Dispute Settlement 争议解决, 2026, 12(8), 77-82
Published Online August 2026 in Hans. https://www.hanspub.org/journal/ds
https://doi.org/10.12677/ds.2026.128239

Hans汉斯

人工智能训练数据池中的ODbL协议合规性研究

吴　若

西南民族大学法学院，四川　成都

收稿日期：2026年7月1日；录用日期：2026年8月3日；发布日期：2026年8月11日

摘　要

开源协议是开源大模型生态治理中的关键要素，对其认识不足、履行不当既可能引发知识产权侵权风险，也可能陷入投入产出失衡的“合规陷阱”，亟需合规指引。随着人工智能技术的飞速发展，高质量、大规模的训练数据成为核心竞争力。数据池作为汇聚多源数据的重要载体，其构建与使用过程中的法律合规性问题日益凸显。开放数据库许可协议(ODbL)作为一种“著佐权”式的数据许可协议，旨在确保数据的共享与后续使用的开放性，其在人工智能数据领域的应用引发了新的法律挑战。

关键词

ODbL协议，开源协议，人工智能训练数据，数据合规

Research on ODbL Protocol Compliance in AI Training Data Pool

Ruo Wu

Law School of Southwest Minzu University, Chengdu Sichuan

Received: July 1, 2026; accepted: August 3, 2026; published: August 11, 2026

Abstract

Open-source licenses are a critical element in the governance of the open-source large model ecosystem. Insufficient understanding or improper implementation of these licenses can not only lead to intellectual property infringement risks but may also result in a “compliance trap” where input and output are unbalanced, creating an urgent need for compliance guidance. With the rapid advancement of artificial intelligence technology, high-quality, large-scale training data has become a core competitiveness. Data pools, as important carriers for aggregating multi-source data, increasingly

文章引用: 吴若. 人工智能训练数据池中的 ODbL 协议合规性研究[J]. 争议解决, 2026, 12(8): 77-82.
DOI: 10.12677/ds.2026.128239

face prominent legal compliance issues in their construction and usage. The Open Database License (ODbL), as a "copyleft" style data license agreement, aims to ensure the openness of data sharing and subsequent use. Its application in the field of artificial intelligence data has sparked new legal challenges.

Keywords

Open Database License, Open-Source License, AI Training Data, Data Compliance

1. 引言

2025 年 1 月，杭州深度求索公司发布 DeepSeek-R1 开源大模型，重构全球人工智能产业"闭源主导"格局[1]，推动智能技术的普惠化进程，为新型产业生态演化提供了系统性支撑，同时，开源生态的勃兴虽预示我国智能产业正迈向认知范式重构新阶段，但蕴含的系统性知识产权风险也需审慎评估。开源协议是开源知识产权生态治理中的关键要素[2]，本质是权利人以其真实意思表示塑造的权利让渡契约，协议条款对于使用、复制、修改和分发"开源资源"的社会公众具有法律约束力。DeepSeek、LlaMA 等开源大模型的成功经验表明，人工智能的技术创新与产业发展不可避免地要依赖于"开源资源库"。对开源协议的认识不足，一方面可能因履行授权义务不当、对开源协议的责任传导效应认识不清，引发知识产权侵权风险。另一方面，过度履行协议义务可能引发合规成本超限，形成投入产出失衡的"合规陷阱"，为我国人工智能产业生态留下隐患。据此，本文立足开源大模型使用端，梳理 AI 领域主流开源协议条款，进行类型化解读，阐明其类型学特征与约束边界，多维度解构开源协议履行不足可能带来的知识产权风险，提出合规对策。

2. ODbL 协议的法律内涵与核心义务

开源是当代最具影响力的协同创新模式之一，通过汇聚创新资源、构建信任与协作的制度环境，开源模式实现了知识、智慧、技术与成果的高效共享，极大提升了创新要素的流转效率，已成为全球科技创新与产业协同发展的重要制度形态[3]。要分析合规问题，必须首先准确理解 ODbL 协议的法律架构与核心义务。ODbL 协议将数据库权利区分为数据库内容本身的版权或类似权利与数据库作为汇编整体的"数据库特殊权利"。

(一) ODbL 协议的法律内涵

开放数据共用许可协议是开放知识基金会(OKF)发布的许可协议类型，目的是规范和约束数据提供者和数据使用在重用、发布、修改和传播数据的责任与义务，是国际通用的许可协议类型。

