脚注 10。

[108] 根据爱德华·李（Edward Lee）教授在 chatgptiseatingtheworld.com 网站上维护的诉讼追踪记录，截至 2026 年 8 月 22 日，美国地图显示有 156 起诉讼，其他司法管辖区另记录有约 30 起诉讼。该数据为动态数据，以本研究撰写之日为准。另见版权联盟维护的诉讼地图。

本研究第一部分已对这一记忆问题的技术性质进行了讨论。过去十八个月内作出的三项裁决给出了截然不同的答案。

在上文第二部分 B.5 节中讨论的 GEMA 诉 OpenAI 案中，慕尼黑第一地区法院于 2025 年 11 月裁定，利用歌词对语言模型进行训练，以及模型随后对这些歌词的记忆，均构成德国实施《信息社会指令》相关规定所定义的复制行为，法院驳回了对《CDSM 指令》第 4 条例外规定的适用。[109]

关于涉及不同技术的相关“模型即复制件”问题，英格兰及威尔士高等法院就在几天前得出了不同的结论。在 Getty Images（美国）公司诉 Stability AI 有限公司一案中，乔安娜·史密斯女爵士法官于 2025 年 11 月 4 日裁定，Stable Diffusion（一种用于图像生成的潜在扩散模型）的权重并非该模型训练所用照片的侵权复制件。[110]原告在庭审结束时撤回了其关于主要侵权和数据库权利的主张，承认现有证据无法证明训练是在英国进行的；因此，争议转向了次要侵权，这需要法院确定该模型是否属于本身即为侵权复制件的物品。法院认定，庭审中的技术证据证明，扩散模型的权重并不存储训练数据，且在缺乏相关意义上的记忆证据的情况下，该模型不构成侵权复制件。[111]法院在附带意见中补充道，倘若法院得出相反结论，Stability AI 将根据“次要侵权”框架承担责任，因为该开发者知道或有理由相信，训练语料库中包含了未经同意从网络上抓取的作品。[112]

第三项裁决涉及一个相关但截然不同的案件事实模式，即杭州互联网法院对所谓“杭州奥特曼案”作出的判决。该案的被告是一个托管开源人工智能模型的平台，用户可以访问这些模型；部分用户曾利用受版权保护的日本漫画《奥特曼》对这些模型进行训练，并将训练后的模型上传供其他用户使用。版权人起诉了该平台，法院认定该平台构成帮助侵权，应承担责任。法院在附带意见中表示，根据中国法律，训练行为本身或许可被视为合理使用，其理由在于：在生成式人工智能模型的训练中使用受版权保护的作品，其目的是为了学习和分析语言特征、风格特征及其他要素，而非复制原有的创意表现形式。法院无需就该问题作出裁决，相关理由对判决结果亦不具有实质影响，但其附带意见认为，根据中国法律，训练可能符合合理使用情形。[113]

从纯技术角度来看，这三项裁决并不完全一致，因为它们涉及不同类型的模型（慕尼黑的 Transformer 语言模型、伦敦的潜在扩散模型、杭州托管用户微调图像生成模型的平台）以及不同的程序阶段。慕尼黑和伦敦的法院都深入分析了关于大型生成式模型记忆的计算机科学文献。这包括库珀、格里梅尔曼（Grimmelmann）、莱姆利（Lemley）等教授在从开放权重模型中提取记忆内容方面所开展的一系列研究。[114]这一系列研究被慕尼黑法院解读为证明了训练构成大规模复

---

[109] GEMA 诉 OpenAI，慕尼黑第一地区法院，2025 年 11 月 11 日（42 O 14139/24），终局判决。该判决也在第二部分 B.5 节进行了讨论。

[110] Getty Images（美国）公司诉 Stability AI 有限公司，[2025] EWHC 2863 (Ch)（史密斯法官，2025 年 11 月 4 日）。该判决同时涉及版权和商标主张；本讨论仅涉及版权方面的分析。2025 年 12 月，批准了就次要侵权裁决提起上诉（[2025] EWHC 3343 (Ch)）；截至本研究撰写之日，该上诉案件仍在审理中。

[111] 关于扩散模型架构，见第一部分 B.1 节及该节所引参考文献。

[112] 相关评论见 E. Derclaye，“Getty Images v Stability AI: A landmark judgment reinforcing the need for the UK government to amend its copyright laws”，Kluwer Copyright Blog（2025 年 11 月 20 日），网址：https://legalblogs.wolterskluwer.com/copyright-blog/getty-images-v-stability-ai-a-landmark-judgment-reinforcing-the-need-for-the-uk-government-to-amend-its-copyright-laws/。

[113] 杭州互联网法院，(2024)浙 0192 民初 1587 号（一审）； 浙江省杭州市中级人民法院，(2024)浙 01 民终第 10332 号（二审）。

[114] A. Feder Cooper and J. Grimmelmann, “The Files are in the Computer: Copyright, Memorization, and Generative AI”, SSRN Working Paper No. 4803118 (2024).关于语言模型中记忆的实证研究，另见本研究报告第一

制；而被伦敦法院则解读支持了以下结论：一般而言，扩散模型的权重不构成侵权复制件。这一比较表明，关于记忆的现有实证证据本身并不能完全决定法律分析的结果。模型所展现的、能够再现其训练数据中序列的能力，究竟应被视为该模型是训练语料库的复制件的证据，还是应被视为一种需通过输出过滤器和机器遗忘技术加以缓解的行为缺陷，这是一个法律定性问题。针对这一问题，不同法院根据证据记录得出了不同的结论，而这些证据记录虽然在形式和实际提交给各法院的证据内容上存在差异，却基于相似的基础科学文献。[115]如同第一部分提到的，某个模型是否包含受保护作品的复制件或者相关权的客体，应当逐模型判断。

“模型即复制件”这一问题在美国以另一种程序上不同的形式被提出：部分原告不仅主张训练过程产生了复制件，还主张训练后模型本身就是一项侵权的演绎作品。在卡德雷（Kadrey）诉 Meta Platforms 公司案中，作者主张，LLaMA 模型若不利用从其书籍中提取的表达性信息便无法运行，这些信息被保留在模型权重中，因此这些模型属于未经授权的演绎作品——这种定性将使模型本身承担法律责任，而无需证明任何特定输出与模型训练所基于的作品之间存在对应关系。法院在诉答阶段拒绝采纳该理论。被告最终以“合理使用”为由胜诉（详见下文第 4 节），其辩称这些权重是从训练数据中推导出的统计参数，而非存储的表达内容。[116]

在 Andersen 诉 Stability AI 有限公司一案中，针对潜在扩散模型提出了一种类似的理论，该案原告将该模型描述为存储训练图像复制件的替代手段；但在缺乏与受保护作品高度吻合的输出的情况下，这一定性在举证方面遇到了困难，这一困难与英国法院在 Getty 案中的结论相呼应，即扩散模型的权重并不存储训练数据。美国的经验表明，仅证明受保护作品在训练中被摄取，并不足以认定训练后模型是这些作品的演绎作品。[117]截至本文撰写之时，上述任何问题均未作出上诉裁决。

2.文本与数据挖掘例外

第二个问题出现在欧洲联盟以及版权法受欧洲法律框架影响的司法管辖区（例如欧洲经济区国家），即《CDSM 指令》第 3 和 4 条中的文本与数据挖掘例外是否适用于生成式人工智能模型的训练。本研究第二部分 B.5 节探讨了这一实体法理争议。为使论述完整，本文在此概述了德国关于该问题的判例法，重点关注其对权利人主张的执行所提供的启示。

