公有领域中大模型开源代码数据训练的惠益分享

张惠彬 许 蕾

摘 要：大语言模型的性能提升以海量数据训练为基础，开源代码数据是其重要的语料来源。开源以代码资源的开放共享为理念，以版权保护与协议授权为手段，在传统开源制度中，用户使用开源代码应当附带开源协议输出。而在大模型数据训练中，大模型的介入切开了开源协议的流动，代码数据的无法溯源和开源协议的不兼容使开源协议难以遵守，大模型数据训练的开源之困由此诞生，进而陷入版权保护与产业进步的两难困境。开源协议仅仅是实现技术普惠的工具，在生成式人工智能时代，大模型可以以一种更为开放的方式践行开源的理念，即走向明确的公有领域。在技术普惠理念的指引下，从开源许可协议走向代码数据开放是大模型开源代码数据训练的惠益分享之策。

关键词：大语言模型；开源许可协议；公有领域；技术普惠

中图分类号：D923.4;D922.17 文献标识码：A 文章编号：1673-5706（2024）02-0048-08

一、问题的缘起

2024 年 1 月 31 日，习近平总书记在中共中央政治局第十一次集体学习时强调，科技创新能够催生新产业、新模式、新动能，是发展新质生产力的核心要素。发展新质生产力，必须进一步全面深化改革，形成与之相适应的新型生产关系[1]。2023 年我国人工智能领域取得了持续创新的成果，尤其是生成式人工智能（AIGC）等大模型的迅猛发展。这种发展势头加速了人工智能与实体经济的深度融合，为各行各业带来了巨大的推动力。实践表明，我国在新质生产力的数据、算法和算力等要素的发展方面总体上领先于数字经济的生产关系。当前，我们应该加快推进数据要素治理的法律体系建设，特别是加强对人工智能创新进行监管的法律法规，为新质生产力的形成提供坚实的制度保障。

为了促进生成式人工智能健康发展和规范应用，我国在 2023 年 7 月出台了《生成式人工智能服务管理暂行办法》（以下简称《办法》）。《办法》强调，生成式人工智能服务提供者应当依法开展预训练、优化训练等训练数据处理活动，遵守以

基金项目：教育部重大攻关课题“我国海外利益保护体系构建研究”（22JZD015）；重庆市教委科学技术研究重点项目“生成式人工智能对著作权登记制度的挑战与应对研究”（KJZD-K202300304）；重庆市社会科学规划一般项目“美国经济制裁介入知识产权领域的影响与应对”（2023NDYB35）。

下规定：使用具有合法来源的数据和基础模型；涉及知识产权的，不得侵害他人依法享有的知识产权。生成式人工智能算法训练的数据分为两种，一种是公开的公共数据，一种是版权法保护的作品数据，而开源代码数据则是后者中的重要组成部分。基于开源代码数据的算法训练极大提升了AI大模型的性能，但同时也因开源协议的遵守问题容易引发纠纷。

回顾历史，开源是大模型发展的基础。1984年，麻省理工人工智能实验室的Richard Stallman为了重现软件界合作互助的团结精神，建立了自由软件基金会，启动了庞大的“自由软件联盟计划”。经过四十年的发展，开源模式凭借强大的资源汇集和协同创新优势，已经成为全球软件开发的主导模式[2]，96%的商业代码中包含开源代码，76%的商业代码库是由开源代码组成的[3]。开源代码为大语言模型的发展提供了丰富的语料库，但开源的自由和开放并不意味着免费和随意使用，开源代码亦是著作权法保护的作品，它的开放共享是以完成开源许可协议规定的条件为前提。生成式人工智能时代，开源打破了技术壁垒，促进了分散资源和算力的整合，为大语言模型的发展提供了助力，因而大量的开源代码被用以数据训练与内容生成，而开源许可协议却被置于一旁，诉讼也就因此产生。从理念上看，开源社区和大语言模型都以文本开放与技术进步为价值追求，两者之间并非有着不可调和的法律矛盾，无论是开源协议构建的开放许可生态，还是大语言模型的数据训练与代码生成，都是实现技术普惠的工具。为此，本文围绕大语言模型与开源协议的关系展开研究。