开放数据共用许可协议是专门针对数据或数据库开放的许可协议类型，具有许可对象单一的特点。开放数据共用许可协议适用于复杂的数据集内容和数据库作品，不适用于计算机软件、电影、音乐等创作作品，许可对象单一；开放数据共用许可协议应用领域不够广泛。开放数据共用许可协议目前应用领域远不及知识共享许可协议广泛，国外政府数据开放中目前有法国使用开放数据库许可协议(ODbL)；开放数据共用许可协议许可层级较多。许可层级较多，公共领域贡献和许可(PDDL)将许可内容完全置于公共领域，开放数据共用署名许可(ODC-By)仅有署名的要求，开放数据库许可(ODbL)要求署名和相同方式

共享。开放数据共用许可协议是国际通用的开放许可协议，其包含的三种许可类型均符合开放的条件，不限制商业领域使用和对数据的修改、演绎等，其中限制最多的开放数据库许可(ODbL)也可以用于商业领域。

(二) “相同方式共享”义务

“相同方式共享”义务是 ODbL 协议最核心的条款，具有显著的“著佐权”特性，该条款强制要求任何基于 ODbL 数据构建的衍生作品，在公开发布时都必须沿用同一开放许可，旨在确保数据的公共属性不被私有化侵蚀。根据 ODbL 第 4.3 条，如果您公开使用一个基于 ODbL 许可的数据库的“衍生数据库”，那么您必须按照 ODbL 协议来许可该衍生数据库。协议对“衍生数据库”的定义非常宽泛，包括“对原数据库的全部或实质部分进行修订、修改、后续翻译、改编、编排，或其他任何能够使全部或实质部分可被重新编排或改编的形式”。这一宽泛的定义是后续所有合规争议的根源，因为它旨在涵盖任何基于原数据库创作的新数据库形式。

(三) 署名义务

署名权是开源许可体系的核心条款之一，也是作者人格权的基本体现[4]。ODbL 第 4.4 条规定，在使用 ODbL 数据库时，必须保留所有原始权利人的署名信息。这些信息通常包含在数据库的“归属文件”中。署名义务不仅要求在使用数据库中保留，在分发衍生数据库时，也必须以合理的方式向接收者传递这些署名信息。这对于由成千上万个数据源组成的数据池而言，构成了巨大的管理挑战。

3. AI 训练数据池使用 ODbL 数据的核心合规困境

将 ODbL 协议应用于 AI 训练数据池，其特殊性在于数据的使用目的和产出物均超出了协议设计时的典型场景，导致了一系列解释上的困境。

(一) “衍生数据库”的界定模糊性

AI 模型是否构成 ODbL 意义上的“衍生数据库”这是所有问题的核心。有学者持反对意见，认为著作权法保护的是“表达”而非背后的“思想”或“事实”，AI 模型不构成衍生数据库。理由在于：第一，训练完成的模型并不直接包含原始数据的副本。它是由权重参数、算法结构组成的抽象数学表示，无法通过反向工程还原出原始训练数据。第二，数据库的核心功能是存储、组织和查询数据，而 AI 模型的核心功能是基于输入进行推理和预测。二者在功能和表现形式上存在根本差异。第三，ODbL 协议中列举的“修订、修改、改编”等行为，传统上指向的是对数据库结构或内容的直接操作，而非通过机器学习进行的间接、抽象化学习。而持肯定意见的学者认为，在某些情况下，AI 模型或其训练过程可能构成衍生数据库。理由在于：首先，AI 训练必然读取并处理了数据库的“实质部分”，甚至全部数据。这种大规模的使用行为，其目的就是创建一个新的、具备价值的知识产品。其次，虽然模型本身不是数据库，但用于训练的数据集，即数据池中经过预处理、标注的子集，本身就是数据库。如果这个训练集大量包含了 ODbL 数据，并且其构建方式，例如数据选择、清洗、标注则体现了独创性的编排，那么该训练集本身就很可能被视为一个“衍生数据库”。而模型是这个衍生数据库的直接产物。最后，从协议促进数据共享的立法目的出发，将利用开放数据训练出的专有、闭源模型视为对共享精神的规避，可能促使法律解释倾向于将此类模型纳入“衍生”范畴，以维护协议的效力。

目前，尚无具有约束力的司法判例对此作出明确裁决，这使得法律风险真实存在且不确定。

(二) “署名”义务的履行困境

即使暂时搁置模型是否构成衍生数据库的争议，仅就训练过程而言，履行署名义务也极为困难。AI 数据池通常是海量数据的聚合体。一个训练集可能混合了来自成千上万个 ODbL 数据源、其他开源协议数据源以及自有数据源的数据。在模型训练过程中，精确记录每一条数据对应的归属信息，并在最终模