德国在 Kneschke 诉 LAION e.V.案中就文本与数据挖掘例外对人工智能训练的适用性作出的两项裁决（一审：汉堡地区法院，2024 年 9 月 27 日；二审：汉堡高等法院，2025 年 12 月 10 日），已在上文第二部分 B.5 节中予以讨论。

部分脚注 18 中引用的库珀等人所做的研究。另见 M.A.Lemley 和 A.F.Cooper，“Probabilistic ‘Copies’ in Generative AI Models”，本研究报告第一部分脚注 25 所引。

[115] 关于 GEMA 案在此问题上说理的批判性评论，见 G. Frosio，“Copyright in Formaldehyde: How GEMA v OpenAI Freezes Doctrine and Chills AI”，Kluwer Copyright Blog （2025 年 12 月 10 日和 11 日），第 1 和第 2 部分。

[116] 如需了解开发者在诉讼情境之外的立场声明，见谷歌，“A Policy Agenda for Responsible Progress in Artificial Intelligence”（人工智能监管白皮书），可查阅：https://blog.google/company-news/outreach-and-initiatives/public-policy/white-paper-ai-regulation/，于 2026 年 7 月 7 日访问。

[117] 关于美国的演绎作品理论，见卡德雷诉 Meta Platforms, Inc.，案号：3:23-cv-03417-VC（加利福尼亚北区联邦地区法院，查布里亚法官，2023 年 11 月 20 日）（在诉答阶段驳回了关于 LLaMA 模型本身构成侵权演绎作品的主张，同时，允许基于训练中复制行为的直接侵权主张继续审理）；以及 Andersen 诉 Stability AI Ltd.，案号：3:23-cv-00201-WHO（加利福尼亚北区联邦地区法院，奥里克法官），2023 年 10 月 30 日和 2024 年 8 月 12 日的命令（除其他事项外，涉及一项主张，即该模型存储了训练图像的压缩副本）。下文 A.4 节将讨论卡德雷案中的合理使用裁决。

以下是关于权利执行的两点启示。首先，在文本与数据挖掘例外适用于人工智能训练的司法管辖区，权利人的权利执行处境受“选择退出”机制约束。在权利人以自动化流程能够识别并据此采取行动的方式有效表达了选择退出的情况下，其可通过普通侵权程序行使其权利。如果权利人未表达此类选择退出，或者该保留条款不符合适用要求，则文本与数据挖掘例外可能适用（须符合其其他条件），从而限制基于复制权提出的侵权主张。在这方面，权利人能否保留权利在一定程度上取决于“选择退出”信号所采用的技术形式。

其次，机器可读性的标准并非一成不变。汉堡上诉法院明确认定，必须根据有争议使用发生时普遍具备的主流技术能力来评估机器可读性。就本案事实而言，这指的是 2021 年下半年的技术发展水平，当时，能够从非结构化自然语言文本中提取语义的大语言模型尚未广泛普及。该裁决表明，选择退出在 2021 年或许不能被机器读取，但如今可能已可被机器读取；而在 2025 年尚不可被机器读取的选择退出，未来也可能可被机器读取。这对权利执行的实际影响是，随着技术的发展，权利人可能需要重新审视并在许多情况下重新声明其“选择退出”；对人工智能开发者的实际影响则是，某项操作在实施之时的合法性，并不因相关能力此后的变化而被否定。[118]

3.对域外训练的管辖权

第三个问题的实际重要性日益凸显，涉及国家法院对在位于另一国境内的基础设施上进行的训练操作是否具有管辖权。加拿大关于该问题的主要裁决是安大略省高等法院基梅尔（Kimmel）法官于 2025 年 11 月 7 日就多伦多星报有限公司诉 OpenAI 公司一案作出的。[119]加拿大主要媒体机构组成的联盟已对 OpenAI 提起诉讼，指控被告使用其内容来训练 ChatGPT，此举违反了加拿大版权法，也违反了保护该类内容的技术保护措施以及相关网站的使用条款。OpenAI 以法院无管辖权为由提出驳回动议，辩称训练行为是在位于美国的服务器上进行的，因此超出了《加拿大版权法》适用的领土范围。法院驳回了这一主张，认为该诉讼与安大略省之间存在真实且实质性的联系，且所指控的行为是否触犯加拿大版权法的问题涉及案情实质，而非管辖权分析的事项。该诉讼于 2026 年初进入证据开示阶段；被告已就管辖权裁决提起上诉，截至本研究撰写之日，该上诉仍在审理中。

“真实且实质性的联系”检验标准在人工智能训练方面的适用范围尚未确立。针对多伦多星报裁决的学术评论指出，该检验是在加拿大法律中关于“向公众传播权”的背景下形成的，例如在 SOCAN 诉 CAIP 案和 SOCAN 诉 ESA 案中，且将其转化为复制权并非必然简单。[120]从管辖权的角度来看，复制权在传统上被认定为位于复制/固定行为发生地。复制权的地域适用范围与《伯尔尼公约》第五条所规定的国民待遇原则之间的关系将在下文 B.2 节中探讨。

加拿大的其他进展则体现了可采取的各种程序性应对措施。在 CanLII 诉 Caseway AI 案中，CanLII 曾于 2024 年起诉位于温哥华的初创公司 Caseway AI，指控其为训练法律信息人工智能助手而大规模下载 CanLII 的司法判决数据库。CanLII 主张，尽管基础判决属于公有领域，但其数据

[118] P. Keller, “LAION Round 2: Machine-Readable but Still Not Actionable - The Lack of Progress on TDM Opt-Outs”, Kluwer Copyright Blog（2025 年 12 月 17 日和 18 日），第 1 和第 2 部分。

[119] 多伦多星报有限公司诉 OpenAI 公司，2025 ONSC 6217（基梅尔法官，2025 年 11 月 7 日）。原告方包括多伦多星报、环球邮报、Postmedia 以及加拿大广播公司。以司法管辖权缺失为由提出的驳回动议，所依据的论点是：被指控的训练行为发生在位于美国的服务器上，因此超出了《加拿大版权法》的地域适用范围。

[120] 加拿大作曲者、作者和音乐出版商协会（SOCAN）诉加拿大互联网供应商协会（CAIP），2004 SCC 45（“SOCAN 诉 “CAIP”）；另见加拿大作曲者、作者和音乐出版商协会诉娱乐软件协会，2022 SCC 30。关于该检验标准系针对向公众传播权而制定、可能不易转用于复制权的法理评述，已在针对多伦多星报案裁定的学术评论中提出。

库的组织结构、元数据以及内容筛选构成了一部受版权保护的汇编作品。双方于 2026 年 3 月以保密条款达成和解；由于双方已达成和解，本案未就上述问题的实体作出司法裁决。集体诉讼也在进行中：在不列颠哥伦比亚省，作家 J.B. 麦金农（J.B. MacKinnon）已对 Meta、英伟达、Anthropic 等公司提起集体诉讼，指控这些公司使用非法的作品复制件（源自 Books3 数据集）来训练其模型；在魁北克省，安妮·罗比拉德（Anne Robillard）已申请针对 Meta 和 OpenAI 提起集体诉讼的授权，要求每部作品判予最高 20,000 加元的法定赔偿金，这是商业侵权情形下的法定最高赔偿额。在上述A.1节中讨论的英国Getty诉讼案中，同样的结构性限制起到了决定性作用：一旦证据无法证明相关训练是在联合王国进行的，主要侵权主张便被放弃。[121]