二、追本溯源：开源代码数据是大模型训练的基础

（一）开源代码与开源协议

开源理念的诞生是为了对抗商业代码的闭源模式，重现合作互助的软件共享精神。在20世纪五六十年代，提供软件的源代码是业内惯例，购买计算机的用户可以自由对软件进行修改，开发者之间亦是无偿共享[4]。到20世纪70年代，以微软为代表的商业化软件开发公司开始出现并不断壮大，他们认为这种共享软件的行为是对软件开发者积极性的打击，并提出了在终端销售中只提供目标程序而不提供源代码的闭源模式。1976年美国大幅修改著作权法，将“电脑程式”纳入法律规制的范畴，软件代码的闭源销售模式逐渐成为主流。该模式下，软件的源代码被开发商垄断式占有，除他们以外的用户和其他程序员都无权查看、改动和完善，这无疑不利于软件的进一步发展。1984年理查德·斯塔尔曼（Richard Stallman）为了重现软件界合作互助的团结精神，在一些企业的支持下建立了自由软件基金会，启动了庞大的“自由软件联盟计划”，也就是GNU项目。自由软件自开发之日就重点强调“自由”的理念。但是自由软件定义中“free”的含义非常模糊，除了“自由”之外还含有“免费”的意思[5]。实际上，自由软件本身并不抵制商业化，自由软件基金会明确表示“自由软件无关价格”。为了避免对自由软件含义的误解，以埃里克·雷蒙德（Eric Raymond）为代表的程序员提出了新的概念——开放源码软件（Open Sourse），简称为开源代码。1998年开放源代码促进会正式成立，“自由软件”也开启了向“开源软件”的转型之路。

开源许可协议是伴随开源软件发布并对其使用、修改和分发等作出规定的文本，用以防止代码垄断、确保自由分发[6]。开源的理念是放弃对代码作品的独占版权利益，将代码置于公有领域，以实现自由分发与使用，从商业闭源的“大教堂”走向自由开放的软件代码市集[7]，但公有领域的作品难免有被他人滥用甚至占为己有的风险。为了防止垄断，保障开源代码能够永久地实现自由的使用及修改发布，开源模式巧妙依靠版权法来构建起一个具有法律约束力的自由作品分发规则[8]。开源社区依托版权作品授权规则，通过制定统一的开源许可协议来实现授权许可。只要是附带开源许可协议公布的开源代码，一律具有普通许可使用的授权，任何人都可以依照许可协议直接对代码作品加以利用，而无需与作者再签订双方合同。自理查德·斯塔尔曼为GNU项目撰写GPL开源许可协议，随着开源运动的扩大，开源协议对开源软件的促进与保护作用愈发凸显，其种类也

在不断增加，当前由开放源代码促进会公开认可的许可协议已达 116 种。开源实现了智慧共享，提高了开发效率，已然成为软件的主流模式，更是推动新时代多产业技术发展的重要动力。

（二）基于开源代码的大模型数据训练

工业化、信息化、网络化的快速发展促进了大数据时代的到来，海量数据信息对自然语言处理技术提出了更高标准的要求，大语言模型应运而生。自 20 世纪 50 年代图灵测试提出以来，人们始终在探索机器处理语言智能的能力[9]。从最初基于概率计算的统计语言模型（SLM），到引入上下文语言分布概率的神经语言模型（NLM），再到通过大规模语料训练产生的预训练语言模型（PLMs）。随着算法算力的不断提高，语言模型的数据参数规模也不断扩大并带来模型性能的提升，当模型规模达到一定量级后，就实现了从量变到质变的跃升，在复杂任务的解决上凸显出令人惊奇的能力[10]，甚至开始具备上下文学习等新的能力。语言模型的发展就此进入大语言模型（LLM）阶段。大语言模型的诞生对人工智能造成了重大影响，特别是基于大语言模型的 ChatGPT 和 GPT-4 的问世更是掀起了相关研究的热潮，但现阶段大语言模型的潜在原理仍然没有得到很好的探索，因此也就无法对其进行准确的定义。许多学者主要依据模型参数量与所利用训练数据规模来界定与评估何为大语言模型[11]，虽然尚未确立一个被广泛认可的临界标准，但至少可以明确“模型参数量”和“训练数据规模”这两个核心要素，基于此，我们将大语言模型（LLM）定位为基于海量文本语料训练、包含至少数十亿级别参数的语言模型[12]。