型中予以体现，在技术上和管理上成本极高，甚至不现实。即便保留了所有元数据，如何在最终发布的模型，例如一个 API 或一个软件库中，以“合理的方式”向最终用户展示这数以万计的署名？这几乎会破坏产品的用户体验和商业可行性。

(三) “相同方式共享”义务对 AI 模型的适用性挑战

假设法院或仲裁机构最终认定，使用 ODbL 数据训练的 AI 模型构成了“衍生数据库”的产出物或本身就是一种衍生形式，那么“相同方式共享”义务将带来颠覆性影响。这意味着整个 AI 模型，包括其架构、权重参数等都必须以 ODbL 或兼容的协议进行开源。这对于投入巨资研发、并将其作为核心竞争力的商业公司而言，是无法接受的。如果企业不直接分发模型，而是通过云服务的方式提供 AI 能力，这是否属于“公开使用”衍生数据库？ODbL 协议对“公开使用”的定义是否涵盖服务提供？这个议题暂无行业共识。如果答案是肯定的，那么即使不分发模型，也可能触发开源义务，这将极大地冲击 AI 即服务的商业模式。

4. 风险转化与责任承担分析

(一) 违约与侵权责任竞合

使用 ODbL 数据的前提是接受其许可条款。违反“相同方式共享”或“署名”义务，首先构成违约。由于 ODbL 许可是以遵守条件为前提的权利授予，违约将导致授权自动终止，其后的所有使用行为(包括训练和模型分发)即构成对原始数据著作权和/或数据库特殊权利的侵权。权利人可据此提起诉讼，主张停止侵害、赔偿损失。

(二) “许可证污染”与资产贬值风险

一个数据池若意外混入 ODbL 数据，可能导致基于该数据池训练的整个 AI 模型被“传染”上 ODbL 义务。这种“许可证污染”会使企业面临两难抉择：要么被迫开源核心模型，导致商业资产价值归零；要么停止产品分发，承担前期研发投入的沉没成本。

(三) 商业信誉与生态信任危机

在开源社区，信誉是无形的资产。违反 ODbL 等著佐权协议，会被视为对开源共同体基本规则的背叛，可能导致社区的口诛笔伐、开发者抵制与合作关系的破裂，对企业长期发展造成深远负面影响。

5. 构建多层次合规路径的思考与建议

(一) 数据池构建阶段许可协议的源头治理

开源生态发展初期，开源社区通过信息清理、技术筛查以及制度性设计有效降低“开源资源”中的知识产权瑕疵可能引发第三方权利冲突风险。人工智能技术的范式变革重构了开源生态的风险图谱。开源大模型通常利用开源数据库、公共互联网抓取的海量数据进行预训练，而后成为下游“服务大模型”的底层基础模型，例如，Stable Diffusion 便采用了预训练多模态模型 CLIP。不仅如此，Hugging Face 等开源平台还允许上传用户训练模型，供全球用户下载。对 Hugging Face 上的 7903 个项目进行实证研究表明，开源大模型的训练数据通常存在未经授权问题。我国学界通说认为，应将预训练阶段的数据使用行为纳入合理使用制度，而将“输出端”的数据传播侵权聚焦于开源模型使用者(服务提供者) [5]。

AI 训练数据规模大、权利主体分散、权益结构复杂且呈现非结构化特征，依托开源社区的传统审查模式难以实现有效溯源，开源大模型使用者必须建构自主性权利治理框架。一是数据源头控制。优先选用来源可追溯、产权明确或经认证的标准化数据集进行预训练的开源大模型。部署前，剔除存在“在先权益”的重复数据，降低下游版权侵权风险。二是技术管理维度。一方面，综合运用水印技术、内容相似度检测、差分隐私和合成数据生成技术，对疑似侵权素材进行标记、隔离与实时过滤，消除原始数据中

的版权表达残留。另一方面，利用人类反馈强化学习技术(RLHF)对开源大模型输出侵权内容予以否定性反馈。三是建立内部合规体系，开源模型使用者与服务模型部署者具有主体层面的同一性，从世界范围内的立法、案例及学说看，服务模型提供者的间接责任以未履行适当的注意义务为限，因此，部署严密且符合技术理性的合规体系，是开源模型使用者避免承担知识产权侵权责任的必要举措，应在组织层面明确法律、技术、业务团队的职责分工，实现“技术过滤 + 权责明示 + 数据合规”的全流程风控[6]。由于举证责任倒置，开源模型使用者还应当保存日志，为权利人提供核查工具。

(二) 数据处理与使用过程中的技术性合规

1) 强化元数据管理与溯源技术

尽管困难，但仍需投入资源建立强大的数据溯源系统。利用区块链、数字水印等技术，记录数据的来源、许可条款和流转路径。这不仅是合规的需要，也是数据治理和模型可解释性的内在要求。