在印度，德里高等法院在“ANI Media Pvt. Ltd.诉OpenAI OPCO LLC”一案中，维持了其对在美国服务器上进行的培训活动的管辖权，并基于初步证据认定，在国外存储作品仅是“一系列事件链条中的最后一步”，而该系列事件始于从印度进行的访问；同时认定开发者的异议“站不住脚，因为这将导致规避印度版权法”；[122]该判决的实质内容将在下文 A.4 节中进行讨论。在巴西，域外训练问题也出现在《圣保罗页报》（Folha de São Paulo）于 2025 年 8 月对 OpenAI 提起的诉讼中。该报在诉状中指控被告未经授权，将其新闻内容用于训练 ChatGPT，并且在向巴西用户提供服务时也使用这些内容。[123]该诉讼程序处于司法保密状态；双方于 2026 年达成和解，法院未就案情实质作出裁决。2026 年 8 月，阿根廷一家主要新闻出版商——佩尔菲尔出版社（Editorial Perfil）向联邦民事和商事法院提起诉讼，指控 OpenAI 和微软在生成式人工智能服务的训练和运行过程中未经授权使用了其新闻内容。[124]更宽泛地讲，还有人通过类比联邦最高法院关于在巴西运营的外国数字平台的判例法，提出了另一项主张——该主张在人工智能背景下尚未经过巴西法院检验。[125]根据这一主张，在巴西提供人工智能服务，意味着基础训练受巴西版权法约束，因为巴西用户由此生成的输出构成了训练行为的延续；从这一角度来看，如果训练是非法进行的，那么基于该训练提供服务，将根据《伯尔尼公约》第五条规定的国民待遇原则，受巴西版权法的管辖。[126]该主张实质上是以下原则的变体：如果某项违法行为始于境外，但却在法院

[121] CanLII 诉 Caseway AI 案于 2026 年 3 月以保密和解方式结案。截至本研究撰写之日，不列颠哥伦比亚省的 J.B. 麦金农针对 Meta、Nvidia、Anthropic 和 Databricks 提起的集体诉讼，以及魁北克省的安妮·罗比拉德针对 Meta 和 OpenAI 提起的集体诉讼，均处于认证阶段。关于商业侵权行为的法定最高赔偿额，见《版权法》，RSC 1985，第 C-42 章，第 38.1(1)(a)条。

[122] ANI Media Pvt. Ltd.诉 OpenAI OPCO LLC，德里高等法院，CS(COMM) 1028/2024（班萨尔法官，2026 年 7 月 24 日判决），2026 SCC OnLine Del 5291，第 51 段载有正文中引用的段落。该裁决是针对原告提出的临时禁令申请作出的；该诉讼仍在审理中，截至本研究撰写之日，上诉已提交至高等法院合议庭审理。

[123] 《圣保罗页报》诉 OpenAI，案号：1107237-96.2025.8.26.0100（TJSP）。该案于 2025 年 8 月提起诉讼，目前正在司法保密程序下审理；截至本研究撰写之日，尚未作出任何公开裁决。

[124] Editorial Perfil 股份有限公司诉 OpenAI 有限责任公司、微软公司和微软阿根廷有限责任公司一案（阿根廷联邦民事和商事法院，2026 年 8 月 24 日提起）。截至本文撰写之时，关于申诉的公开信息寥寥无几，但可见 Editorial Perfil, “Perfil demanda a OpenAI y Microsoft por el uso de contenidos periodísticos sin autorización”（2026 年 8 月 24 日）<https://www.perfil.com/noticias/judiciales/perfil-demanda-a-openai-y-microsoft-por-el-uso-de-contenidos-periodisticos-sin-autorizacion.phtml>，于 2026 年 9 月 29 日访问。另见 EFE, “El diario argentino Perfil demanda a OpenAI y Microsoft por uso indebido de sus contenidos”（2026 年 8 月 25 日）<https://www.swissinfo.ch/spa/el-diario-argentino-perfil-demanda-a-openai-y-microsoft-por-uso-indebido-de-sus-contenidos/91953813>，于 2026 年 9 月 29 日访问。

[125] 巴西联邦最高法院，第 1037396 和 1057258 号特别上诉案（合并审理）。联邦最高法院认定，如果平台在收到通知后未采取行动，则可能须对用户发布的内容承担责任；在巴西提供服务的平台必须遵守巴西法律；此类公司必须在巴西设有法定代表人。

[126] 《伯尔尼保护文学和艺术作品公约》第五条第 1 款（国民待遇）；另见第五条第 2 款（保护的独立性）。

所在司法管辖区产生影响，则可在该司法管辖区寻求救济；该原则被适用于这样一种技术场景——相关固定行为发生在面向用户的效果的上游。

4.替代问题

第四个问题涉及替代模型训练所使用的受保护材料或与之形成竞争关系的输出的法律定性。这个问题之所以重要，是因为它决定了权利人所主张的经济损害能在版权现有的类别内得到解决——这些类别涉及复制和向公众传播行为——还是需要采用第二部分所述的另一套机制，即针对那些单一输出不足以构成可诉行为的替代效应给予报酬。技术基础在见第一部分 C.3 节。

在美国，很多法院已经就这一议题进行了讨论。纽约时报公司诉微软公司案现已并入“关于 OpenAI 公司版权侵权诉讼案”，在该案中，被告在诉讼初期即提出动议，请求驳回直接版权侵权主张，理由是原告使用了被告定性为“精心设计且经过操控的提示词”来复现其所投诉的内容。2025 年 4 月，纽约南区地方法院拒绝以此为由驳回上述主张，该案随后进入证据开示阶段。在证据开示阶段，争议主要集中于是否需要提供面向消费者的使用日志。法院在一系列命令中——最终由斯坦（Stein）法官于 2026 年 1 月 5 日维持了 Wang 治安法官此前的一项命令——下令要求提供 2,000 万条 ChatGPT 用户交互日志的随机、去标识化样本。[127]该命令将证据重点从“能否通过对抗性提示诱发复现”转向“在普通消费者使用中复现发生的频率”。法院将广泛且非对抗性的用户输出证据视为与合理使用分析相关，特别是与模型的输出是否构成原告作品的市场替代相关。

加利福尼亚州北区于 2025 年 6 月相隔两天分别就巴茨（Bartz）诉 Anthropic PBC 案以及卡德雷诉 Meta Platforms 公司案作出的两项简易判决裁定中，两家联邦地区法院就主要问题得出了相同的结论，但其依据的理由却存在实质性差异。在巴茨案中，阿尔苏普（Alsup）法官认定，利用合法获得的书籍对大语言模型进行训练具有极强的转换性，构成《版权法》第 107 条所规定的合理使用；他同时认定，从 LibGen 和 PiLiMi 下载并留存这些书籍的盗版复制件不属于合理使用，且该案于 2025 年 7 月 17 日仅就盗版主张获得集体诉讼认证。此后，双方达成了 15 亿美元的和解，该和解于 2026 年 7 月获得最终批准。[128]在卡德雷案中，查布里亚（Chhabria）法官同样就合理使用问题作出了有利于被告的简易判决，[129]但其论证更侧重于合理使用的第四项因素（市场损害），而较少关注第一项因素（转换性使用），这与美国最高法院在“安迪·沃霍尔基金会诉戈德史密斯”案中的方法相呼应。此后的评论文章将本案解读为：其裁决主要取决于原告未能提交市场替代方面的证据，这意味着，若证据记录不同，结果也可能不同。[130]这两项裁决并未确立一项普遍原则，即人工智能训练是否属于合理使用。