大语言模型发展的关键是海量文本数据训练，这些数据中包括公开的公共数据和受版权法保护的作品数据，本文讨论的开源代码数据属于后者。基于大语言模型的生成式人工智能有着高效率学习和高质量输出的超凡优势，在包括代码生成的诸多领域发挥作用，但生成式人工智能不是空中楼阁，能够实现从 1 到无数的飞跃，却无法实现从无到有的创造，其训练过程中输入数据的质量和规模直接决定了生成内容的优劣和应用场景的多寡[13]。海量高质量的语料基础是 Chat GPT 技术突破的关键要素之一[14]。而在提供代码编辑和修改服务的生成式人工智能工具中，开源代码数据则是其最重要的训练文本来源。中国信通院发布的《开源生态白皮书》显示：开源软件对我国的渗透率达到 88.2%。新思科技公司 2022 年发布的《开源安全与风险分析报告》发现，开源成分在各行业代码库中的占比从 46% 增长到 83%，行业越新，比重越大，总计开源成分占被审代码库的 70% 左右。

三、现实困境：大模型数据训练与开源协议的矛盾

开源社区的代码数据可以成为人工智能训练的养料，人工智能的使用可以帮助开源代码的开放，两者应当是互利共赢的发展模式。但在实践中，大语言模型的使用又存在侵犯开源代码权利、阻碍开源社区发展的情况。著名的软件托管平台 GitHub 与人工智能研究实验室 OpenAI 就因使用了程序员储存在 GitHub 上的开源代码进行数据训练，并且在向用户输出内容中复制了相关源代码但未标明归属而面临纠纷。坐拥一亿用户的 GitHub 是目前世界上最大的软件开发平台，依靠 ChatGPT 一炮而红的 OpenAI 在人工智能领域亦有着举足轻重的地位，这场诉讼虽尚未定音，但已然引起开源与人工智能领域的热烈讨论，充分凸显了大模型开源代码数据训练中的开源之困。

（一）大模型的介入切开了开源协议流动

开源许可协议是开源生态的法律基础，也是开源产业可持续发展的制度性保障。基于开源许可协议条款的权利义务设计，开源代码设计者与使用者之间实现了双向联通。在传统版权模式下，著作权人在作品创作完成后依法享有独创性利益，除非具有法定许可、合理使用等法律上的依据，否则其他任何组织、个人要使用其作品都需要从权利人处获得授权许可，著作权人与使用者通过授权许可实现双向互动，以实现作品的传播与进一步利用。开源社区的这种模式并未突破著作权法的作品授权规则，只是在程序上进行了简化，通过制定统一的开源许可协议来实现授权许可。只要是附带开源许可协议公布的开源代码，一律具有普通许可使用的授权，任何人都可以依照许

可协议直接对代码作品加以利用，而无需与作者再签订双方合同。对于这些开源许可协议的法律性质，虽然法律并未明确界定，但在司法实践中始终默示承认其具有合同效力。2021 年 6 月，由广州知识产权法院审理的罗盒系列案件中，法院更是明确指出 GPL 开源许可协议具有合同性质，可认定为授权人与用户间订立的著作权协议，属于我国合同法的调整范围。开源社区采取“先授予软件版权，再提供许可证”的模式，既明确了开源软件享有著作权，又实现了代码开放共享的社区理念。