2) 探索“署名”义务的履行创新

对于最终模型，可以探索在模型文档、官方网站或通过 API 返回的元数据中，以电子化的方式完整列出所有重要的数据贡献者。虽然这未必完全符合协议的“理想”要求，但可以作为证明已尽合理努力履行义务的证据。

(三) 法律解释与协议演进层面的应对

1) 推动 ODbL 协议的明确化与现代化

开源社区和组织应考虑发布官方的指导性文件或对协议进行修订，明确其在 AI 训练场景下的适用边界。开源协议是约束参与主体各项行为并推进大模型广域应用的核心工具[7]。例如，可以明确区分“用于训练的数据集”，可能被视为衍生数据库和“训练所得的模型”，可能不被视为衍生数据库，并为后者设定特定的署名方式。

2) 采用替代性许可协议

对于新的数据开放项目，如果希望避免 ODbL 在 AI 场景下的不确定性，可以考虑采用更为清晰的替代协议。例如，Creative Commons 系列协议中的 CC-BY 或 CCO 可能更适合于旨在最大限度促进 AI 创新的数据发布。

3) 使用合规的全流程防控

鉴于开源大模型输出管理阶段风险应对需求的特殊性，企业需从数据预处理、模型训练、生成控制及用户监管四个维度构建风险管理框架。探索将合规义务融入大模型算法的自动化治理范式。数据预处理环节，通过强化数据清洗与许可协议识别策略，构建代码与开源许可证元数据关联架构，利用大模型的自回归特性将许可声明作为代码片段的后续标记进行联合训练，增强模型对代码与协议关联性的语义理解。模型训练阶段，融入差分隐私训练机制降低代码记忆效应，或构建协议敏感性上下文学习模块，实现高风险协议内容的动态规避。生成控制环节，设置特殊的输出过滤范式，避免对开源协议保护内容的简单抑制，动态注入版权声明、原始许可证标识及衍生作品修改记录，同步满足 MIT 协议的署名要求与 GPL/LGPL 协议的传染性约束条件；模型部署前需开展系统性合规能力评估，重点验证其协议约束代码的识别与处置效能。用户监管层面，优化用户协议，禁止使用者实施未经授权的使用、修改及传播行为；建立对抗性提示工程防御机制，阻断用户通过诱导式输入获取违反开源协议内容。

6. 结语

人工智能训练数据池与 ODbL 协议的碰撞，是前沿技术发展对现有法律框架提出挑战的典型例证。ODbL 协议设计的初衷是保障数据开放的良性循环，但其法律条款在 AI 这种新型使用方式面前，显现出明显的不适应性和模糊性。核心问题在于，“衍生数据库”这一关键概念能否涵盖 AI 模型，目前在法律

上悬而未决，这给所有利用开放数据训练 AI 模型的企业带来了显著的不确定性与合规风险。本文通过系统分析指出，规避风险不能依靠侥幸心理，而必须通过综合治理。在操作层面，企业需加强数据源的许可协议管理，实施严格的数据分类与隔离，并利用技术手段强化溯源。在行业与法律层面，则需要推动开源社区对 ODbL 等协议进行适应性解释或修订，以适应技术发展的现实。从更宏观的视角看，这一问题的探讨也预示着，未来数据立法和许可协议的设计需要更具技术前瞻性，以在鼓励开放共享与保护商业创新之间找到一个可持续的平衡点。

参考文献

[1] 郭亚军, 徐苑茜, 梁艳丽, 等. 从 ChatGPT 到 DeepSeek: 生成式人工智能迭代对图书馆的影响[J]. 图书馆论坛, 2025, 45(7): 140-149.

[2] 喻玲, 邵滨. 开源社区知识产权治理模式及变革——基于 36 个开源社区使用协议的考察[J]. 科学学研究, 2024, 42(9): 1938-1945.

[3] 希瑟·米克. 商业开源: 开源软件许可证实用指南[M]. 刘伟, 译. 北京: 人民邮电出版社, 2021: 4.

[4] 朱鸿军, 王涛. 开源许可协议: 人工智能作品保护的另一种制度安排[J]. 探索与争鸣, 2025(12): 136-145+211.

[5] 王利明. 生成式人工智能侵权的归责原则与过错认定[J]. 中国法律评论, 2025(4): 15-30.

[6] 黄如花, 李楠. 开放数据的许可协议类型研究[J]. 图书馆, 2016(8): 16-21.

[7] 王丹. 开源大模型嵌入档案数据空间的基本架构、安全风险与规范路径[J]. 情报科学, 2025, 43(11): 20-26+117.