---

[127] 关于 OpenAI 公司版权侵权诉讼案，案号：1:25-md-03143-SHS-OTW（纽约南区联邦地区法院）。多地区诉讼（MDL）合并审理新闻机构、作者及其他原告针对 OpenAI 提起的若干案件。关于 2,000 万条 ChatGPT 日志的证据开示命令，见本研究报告第一部分脚注 20 所列参考文献。

[128] 巴茨诉 Anthropic PBC，案号：3:24-cv-05417（加利福尼亚北区联邦地区法院，阿尔苏普法官，2025 年 6 月 23 日）（基于合理使用理由，就使用合法获得作品进行训练部分准许被告的简易判决请求，并驳回被告就下载和留存来自 LibGen 与 PiLiMi 的作品盗版复制件所提的简易判决请求。）。该集体诉讼仅就盗版相关主张获得了认证。2025 年 9 月 25 日，一项 15 亿美元的和解协议获得初步批准；2026 年 5 月 14 日，马丁内斯-奥尔金（Martínez-Olguín）法官主持了最终公平听证会，随后将该案留待裁决。截至本研究撰写之日，法院尚未作出最终批准。该和解协议于 2026 年 7 月获得地区法院的最终批准，该案现已结案。

[129] 卡德雷诉 Meta Platforms 公司，案号：3:23-cv-03417-VC（加利福尼亚北区联邦地区法院，查布里亚法官，2025 年 6 月 25 日）。

[130] 关于两位法官在说理上的差异，见（众多评论中）Wiggin 和 Dana 于 2025 年 9 月 18 日发布的文章中的分析，可查阅 wiggin.com。

在汤森路透企业中心有限责任公司诉罗斯智能公司一案中，美国第三巡回上诉法院在中间上诉中维持了部分简易判决，作出了美国首例关于合理使用与人工智能训练的上诉裁决。被告曾使用原告的判例要旨（即为 Westlaw 研究平台编制的法律要点编辑摘要），来训练一款非生成式人工智能法律研究工具。关于第一项因素，法院认定，该使用至多仅具有最低限度的转换性，因为其与原告使用判例要旨所达到的法律研究目的基本上相同；这一结论，加之该使用的商业性质，使得第一项因素不利于认定合理使用。法院将“中间复制”类案件与本案区分开来，理由是：那些案件中的复制对于获取软件中不受保护的功能性要素是必要的，而本案中并不存在类似必要性来为复制判例要旨提供正当理由。第二项因素对被告有利，但正如法院所指出的，该因素通常不会产生重大影响；第三项因素则不利于认定合理使用，因为整篇判例要旨被原样复制，缺乏转换性方面的正当理由。关于第四项因素，法院认定损害体现在三个方面：一是判例要旨作为吸引订阅原告平台的内容，其价值受到损害；二是被告直接参与竞争的法律研究平台原有市场受到损害；此外，在完成合理使用分析后，许可判例要旨作为人工智能训练数据的潜在衍生市场也受到损害。法院指出，该市场正在迅速发展，并认为原告曾利用其判例要旨来训练自身的人工智能产品，其决定不向他人授予许可并不能证明该市场不存在，且该未经授权的复制行为剥夺了原告进入该市场的机会。法院对巴茨案和 OpenAI 案中有争议的生成式模型进行了区分，因此，在生成式模型案件中，对转换性的评估可能有所不同。尽管如此，该判决在上诉阶段确认，对创作者通常会自行开发或许可他人开发的内容进行许可的这一新兴市场，是第四项因素下可予考量的因素。[131]音乐行业的一系列和解并未终结该领域的诉讼。2026 年 1 月，一群音乐出版商针对巴茨诉讼中的开发商提起了另一项诉讼，指控其使用了两万多部音乐作品，并请求判予超过 30 亿美元的法定赔偿金；同时，德国和丹麦的集体管理组织仍在继续对另一家音乐生成软件开发商提起诉讼。[132]

在美国境外，也有法院涉及到了这一议题。上述《圣保罗页报》诉 OpenAI 案曾提出这一问题，但双方最终达成和解。在撰写本文时仍待决的一个案例是北京互联网法院审理的一起案件，在该案中，四名插画师起诉小红书，指控该平台未经授权使用他们的原创作品来训练其人工智能图像生成模型 Trik，且该模型能够生成与原告作品实质性相似的图像。[133]

---

[131] 汤森路透企业中心有限责任公司诉罗斯智能公司案，案号：1:20-cv-613（特拉华州联邦地区法院，比巴斯（Bibas）法官，2025 年 2 月 11 日）（就直接侵权问题部分准许原告的简易判决请求，并驳回了合理使用抗辩）。地区法院的裁决在中间上诉中得到维持：案号：25-2153（第三巡回上诉法院，2026 年 6 月 11 日进行辩论，2026 年 9 月 29 日作出裁决）（由蒙哥马利-里夫斯（Montgomery-Reeves）法官出具意见）。上诉法院在一条脚注中指出，美国司法部在“关于 OpenAI 公司版权侵权诉讼案”，案号：1:25-md-3143（纽约南区联邦地区法院，2026 年 9 月 1 日）中提交了《利益声明》，该声明主张生成式大语言模型的训练属于转换性使用，并区分了生成式语境与本案待审情形。

[132] 2026 年 1 月 28 日，环球音乐出版集团、协和音乐集团和 ABKCO 在美国加利福尼亚北区联邦地区法院对 Anthropic 提起第二起诉讼，指控其侵犯了 20,000 余部作品的版权。该诉讼不同于这些出版商于 2023 年针对 Anthropic 提起的首起歌词案；后者最初在田纳西州中区联邦地区法院提起，之后移送至加利福尼亚北区联邦地区法院。见 Music Business Worldwide，“UMG, Concord and ABKCO sue Anthropic for $3bn in what could be single-largest non-class action copyright case”（2026 年 1 月 28 日）<https://www.musicbusinessworldwide.com/umg-concord-and-abkco-sue-anthropic-for-3bn-in-what-could-be-single-largest-non-class-act/>，于 2026 年 7 月 3 日访问；关于 2023 年的案件，见 Rolling Stone，“Music Publishers Sue Amazon-backed AI Platform for Copyright Infringement”（2023 年 10 月 18 日）<https://www.rollingstone.com/music/music-news/umg-concord-abkco-sue-anthropic-ai-company-copyright-infringement-1234856761/>，于 2026 年 7 月 3 日访问；见 Music Business Worldwide，“Universal, Concord and ABKCO ask court for injunction to stop Anthropic AI from using their lyrics”（2023 年 11 月 16 日）<https://www.musicbusinessworldwide.com/universal-concord-and-abkco-ask-court-for-injunction-to-stop-anthropic-ai-from-using-their-lyrics/>，于 2026 年 7 月 3 日访问。尽管唱片公司已于 2025 年底与该开发商以及 Udio 达成了和解，但 GEMA（德国）和 Koda（丹麦）等仍继续对 Suno 提起诉讼。

[133] 北京互联网法院（2023 年 11 月提起；截至本研究撰写之日，案件仍在审理中）；四名原告均为插画师，指控小红书未经授权使用其原创作品来训练人工智能图像生成模型 Trik。