基于开源许可协议，代码数据从权利人流向使用者的同时，版权声明和许可协议声明也附带传递，使用者能够依此知晓其需要履行的义务，实现了两方主体之间的联动互通。但在生成式人工智能时代，开源代码数据的使用呈现出三方主体参与的情形。通过大语言模型训练过程，开源代码数据从代码设计方流向大模型；通过生成式人工智能的用户使用内容生成过程，开源代码数据再流向使用方。大语言模型在这个环节中不仅仅是作为一个中转平台存在，而是有着双重身份。一方面，大语言模型本身也是代码的使用者，从开源许可协议的规定来看，对开源代码的使用方式一般表现为复制、修改和发行行为。大语言模型使用代码的基本流程为：将开源代码复制到数据库进行训练，根据用户指令对数据库中的代码加以选择和修改并输出。在这个过程中，大语言模型往往在复制时或者复制后通过特定对比过滤机制筛除许可协议相关规定，构成了义务的不履行。另一方面，大语言模型是代码“传送”的中间商，但其在未获得授权的情况下，移除或更改开源代码附带的版权管理信息，不向终端用户提供代码原作者的任何属性，也不向终端用户提供关于其许可证的任何要求，致使用户在不知情下违反了开源许可协议中的义务要求。在大语言模型的开源代码纠纷中，不仅存在生成式人工智能自身违反许可协议规定，构成版权侵权的问题，同时其故意擅自删除、更改版权管理信息，并传播作品的版权管理信息或复制件的行为亦属于版权法所禁止的。大语言模型的介入，切断了开源协议的流动，造成了版权侵权的纠纷。

（二）输出的代码数据无法附带开源协议

大语言模型数据训练的介入中断了开源协议从代码设计者向用户的流动，开源协议与代码作品的分离使版权授权许可失去基础依据，而矛盾的根源在于大语言模型的训练发展离不开丰富的开源代码数据，但从现实层面又无法实现附带开源协议的输出。

一方面，代码数据无法溯源。大语言模型数据训练过程主要包括预训练环节和监督微调环节，通过无监督的预训练来获得模型的初始参数，并经过受监督的微调适应不同的下游任务[15]。预训练模型首先通过各种途径获取大规模的无标注数据，这种自监督训练策略可以简单理解为一种“完形填空”练习，大语言模型通过随机掩藏模型中的某些词再让模型预测被掩藏的词来完成预训练[16]。代码生成类大语言模型的预训练机制是在自然语言模型上进行的改进，这类模型不仅能够掩藏文本信息来完成训练，而且会在代码的数据流图中随机掩藏某些数据节点然后让模型去预测，基于足够大数据基数的预训练练习，模型对于无标注数据的理解能力不断提升，“完形填空”的正确率也越来越高，一个有着代码生成功能的大语言模型便初具雏形。经过预训练的模型具备了较强的语言生成能力，但缺乏对人类指令的理解。为了让大语言模型更加准确了解人类意图，需要在人类监督下对其进行微调。基本流程为：先形成一定数量的“人类表达—任务结果”的标注数据[17]，将这些人工标注过的“人类偏好数据”注入大语言模型进行优化训练，接着随机选取部分问题与大语言模型进行问答，将模型输出的内容打分排序，依据人类偏好进行定向反馈。循环进行上述训练，不断强化提升大语言模型对人类意图的理解，形成更加“类人”智能的模型机制。事实上，大语言模型并没有自己的“核心信念”，它的代码输出活动本质上是在玩填词游戏，以用户输入的提示为样本，在数据库中进行概率预测，将最大可能的结果作为答案输出。大语言模型进行的是最大概率的填词游戏，这就决定了它对输出的最佳答案无法进行单一性的溯源。在开源代码中，开源许可协议的相关文本一般位于代码文本相并

列的文件夹中，或者置于全部代码文本的首端，大语言模型的训练通常以语句或代码段为单位，因此在其数据库中，开源协议文本与开源代码文本往往是割裂存在的。如果要求大语言模型输出代码溯源至附带有开源协议的原始数据，就需要模型另行建立完整保存训练语料的数据库，以作对比文本使用，这种方案仅具有理论上的可行性，但从技术层面对模型效能的提升并非利事。

另一方面，开源协议互不兼容。大语言模型生成的代码是从整个数据库“拼凑”而来，不同的代码数据片段往往附带不同的开源协议，协议的不兼容性必然会影响用户的使用。大语言模型辅助代码开发的版权问题已经引起了企业的关注，开源代码识别技术也在不断进步，新思科技基于Black Duck平台开发了代码片段分析功能，可以将代码片段与所属的开源项目进行匹配，通过向用户披露代码所涉及的开源项目和许可协议来提醒规避侵权风险。这一技术在一定程度上解决了大语言模型难以进行代码溯源的困境，但问题仍然存在。代码识别技术的基本机理是将AI输出的代码片段与开源项目数据库进行相似对比，将类似开源项目附带开源协议输出，然而大语言模型输出的代码文本往往是不同开源项目的组合，附带的不同开源许可协议之间并非都具有兼容性。比如ApacheV2.0中含有专利中止和侵害保护条款，而GPLV2.0禁止专利许可，GPLV3.0的“强传染性”要求任何衍生作品都要基于GPLV3.0分发并提供相应的源代码。从用户的角度来看，使用大语言模型大多是为了完成一些基础但较为繁琐的简单代码工作，从而提高软件开发效率，相较于直接从不同开源社区中寻找代码，大语言模型仅仅是在检索与整理方面提供了便利，很难在代码质量上有所突破。如果大语言模型输出的代码文本附带有多个开源协议，用户在使用前就需对不同开源协议的不同条款进行理解，对协议之间兼容性进行判断，反而增加了用户使用的负担。如果大语言模型因代码附带开源协议的兼容性问题将困难转嫁用户，将诉讼的“达摩克里斯之剑”悬在其头顶，必然会失去用户群体。对于这一困境，微软在Copilot上做出了版权承诺：如果用户因使用Copilot或者微软AIGC服务产生的作品被控侵犯版权，微软将承担可能涉及的法律风险，版权巨头Adobe也承诺，为使用“萤火虫”AI的商业用户提供知识产权纠纷的保障。这样的“售后服务”或许能为这些基于大语言模型的生成式人工智能产品留住用户，但这一无奈之举也从侧面说明从技术上解决大语言模型数据训练与开源协议的矛盾的不可行性。

四、破局之维：探寻开源代码数据使用的惠益分享

大模型数据训练与开源协议之间的矛盾在于不遵守协议条款，但新技术发展必然会对传统制度规范造成一定冲击，一个更基本的问题是制度背后的目的在新的时代应当如何实现，开放源代码运动的互惠期望能否得到应有的重视。对于许多选择在各种“开放”许可证下分享代码作品的创作者来说，这种选择是他们为实现政治和道德理想而做出贡献的唯一途径，但是大模型训练提供了新的选择。人类不能无视开源代码开发者的互惠预期，机械性依据开源协议进行判断，落入工具主义的陷阱，必须从开源代码的技术普惠理念出发，在遵循权利人期待的前提下，探寻更利于人工智能产业发展的两全之策，那就是从法律层面明确大模型开源代码数据训练的合法性，使大模型据此生成的代码直接走向共有领域。

（一）开放之基：开源代码协议设计的普惠期望

技术普惠是开源的追求。开源理念诞生于对软件私有化的抗争，正是在软件代码自由分享、技术普惠的共同追求之下，开源社区得以成立并不断发展壮大。开源软件自开发之日就重点强调“自由”的理念，即开发者同意将源代码以共享的方式开放，在满足自由软件许可协议的前提下，鼓励用户之间互相复制，通过网络在线发布和自由传播。开源运动旨在利用开源软件的价值和分散的生产模型，为其社区和行业的问题寻找新的解决方法。在开源模式下，开发者处于一个开放共享的环境当中，以合作开发的方式提高代码质量与推广软件，大规模“人人生产”，能够更加完全、更加有效地利用人的技能、天赋和智力，