2025 年 1 月，大韩民国三大地面广播公司——KBS、MBC 和 SBS——向首尔中央地方法院起诉 Naver，指控其未经授权使用这些广播公司的新闻内容来训练Naver的HyperCLOVA系列语言模型。这是韩国首例将训练问题直接提交法院的案件；鉴于《韩国版权法》第 35-5 条与《美国版权法》第 107 条具有共同渊源，上文所述的美国合理使用判例法已成为向韩国法院提交的论据的一部分。[134]在巴西，一名配音演员起诉了一家购物中心，称其未经授权在广告中使用了他的声音；该购物中心辩称，涉案的声音是由人工智能生成的，这一抗辩获得一审法院支持。然而，在 2024 年，圣保罗上诉法院推翻了该裁决，认为“人工智能的使用本身并不会消除（恰恰相反，反而加剧了）滥用第三方权利的风险”，因为此类工具是使用未经权利人授权、通过抓取从网上提取的内容进行训练的。该法院还认定需要进一步收集证据，并将该案件发回一审法院；截至本研究撰写之日，该法院尚未就案情实质作出最终裁决。[135]

相比之下，在印度，关于训练的问题迄今为止已得到了除美国和德国以外最全面的司法审理。在“ANI Media Pvt. Ltd. 诉 OpenAI OPCO LLC”一案中，一家新闻社指控其作品未经授权被用于训练被告的语言模型以及生成输出内容。在就该新闻社提出的临时禁令申请作出裁决时，德里高等法院在确认其对在美国境内服务器上进行的训练行为具有属地管辖权（见上文 A.3 节）后， 德里高等法院认定，训练行为涉及复制权，因为《1957 年版权法》第 14 条(a)款第(i)项将该权利延伸至“通过电子手段在任何介质上”存储作品的情形，且训练后删除副本并不能取消存储行为。然而，法院基于初步证据认定，该存储行为属于第 52 条第(1)款(a)项公平交易条款的适用范围，该条款涵盖以“私人或个人使用，包括研究”为目的的使用行为。“研究”一词被赋予了与时俱进的解释：“ChatGPT 底层大语言模型（LLM）的训练过程可视为一种旨在产生新知识的研究形式”，且“进一步研究的行为不能仅限于人类的行为”，因为“归根结底，研究是应人类要求并为人类利益而进行的”。被告的商业性质亦未使该使用行为超出该条款的适用范围：法院将该使用定性为“纯粹私人的”，训练副本保留在封闭环境中，仅模型可访问而非公众可访问，并且第 52 条第(1)款(a)项并不包含对商业用途的排除。

在公平性方面，法院认定，对作品的使用仅限于模型的训练；案卷中除无证据支持的断言之外，没有证据表明存在市场替代或收入损失；且该技术带来的公共利益在权衡中占据重要地位；法院还指出，原告本可以阻止被告的爬虫程序，但并未采取此类措施。法院驳回临时禁令申请的另一理由是，即使在审判中证实存在损失，也可以通过损害赔偿予以补偿；法院指出，如果要求从每个同类新闻来源获取许可，将使大语言模型的开发在经济上不可行。该裁决为临时裁决：诉讼将进入审判阶段，截至本研究撰写之日，上诉案正在合议庭审理中。尽管如此，该案仍具有相当大的比较研究价值。德国法院曾探讨“针对特定目的的文本和数据挖掘例外”是否涵盖人工智能训练，美国法院则审视了训练是否属于“合理使用”范畴；德里高等法院则通过一项列举“私人或个人使用，包括研究”情形的“公平交易”条款，在现阶段，倾向于开发者一方，处理了同一问题，其论证逻辑（包括未证实存在市场损害、存在未被利用的“退出机制”、该技术所带来的公共利益等）与前述美国司法判例中所考量的因素如出一辙。

---

[134] KBS、MBC 和 SBS 诉 Naver 公司，首尔中央地方法院（2025 年 1 月 13 日提起）。这些主张是根据《版权法》和《防止不正当竞争法》提出的，原告寻求赔偿损失和禁令，以禁止进一步使用。2026 年 2 月，这些广播公司在同一法院对 OpenAI 提起了类似诉讼。

[135] 圣保罗州法院，民事上诉案第 1119021-41.2023.8.26.0100 号 （私法第六庭，主审法官 Costa Netto，2024 年 10 月 31 日判决），于 2025 年产权组织知识产权法官论坛上介绍；见 <https://www.wipo.int/wipolex/zh/judgments/details/3217>，于 2026 年 9 月 19 日访问。

B.针对人工智能开发者执行权利

A 节中梳理的判例法反映了过去十八个月来的实质性进展。B 节探讨了一个结构性问题：鉴于技术的跨境特性以及此类争议中当事方之间通常存在不对等，权利人如何向人工智能开发者主张其权利。

1.权利执行格局的结构性特征

有三个结构特征反复出现。首先，上文 A.3 节中指出的跨境问题持续存在且严重。生成式人工智能模型的训练通常在位于权利人所在司法管辖区以外司法管辖区的基础设施上进行；随后，该模型会在多个司法管辖区同时部署和使用；其输出则在每位用户的司法管辖区内生成。上述每个阶段都可能涉及不同国家的法律。迄今为止，判例法尚未就不同的国家法律制度之间如何相互作用这一问题形成统一结论，开发者在其提供服务的各司法管辖区可能面临不一致的义务。

第二，《伯尔尼公约》第五条第 1 款所规定的国民待遇原则发挥着作用。根据该原则，作者在作品起源国以外的联盟成员国中，享有这些国家法律赋予本国国民的权利，以及本公约特别赋予的权利。该原则属于一项非歧视原则，其本身并不决定国内法的地域适用范围。

第三，利用程序性机制对主张进行合并处理，其中以美国的多地区诉讼为主要例证。将新闻机构、作者及其他原告在纽约南区法院针对 OpenAI 提起的多起案件合并为一项多地区诉讼，使原本可能分散在各单独诉讼中的程序得以集中审理。同样，巴茨集体诉讼汇总了涉及约 482,000 本书籍盗版复制件的各项主张，并如上文所述，达成了规模巨大的和解协议。在缺乏集体诉讼（或“集体索赔”）机制或多区诉讼机制，或者此类机制发展尚不完善的司法管辖区，个人权利执行所需的成本可能高得令人望而却步，特别是对于个人作者和小型权利人而言。A.3 节中所述的魁北克集体诉讼试图在不同的程序环境下运用相同的机制；不列颠哥伦比亚省的集体诉讼也呈现出类似的情况。在缺乏此类机制的司法管辖区，个人权利人针对资金雄厚的开发者行使权利的实际能力受到显著限制。

2.当事方之间的不对等

除了上述特征之外，判例法还指出了使实际权利执行变得复杂的若干不对等因素。首先是信息不对称：开发者掌握着技术记录（数据集清单、爬取日志、训练配置），原则上，即使这些记录并不完整，也可据此确定训练语料库的构成；而权利人则无法获得同等程度的访问权限，除非通过提取或其他实证演示产生间接证据。这是下文 C.1 节中透明度讨论的主题。其次是资源不对称：开发者，尤其是大型前沿模型开发者，所掌握的资源远超大多数权利人。鉴于如今此类案件所预期的技术专家工作，开展一项持续的权利执行行动所需成本相当高昂。第三是法律上的不确定性：当法律规定尚不明确时，已公布的判例法可能会受到选择效应的影响，因为当事方会就某些争议达成和解，而就其他争议提起诉讼；与此同时，权利人发起试验性案件的能力可能受到诉讼成本的限制。因此，在评估由此产生的判例法时，应将这些选择效应纳入考量。