可以综合群体知识、群体能力、群体资源，完成的成果远远超过一个单独的个体所能完成的[18]。

而开源软件协议是以实现软件自由、消除著作权壁垒为目标形成的开放式、集体主义色彩的协议，该协议以协议主体让渡著作权中部分权利为代价，通过责任条款形式鼓励软件传播[19]。开源许可协议的设立与附带使用是为了避免使用者个人垄断，确保开源代码的永久自由，避免权利人放弃版权利益的代码重新被私有化。从理查德·斯塔尔曼构建开源理念，到今天呼喊信息时代的开放革命，资源的共享与技术的普惠才是目的，而制度设计只是实现目的的手段。在违反开源协议的版权纠纷中，决不能出现以开源开放之名行垄断自利之事的情形，只有明确这一点，才能使开源的理念在不同时代持续延伸，也只有基于这一点，才能寻求大语言模型与开源之间矛盾的解决方案。

开源协议是普惠的工具。在技术普惠的共同追求之下，开源协议和大语言模型都是实现目的的工具。无论是版权保护还是开源协议，都只是开源理念实现的工具。开源希望开放源代码能够不受限制地在公共领域自由分发与传播，但恐于权利人的让步成为他人谋利的嫁衣，因而巧妙地设计出一套 Copyleft 制度。这种独特的模式对代码技术的发展与代码数据的传播进行了助力，但并不意味着这种模式是不可突破、不可替代的。知识产权法的“私人定制”就如同在数据河流的上游疯狂建立权利的数字磨坊，但其存在并不必然代表着合理。开源社区在理念上有反抗传统知识产权法之处，但没能坚守“对知识产权制度批判”的初心，开源许可协议的天然垄断性对创新和市场竞争控制的负面影响已经显现[20]。人工智能大语言模型虽然改变了“版权保护 + 开源许可”的传统范式，但从技术普惠与代码共享的角度而言，开源的理念并不是被突破，反而是以一种更好的方式实现。在讨论开源协议能不能传递之前，应当明确开源协议是不是确有必要传递。开源社区与 GitHub 的这场纠纷恰时敲响了警钟，在人工智能大语言模型引领下的新技术时代，国家提倡的开放创新，不再是简单沿袭传统的开源规则，而是要在开源精神下找寻更适合的技术工具。

（二）惠益分享之策：从开源协议走向代码数据开放

大模型开源代码数据训练的合法性论证。开源协议通过“版权保护 + 开源许可”模式实现技术普惠，同时这也是获得法律保护的两个条件，但是现有关于大模型开源代码数据训练的争论大多聚焦于后一条件，即是否遵守开源协议，而忽视了前一条件。自由软件之父理查德·斯托曼之所以创造出 GPL 开源协议，目的是防止那些阻碍自由软件分发传播的行为，为源代码共享计划带来福音。在关注开源许可协议的同时，我们必须基于版权保护的基础。换言之，源代码的版权保护是开源协议授权许可的前提。但是，在大语言模型数据训练与代码生成的活动中，这个前提不再存在。从立法论的角度，人工智能自主生成物不符合自然权利学说、激励说和投资说等版权正当性理论要求，同时因可通过收取服务费等方式获得激励而缺少版权保护的必要性[21]。从主体论的角度，著作权法以鼓励创作为目的，只有人才能理解和利用著作权法的激励机制，因此只有人的创作成果才能作为作品受到著作权法的保护[22]。美国版权局认定人工智能软件 Midjourney 生成的图像不受版权保护，在我国，“腾讯案”“菲林案”以及北京互联网法院一审的“AI 图片侵权第一案”也都将以人类作者为中心的理念贯彻于作品构成要件判断标准的始终，即使认定对 AI 生成内容加以保护，也是对作品中人类独创性的认可。无论是学理研究，还是司法实践，都未认可人工智能生成物的可版权性。如果脱离版权保护，大语言模型生成的代码文本不再属于作品，不再具有独占性利益，而是落入公有领域，成为社会公众可以使用与分发的“自由代码”，而不会被任何主体垄断。

只要确保大模型开源代码数据训练后生成的内容直接落入公有领域，就能够从法律层面避免垄断的发生，进而确保代码数据能够被普惠性使用。基于此，在大模型开源代码数据训练活动中，开源协议既没有存在的必要，也失去了存在的基础，两者的纠纷也能从根源上化解，破解版权保护与产业发展相冲突的僵局。跳出传统开源规则的束缚，大语言模型以更简便的使用、更高效的回应、