C.执行权利面临的挑战

在当前的法律和技术状况下，权利人若想针对人工智能开发者执行其权利，将面临三大主要挑战。这些挑战将依次予以讨论。

1.透明度

第一项挑战是训练过程的透明度有限。除极少数例外情况外，人工智能开发者并未披露其训练语料库的构成；即使进行了概括性披露，其笼统程度也不足以让权利人确定其特定作品是否被使用。这种不透明性因商业秘密主张而进一步加剧：例如，2026 年初，某音乐生成系统的开发商在大型音乐公司提起的诉讼中，以竞争性损害为由，提出动议，请求就连其训练数据集的规模也予以保密。[136]其中一个原因是，训练语料库的构成可能属于商业资产，全面披露其内容可能会影响开发者的竞争地位。其结果是造成了信息不对称，这可能会限制权利人识别潜在侵权行为并采取权利执行行动的实际能力。

欧洲联盟已通过一项明确的法定披露要求来应对这一问题。《人工智能法案》第 53 条第 1 款 d 项规定，通用人工智能模型的提供者必须按照欧盟人工智能办公室提供的模板，就用于训练该模型的内容编制并公开发布一份“足够详细的摘要”。[137]具体实施在一定程度上取决于要多详细才能满足该条款的要求。此外，该条款受到限定，明确承认有必要尊重商业秘密和机密商业信息；而该要求的实施，目前是持续的政策与标准化工作的主题。[138]披露义务与商业秘密例外之间的这种矛盾引发了广泛的讨论。该义务最终的适用范围，很可能既取决于 2025 年 7 月人工智能办公室发布的模板的适用情况，也取决于各国司法机关及其他主管部门对该例外的解释。

其他司法管辖区通过不同的手段来处理透明度问题。在中国，有一套广泛的标识义务制度适用于人工智能生成内容。[139]中国的标识制度主要目的并非保护版权，而是为了防止欺诈、误导和公众混淆；然而该制度对权利人产生的实际作用是使一定程度的反向推断成为可能，因为对人工智能生成内容进行标识，既能识别特定输出为人工智能生成，又能据此调查训练中是否可能使用了受保护作品。

在大韩民国，自 2026 年 1 月 22 日起生效的《人工智能发展与信任基础构建基本法》要求事先告知生成式人工智能服务用户，并必须对人工智能生成的内容进行标识。[140]与中国的规定类似，这些标识义务并非主要旨在保护版权；但可能有助于基于输出进行类似的反向推断。版权问题依据《版权法》而非上述框架性立法处理。这种对法律文书进行分工的做法反映了欧洲联盟《人工智能法案》与版权现行法之间的关系。

在巴西，相关立法措施载于第 2338/2023 号法案。该法案的现行文本规定，人工智能开发者必须通过发布摘要，披露其在系统开发过程中使用受保护材料的情况；但需注意，此项义务须在

---

[136] Music Business Worldwide，“Suno moves to keep size of its AI training data sealed in UMG and Sony’s copyright case, citing competitive harm”（2026 年）<https://www.musicbusinessworldwide.com/suno-moves-to-keep-size-of-its-ai-training-data-sealed-in-umg-and-sonys-copyright-case-citing-competitive-harm/>，2026 年 7 月 7 日访问。

[137] 本研究报告第二部分脚注 82 中提及的第(EU) 2024/1689 号条例第 53 条第 1 款 d 项。

[138] 《人工智能法案》序言第 107 段明确指出，公布摘要的义务应以尊重商业秘密和机密商业信息的方式进行规定，且不得要求披露此类信息。

[139] 《互联网信息服务算法推荐管理规定》（中华人民共和国），2022 年 3 月 1 日起生效；《互联网信息服务深度合成管理规定》（中华人民共和国），2023 年 1 月 10 日起生效；《生成式人工智能服务管理暂行办法》（中华人民共和国），2023 年 8 月 15 日起生效；《人工智能生成合成内容标识办法》（中华人民共和国），2025 年 9 月 1 日起生效；以及由国家市场监督管理总局和国家标准化管理委员会于 2025 年批准的同名强制性国家标准。

[140] 《人工智能发展与信任基础构建基本法》（大韩民国），于 2025 年 1 月颁布，自 2026 年 1 月 22 日起生效。

遵守商业和工业秘密的前提下履行。[141]该披露义务的实际效果将取决于对商业秘密例外作的解释有多宽泛。这一结论经必要调整后也适用于欧洲框架。

有两点结论与权利执行中的透明度问题相关。首先，透明度虽然可能有助于权利执行，但仅靠透明度本身还不够。无论摘要多么详细，其公开本身并不会构成诉讼理由；权利人仍须证明其作品被使用、该使用侵犯了其专有权，且不适用任何例外。透明度义务虽然能减少信息不对称，但无法将其彻底消除。其次，透明度义务的设计与“选择退出”机制的设计相互作用。已表明选择退出且其作品在透明度摘要中被标注为已纳入训练语料库的权利人，在符合适用法律及相关例外的其他条件的前提下，可能拥有更明确的依据来提出权利执行主张；实际问题在于，该摘要是否足够具体，以使其作品能够被识别。

2.数据攻击、过滤及选择退出标准激增

第二项挑战涉及权利人、开发者和中间商为应对影响权利执行的训练和推理技术特征而采取的技术手段。可以识别出四类不同的技术手段。

第一类是提取攻击，此类技术源自安全研究文献，如今在诉讼中被援引。提取攻击是指通过精心设计的提示，诱使训练后模型复现其训练数据中内容的技术。本研究第一部分所述的该领域研究成果，已被安全研究文献引用为证据，表明可提取性可能是大型生成式模型的一种结构性属性，而不仅仅是特定提示策略造成的结果。[142]从权利人的角度来看，提取攻击是一种证据收集技术：可用于证明特定作品被纳入了训练中或可以从模型中还原。从开发者的角度来看，此类攻击是一种漏洞，可通过训练数据去重、输出过滤和遗忘技术来加以缓解。正如本研究第一部分 C.2 节所述，一个法理层面的问题现已提交若干司法管辖区法院审理：引出被记忆的输出的方式，是否应当影响版权实体分析。

第二类是输出过滤器和输入过滤器这组技术。开发者越来越频繁地部署过滤器，以阻止模型复现某些已识别的内容（输出过滤器），或在数据准备阶段将某些内容从训练语料库中排除（输入过滤器）。[143]此类过滤器在实质分析中的作用尚无定论。在美国，巴茨案和卡德雷案的裁决以不同方式论及了过滤器的作用，其中查布里亚法官特别指出，市场替代问题在一定程度上取决于过滤器能否有效防止模型生成将替代特定原作品的输出。在欧洲联盟，《人工智能法案》中的透明度义务涉及开发者在版权方面所遵循的政策；这些义务的履行可能影响对过滤做法是否构成开发者整体合规状况一部分的法律定性。

第三类是数据投毒技术，即权利人或第三方将恶意、错误或混乱的数据引入数据集，从而破坏使用这些数据集训练任何模型的性能。在中国，国家互联网信息办公室已将数据投毒视为一种安全威胁，并启动了一项专项行动（“清朗·整治 AI 应用乱象”）予以打击；推荐性国家标准《生成式人工智能服务安全基本要求》要求对训练数据进行过滤，以剔除非法和有害信息。[144]在调