更自由的分发，向我们展现了开源精神实现的不同方式，大语言模型作为新的工具，将代码真正置于公有领域。从开源协议到大语言模型的工具变迁，不仅理清了开源理念与技术发展共存之道，也为在人工智能带来的新技术革命中，新技术与旧制度的关系梳理提供了新的思考之路：拨开技术迷雾，跳出工具掣肘，以共同理念指引寻找出路。

大模型开源代码数据使用的规范性要求。代码数据走向开放仅仅意味着大语言模型使用开源代码数据可以不受开源协议的约束，而非完全自由地使用，大模型代码数据训练的过程中，仍应当遵守一定的规范。

首先，从法律层面声明大模型开源代码数据训练活动的合法性与生成内容的公共性。在大模型的迅速发展与广泛应用推动下，人工智能相关法律规范不断推出，起着重要的引导和规范作用，明确大模型技术的研发、应用和管理的标准和原则，促进人工智能技术的健康发展。一方面，在大模型规范相关文件中表明大模型开源代码数据训练活动的合法性与生成内容的公共性，为大模型利用开源代码数据进行训练提供法律层面的依据，从根源上为大模型数据训练的开源纠纷解围，推动大模型技术发展与人工智能产业进步。另一方面，在相关司法案件中表明立场与强化说理，明确大模型开源代码数据训练不构成对开源协议的违反，并充分阐明开源协议的工具性质与大模型训练的开放属性，帮助社会形成大模型开源代码数据训练生成内容属于公有领域的基本意识。

其次，在技术层面标明大模型开源代码数据训练的数据来源与禁止垄断要求。虽然大模型无法对其生成代码内容逐一溯源并标明来源，但由于在大模型训练过程中，开源代码数据在数量上和质量上都发挥着不可替代的作用，因此大模型在面向用户生成特定代码文本时，应当向用户声明：生成内容中可能含有与开源代码数据相同或相似的部分。这里的声明并非是为了让大模型用户不得再以任何方式使用这些代码，也不是转嫁责任与风险，让用户自行对比寻找相关代码附带的协议文本并遵守，而是为了结合前述法律规范，进一步提醒用户不得进行垄断性使用。

最后，从诉讼层面表明大模型开源代码数据训练生成内容的不得垄断性。我国著作权采取的是版权自动取得原则，著作权的取得无需进行登记公示，而是自动享有。因此为了确保大模型开源代码数据训练生成内容持久性存在于公有领域，必须在诉讼程序上加以规范。如若有主体提起代码的计算机软件作品著作权侵权诉讼时，被告方可以以相关代码为开源代码作为抗辩事由。这样是否会有加重被告举证责任的嫌疑？在一般类型的著作权诉讼中，当原告提出版权侵权诉讼时，法律亦是要求原告进行侵权举证，而作品进入共有领域作为免责事由则是由被告方举证，因此在诉讼程序上表明被告可以以相关代码为开源代码作免责抗辩符合现有法律规范。

五、结语

新质生产力与 AIGC 之间存在着密切的关系。AIGC 作为人工智能领域的一种重要技术手段，具备生成文本、图像、音频等能力，能够自主进行创作和创新。AIGC 的发展不仅推动了数字经济的蓬勃发展，也为各行业赋能，促进了经济的创新和增长。从公共利益保护工具变迁的角度为大模型数据训练的开源纠纷解围，并非是对人工智能技术发展的一味偏袒，而是基于现有人工智能宽松监管的政策与人工智能工具属性的界定，对大模型使用开源代码进行数据训练的合理性进行正名。新技术对旧制度的冲突是不可避免的，人工智能时代已经到来，生产生活方式都在发生翻天覆地的变化，积极思考与主动作为方是应对之道。开源理念以代码共享、技术进步为目标，大模型正在以更好的方式实现它，技术普惠与工具变迁分别从理论与制度上破解了协议遵守的开源之困。此外，站在开源巨人的肩膀上眺望世界的 AI 大模型又在成为新的巨人。基于开源代码进行数据训练的大语言模型正在进一步开源，2023 年 7 月，Mate 和微软深度合作，推出开源大语言模型 Llama 2，同年 9 月蚂蚁集团正式开源代码生成大模型 CodeFuse。开源的力量在于代码的使用，开源的追求在于技术的进步，无论在什么时代，开源都不应成为新技术发展的阻碍，人工智能与开源应当以互助谋求共存、以共存实现普惠。