---

[141] 第 2338/2023 号法案（巴西），第 62 条（提交众议院的版本）；关于该法案的立法现状，见本研究报告第二部分脚注 100。

[142] 关于提取攻击，见第一部分脚注 18（Cooper、Gokaslan、Ahmed、Cyphert、De Sa、Lemley、Ho 和 Liang）和脚注 22（Carlini 等(2021)以及 Nasr 等(2025)）。

[143] 输入阶段的过滤可能涉及排除元数据中注明保留权利的作品、对语料库进行去重以降低记忆风险，以及排除来自已知未获许可来源的内容。输出阶段的过滤可能涉及检测并屏蔽与已识别受保护作品高度吻合的输出。不同开发者以及不同模型类型对输入阶段和输出阶段过滤的重视程度各不相同。

[144] 中国国家互联网信息办公室，“清朗·整治 AI 应用乱象”专项行动；推荐性国家标准《网络安全技术——生成式人工智能服务安全基本要求》，2025 年 11 月 1 日起生效，第 4.2.1 条（训练数据内容过滤）。

查的其他司法管辖区中，目前尚未出现针对数据投毒的类似监管措施。作为权利人的一种自助机制，数据投毒的实际重要性在任何司法管辖区尚未得到权威评估；其引发的技术和法律问题基本仍是开放的：针对人工智能开发者部署投毒技术，是构成对开发者所享任何权利的可认定侵权，还是权利人对其自身权利的正当行使。

第四类是与欧洲权利执行特别相关的一类，即以机器可读形式的选择退出标准的问题。[145]上文第二部分 B.5 节讨论的汉堡上诉法院在 Kneschke 诉 LAION e.V. 案中的裁定认为，根据《CDSM 指令》第 4 条第 3 款，有效的退出选择必须既可被机器读取，又可被机器处理。该要求的技术实现方案尚未确定。互联网工程任务组的人工智能偏好工作组正在制定人工智能相关使用偏好的通用词汇表，截至 2026 年中期，该工作仍在进行中；工作草案仍在制定中，目前仅界定了少量使用类别（实质上是人工智能训练和搜索相关用途，另有一个“检索增强生成”类别正在讨论中），且尚未通过任何标准。[146]一个值得探讨的相关问题是，选择退出义务与《伯尔尼公约》所保护的权利的产生或行使不得要求任何手续的规定之间如何协调。这一问题的答案可能取决于选择退出机制的具体设置及其法律效力。

若干相互竞争的标准已同时出现，且可分为两组，而这两组容易被混淆。第一组仅表达权利保留，即选择退出。该组包括文本与数据挖掘权利保留协议（TDMRep）和 TDM•AI，两者均专门用于按照《CDSM 指令》第 4 条第 3 款实现选择退出；Cloudflare 的内容信号机制，该机制是对机器人排除协议（robots.txt）的一项专有扩展；以及 C2PA 和 IPTC PLUS 等更广泛的倡议，这些倡议通过各自的词汇表提供选择退出功能。此类信号已在实践中用于限制人工智能对话式搜索引擎的爬虫：据报道，相当一部分主流新闻网站已屏蔽某款此类搜索引擎所使用的爬虫。[147]第二组标准更进一步，允许权利人在同一网络基础设施中嵌入明示的、机器可读的许可条款。因此，问题不再仅仅是某部作品能否被使用，而是以何种条件使用，包括价格，以及按爬取付费或按推理付费两种模式之间的选择。“真正简单许可”（RSL）标准——一项由美国在线出版商和内容分发网络支持的倡议——便是其中的典范；Cloudflare 的按爬取付费模式则是另一个例证，该模式在访问请求本身这一层面运作，对未获许可的爬虫返回“需付费”响应。在第二组中，自动化代理所接收到的信号不仅起到“选择退出”的作用，同时也构成许可要约。引入人工智能代理有望促进该领域的自动化许可。

合规与裁决因这两类权利保留并存于同一项尚无定论的标准讨论中而变得复杂。正如 Cloudflare 通过其内容信号和按爬取付费机制所体现的，同一家服务提供商可能同时采用这两组标准，因此某一信号究竟仅表达权利保留，还是构成许可要约，从表面上看，并不总是显而易见的。欧盟委员会已就《人工智能法案》和《通用人工智能实践准则》中关于文本与数据挖掘权利

[145] 关于 LAION II 案裁定以及“选择退出”标准激增，见第二部分 B.5 节，以及其中引用的 Keller 评论。

[146] 关于互联网工程任务组（IETF）的人工智能偏好工作组，见 https://datatracker.ietf.org/wg/aipref/about/。文本与数据挖掘权利保留协议（TDMRep）、TDM•AI、“真正简单许可”（RSL）集体协议、Cloudflare 的 contentsignals.org 机制、C2PA 以及 IPTC PLUS 等，都是该领域的主要倡议。

[147] 2026 年发表的一项研究发现，美国和联合王国许多主要新闻网站在其 robots.txt 文件中屏蔽了 Perplexity 的爬虫 PerplexityBot。见 BuzzStream, “Most Major News Publishers Block AI Training & Retrieval Bots”（2025 年 12 月 17 日，2026 年 1 月发布）<https://www.buzzstream.com/blog/publishers-block-ai-study/>，于 2026 年 7 月 3 日访问。

保留的协议启动磋商，旨在制定一份普遍认可的机器可读协议清单，供通用人工智能模型提供商遵守，以履行其在《人工智能法案》下的义务。[148]

标准体系的碎片化本身就对权利执行构成了挑战。汉堡上诉法院的裁决表明，应根据有争议使用发生时的技术能力来评估“选择退出”的机器可读性；但在多种互不兼容的标准并存的背景下，开发者必须遵守哪些标准才能符合法律要求这一实际问题并没有明确答案。针对该裁决的评论中指出的风险是：若缺乏统一、开放治理的标准，机器可读的“选择退出”可能为新的中介层奠定基础，而这个中间层会使按爬取付费或按点击付费的开放网络访问成为常态。该风险是否会成为现实，取决于未来几年标准化机构和政策制定过程中作出的选择。

选择退出机制构成了多层权利管理基础设施的一部分。第一层由标识符和元数据层构成，包括针对作品和贡献者的既定国际标准标识符，以及可以将作品与其权利状态相关信息联系起来的、机器可读的权利管理和来源元数据。目前，相关工作正在开展，旨在汇总适用于人工智能环境、具有互操作性、可信且机器可读的版权数据良好做法，尤其是由芬兰教育与文化部召集的版权基础设施工作队开展的工作，[149]其前提是，可靠的识别既是表达使用偏好的前提，也是事后追踪使用情况的前提。这一层以上述 WCT、WPPT 和《北京条约》中关于权利管理信息的条款为支撑。

第二层由访问控制层构成，包括权利人和平台为确保只有被授权方才能访问其内容而采取的应用程序接口和反爬虫措施，以及流媒体和出版行业为防止未经授权提取内容而采用的技术保护措施。这一层以上述 WCT、WPPT 和《北京条约》中关于反规避技术的条款为支撑。