参考文献：

[1] 习近平在中共中央政治局第十一次集体学习时强调 加快发展新质生产力 扎实推进高质量发展 [EB/OL].（2024-02-01）[2024-03-01].http://politics.people.com.cn/n1/2024/0201/c1024-40171157.html.

[2] 何婷，徐峰 . 国外人工智能开源生态运营模式剖析 [J]. 全球科技经济瞭望，2022，37，（1）：55-63.

[3] 新思科技 .2023 年开源安全和风险分析报告 [R/OL].（2023-03-03）.https://www.synopsys.com/content/dam/synopsys/china/software-integrity/reports/rep-ossra-2023-ch.pdf.

[4] Richard Stallman.Why Software Should Not Have Owners[EB/OL].（1994）[2024-01-01].https://www.gnu.org/philosophy/why-free.en.html.

[5] 张平，马骁 . 共享智慧——开源软件知识产权问题解析 [M]. 北京：北京大学出版社，2005.

[6][7] RAYMOND E S. 大教堂与集市 [M]. 北京：机械工业出版社，2014.

[8] 何炼红 . 从 Copyright 到 Copyleft: 作者观念的反思与超越 [J]. 甘肃社会科学，2005，（5）：61-67.

[9][12] Zhao W X, Zhou K, Li J, et al. A survey of large language modelsl[EB/OL]. arXiv preprint arXiv:2303.18223，2023.

[10] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus.Emergent abilities of large language models[J].CoRR, vol. abs/2206.07682, 2022.

[11] 文森，钱力，胡懋地，等 . 基于大语言模型的问答技术研究进展综述 [J/OL]. 数据分析与知识发现，1-17[2023-12-21].

[13] 顾男飞，方舟之 .ChatGPT 等生成式人工智能使用作品的合理边界与侵权规制 [J]. 数字图书馆论坛，2023，19，（7）：1-8.

[14] 钱力，刘熠，张智雄等 .ChatGPT 的技术基础分析 [J]. 数据分析与知识发现，2023，7(3)：6-15.

[15] 郭春镇 . 生成式 AI 的融贯性法律治理——以生成式预训练模型（GPT）为例 [J]. 现代法学，2023，45（3）：88-107.

[16] 杨泽洲，陈思榕，高翠芸等 . 基于深度学习的代码生成方法研究进展 [J/OL]. 软件学报，1-25[2024-01-01].https://doi.org/10.13328/j.cnki.jos.006981.

[17] 朱光辉，王喜文 .ChatGPT 的运行模式、关键技术及未来图景 [J]. 新疆师范大学学报（哲学社会科学版），2023，44，（4）：113-122.

[18] 齐佳音，张国锋，王伟 . 开源数字经济的创新逻辑：大数据合作资产视角 [J]. 北京交通大学学报（社会科学版），2021，20（3）：37-49.

[19] 杨守晶 . 开源软件权利性质及保护路径探索——从“2022 开源软件保护第一案”切入 [J]. 理论界，2023，（6）：65-71.

[20] 张平 . 开源规则：案例、许可证及开源组织 [M]. 北京：知识产权出版社，2022.

[21] 张金平 . 论人工智能生成物可版权性及侵权责任承担 [J]. 南京社会科学，2023，（10）：77-89.

[22] 王迁 . 再论人工智能生成的内容在著作权法中的定性 [J]. 政法论坛，2023，41（4）：16-33.

作者：张惠彬，西南政法大学知识产权研究院副教授、博士生导师

许　蕾，西南政法大学知识产权法专业硕士研究生

责任编辑：钟晓媚