第三是归属层，即评估特定作品对某项输出的贡献程度，以便在适用情况下分配报酬。从这个意义上讲，归属应与检测区分开来。检测技术旨在以一定的概率确定文件是由人还是机器生成的；这类技术不能识别源作品，且仅凭这些技术本身无法为版权主张提供支持。相比之下，归属技术旨在估算已识别作品在生成输出中所占的比例，因此需要开发者在一定程度上予以配合并保持透明度。对该能力的投资已经开始：一家已参与第二部分所述的瑞典集体许可安排的归属技术提供商，于 2026 年被一家大型唱片公司收购。[150]归属层与第二部分 B.6 节中提到的输出层面报酬模式直接相关。其中一些模式并非以输出阶段的侵权为基础，而是基于特定模型输出所产生的替代效应。

另一类来源追溯和水印工具，以及针对人类创作内容标识的倡议，涉及输出的真实性和来源，而非训练过程中对受保护作品的使用；这些工具主要涉及本研究未讨论的输出问题，在此提及这些工具仅为了将归属层置于更广泛的架构之中。

3.判例法概况

根据所调查的判例法体系，在生成式人工智能训练背景下，可以就当前版权权利执行得出若干跨领域的观点。

---

[148] 欧盟委员会，Consultation on protocols for reserving rights from text and data mining under the AI Act and the GPAI Code of Practice（2025 年）；见 https://digital-strategy.ec.europa.eu。

[149] 由芬兰教育与文化部牵头成立的版权基础设施任务组（CITF）于 2025 年发布了首份报告“Interoperable, trustworthy and machine-readable copyright data in the AI era”。

[150] 华纳音乐集团于 2026 年 6 月收购了归属技术提供商 Sureel。该提供商的技术亦用于本研究第二部分 B.3 节所述涉及 STIM 的许可安排。

首先，法院总体上采取了一定程度的审慎态度，通常不愿作出意在全面厘清法理框架的裁决。迄今为止的裁决涉及特定技术（Getty 案中的潜在扩散、GEMA 案中的语言模型、巴茨案中合法和非法获取的书籍），特定程序性立场（《纽约时报》诉讼中的驳回动议、巴茨案和卡德雷案中的简易判决）， 以及特定问题（GEMA 案和 Getty 案中的“模型即复制件”，巴茨案和卡德雷案中的合理使用等）。目前，这些裁决尚未在任何单一法律体系中建立起一个全面的框架。尽管德国的立场目前已体现在三项判决中，但其内部并不统一：LAION 系列案件认可将文本与数据挖掘例外适用于训练，而 GEMA 系列案件则采取了不同的做法。美国的立场同样呈现出不同层级，目前正等待上诉审查，最终将由美国最高法院作出一项或多项判决意见。

其次，尽管如此，判例法体系已形成了一些共识。要保护人工智能输出，必须由人类作者生成独创性这一根植于《伯尔尼公约》这一要求已形成了共识。在美国、肯尼亚[151]、德国的一审法院一级以及在欧洲联盟都可以看到这一点，欧盟法院在涉及家具版权保护的 Mio 案中重申的独创性标准也以人类作者为前提。[152]对于制作训练数据集是否适用复制权（在德国、中国和加拿大）也已有共识。最后，对于输出替代与训练分析的相关性（在美国，以及未来在欧洲联盟通过三步检验法的适用而可能具有的相关性）也有一些共识。判例法中仍存在分歧的若干问题包括训练后模型是否包含用于训练、再训练和微调的作品复制件，文本与数据挖掘例外是否以及如何适用于人工智能训练，以及跨境管辖权的适用范围边界在哪里。

第三个意见是上文 C.1 和 C.2 节中所述的实际挑战——即透明度、“选择退出”标准激增，以及开发者与权利人之间在成本和信息方面的不对等问题——很可能在未来数月乃至数年内影响判例法的发展。即使法院明确了版权法的实质性问题，权利人主张的实际可执行性仍将取决于权利人通过何种机制来核实其作品的使用情况、表达选择退出，并以法院可据以采取行动的形式提交证据。在这一领域，技术设计与法律分析紧密相关，权利执行将受到这两方面所作选择的影响。

结　论

本研究查明了两个领域中迅速且不均衡的发展情况：各司法管辖区正在尝试采用不同组合的例外、许可机制和报酬方案；判例法体系已阐明相关问题，但尚未形成一个稳定的框架。目前正在考虑采取立法或监管措施的成员国，所处的环境比十二个月前更为成熟；经验和判例法已达成共识的问题提供了一定基础；而仍存在分歧的问题，则需要作出政策选择——本研究力图呈现这些选择，但并不给出定论。这些选择将继续在国家层面作出；如果成员国作出决定，可通过版权及相关权常设委员会以及更广泛的产权组织框架在国际层面进行讨论。

---

[151] Aryeh Movement 有限公司诉 Cynthia Beldina Akoth Okello，COPTA/E001/2025 （肯尼亚版权法庭）（认定，《版权法》（第 130 章）第 22(3)条预设了在“足够努力”的投入中须有人类参与，因此由人工智能系统自主生成的作品无法达到该法定门槛，除非有充分的人为干预使作品具备独创性）。

[152] Mio 等案，C-580/23，与 USM 诉 konektra 案，C-795/23 合并审理（欧盟法院，2025 年 12 月 4 日判决）。这些合并审理的案件涉及应用艺术作品（具体而言是家具）的版权保护，（在 C-580/23 案件中为“Palais Royal”餐桌）。法院重申，只有当某一客体具有原创性——即作者本人的智力创作、体现了表达作者个性的自由与创造性选择——时，才构成受保护的作品；实用艺术作品与其他作品适用相同的原创性要求。该判决并未涉及人工智能；关于人工智能输出需由人类作者创作才能获得保护的要求，源于这一原创性标准，而非法院针对人工智能作出的任何裁决。此外，作者在权利执行程序中可能还会面临一个新障碍，即提供证据证明其作品并非由人工智能创作，或者至少证明其做出了足以获得版权保护的贡献。

本研究表明，各成员国在权利人与人工智能开发者利益之间的平衡方面可能采取不同的做法。分析中提出了五个更具体的问题。

第一个问题是，在构建语料库和训练模型过程中产生的复制行为是否需要获得授权，还是应属于例外或限制的范畴——无论是针对人工智能的特定规定、文本和数据挖掘条款，还是一般性条款；以及应满足哪些条件，包括合法获取、训练目的以及使用行为的商业性或非商业性。

第二个问题是，对任何保留机制（在采用这种机制的情况下）的设计。这包括“选择退出”的形式，同时需考虑到，机器可读的保留机制预设了个人作者和表演者可能并不具备的能力。反之，如果保留机制无法在较大规模上实现机器处理，则无法为开发者提供明确的合规路径。还应考虑“选择退出”机制是否符合《伯尔尼公约》第 5 条第 2 款的规定。

第三个问题是，补偿应针对内容摄取、输出替代，还是两者兼而有之；应采取基于专有权的市场许可形式，还是法定报酬主张形式；无论哪种情况，补偿应通过哪些机构进行流转。这包括集体管理授权是否涵盖训练用途、是否需要延伸的或新的集体许可形式，以及如何设计合同链，以确保价值流向其作品被使用的作者和表演者。

第四个问题是透明度：为了行使权利和管理报酬，需要对训练材料进行何种记录保存和披露，以及如何将这些义务与合法的商业秘密和竞争关切相协调。

第五个也是最后一个问题是，当从经过训练的模型中移除作品在技术上不可行时，补救措施是否仍然有效且相称；以及当训练在国外进行但由此产生的服务在国内提供时，应该如何行使管辖权。

[文件完]