

引用格式: 温馨, 张超, 郭锐, 等. 推动我国大模型开源创新生态建设的挑战与建议. 中国科学院院刊, 2024, 39(8): 1313-1326, doi: 10.16418/j.issn.1000-3045.20231207004.

Wen X, Zhang C, Guo R, et al. Challenges and recommendations for building open source innovation ecosystem for large-models in China. Bulletin of Chinese Academy of Sciences, 2024, 39(8): 1313-1326, doi: 10.16418/j.issn.1000-3045.20231207004. (in Chinese)

推动我国大模型 开源创新生态建设的挑战与建议

温馨¹ 张超^{2*} 郭锐³ 陈凯华³ 冯泽¹ 朱其罡^{4,5}

1 中国科学院科技战略咨询研究院 北京 100190

2 上海交通大学 国家战略研究院 上海 200052

3 中国科学院大学 公共政策与管理学院 北京 100049

4 上海开源信息技术协会 上海 201620

5 对外经济贸易大学 全球开源协作研究中心 北京 100029

摘要 解决当前制约我国大模型产业发展的科技创新问题, 推动我国大模型产业持续繁荣发展和国际竞争力提升, 需要构建大模型开源创新生态。文章分析了开源创新生态对我国大模型发展的重要意义, 在回顾开源创新生态构建国际经验的基础上, 进一步剖析了我国大模型开源创新生态构建面临的问题和挑战, 并提出针对性建议。文章发现, 我国大模型开源创新生态面临技术能力制约生态形成、数据算力明显限制技术发展、创新主体无序竞争制约整体发展速度、开源支持体系建设水平较低、系统协同政策架构设计缺失等问题, 阻碍了大模型快速发展和竞争力提升。文章提出需加强顶层设计统筹协调发展、打造共享的大模型研发基础体系、强化全产业链开源开放体系建设、完善大模型开源创新体系治理等对策建议。

关键词 大模型发展, 开源创新生态, 对策建议

DOI 10.16418/j.issn.1000-3045.20231207004

CSTR 32128.14.CASbulletin.20231207004

大模型的涌现和同质化能力不仅将大幅提升人类的认知效率^[1], 还将引发经济、社会、文化等领域的

变革与重塑。世界主要国家争相加快推进大模型发展, 探索大模型发展的有效路径成为当前关注的焦

*通信作者

资助项目: 研究阐释党的二十大精神国家社会科学基金重大项目 (23ZDA060)

修改稿收到日期: 2024年7月10日; 预出版日期: 2024年7月11日

点^①。美国大模型开源创新生态的繁荣是其技术和产业发展始终走在前列的重要原因。一方面,大量开源的基础大模型层出不穷,不断推动底层技术性能的进步。例如,以开放式大语言预训练模型 OPT、GPT-NeoX-20B 等为代表的早期开源大模型的推出促进了开源社区对大模型的研究,美国 OpenAI 公司推出的 GPT 大模型的早期版本也完全开源。开源情况下,研发者能直接接触具有前沿性能的大模型,通过对已有开源大模型进行微调或者采用更大、更高质量数据集及更大规模模型参数创建性能更优的基础大模型,推动开源大模型技术性能快速进步。另一方面,以开源大模型为基础的开源应用不断出现,推动大模型产业的壮大。以 AI (人工智能) 绘画生成工具 Stable Diffusion 为代表的开源大模型形成了广泛的用户社区,衍生出极具多样性的应用场景,打开了大模型产业应用的想象空间。

与之相比,尽管我国部分大模型性能突出,但大模型上下游产业链各个环节缺乏协同,存在竞争无序和资源浪费现象。一方面,存在大量未开源的低质量大模型,导致低水平重复建设,难以真正推动我国大模型的发展;另一方面,大模型上游涉及的数据、算力,以及下游涉及的应用,均未能建立起真正的开源开放生态,阻碍了我国大模型产业的发展。这一状态将影响我国大模型产业的可持续发展,难以保障我国科技安全和产业链安全。

经验表明,开源创新生态能帮助汇聚全球开发者智慧以推动大模型技术进步,并激发社会创新活力加快大模型应用落地,能够凭借开源开放这一全球公认的突破科技垄断或制约的有力手段推动我国大模型及相关产业发展。然而,现有研究缺乏对大模型开源创

新生态的关注。本文从上游供应生态、下游应用生态和治理协调生态 3 个维度回顾开源创新生态构建的相关经验;从关系到大模型性能的底层算法、数据和算力维度,大模型下游产业生态搭建现状,大模型开源治理体系,以及政府系统协同政策推动方面,分析目前我国大模型开源创新生态构建存在的问题;在此基础上,提出构建开源创新生态推动大模型产业发展的相关对策建议。

1 开源创新生态对发展我国大模型的重要意义

大模型是指包含超大规模参数(通常在 10 亿个以上)的深度学习或机器学习模型,具有基础资源门槛高、产业集群效应强和潜在垄断性大等特点,后发企业难以快速形成行业积累实现追赶。开发贡献者、行业开源者、开源使用者等多元创新主体基于开放、协作和共享理念^[2],围绕数字基础设施构建协同创新和价值共创的开源创新生态^[3,4],有助于整合资源降低大模型研发成本,汇聚众智促进大模型技术迭代演进,形成相对竞争优势,从而有效推动大模型的发展与赶超。

1.1 整合底层基础资源,降低行业研发成本

大模型往往需要大量的训练数据、多种不同的学习任务及强大的计算资源支撑,致使训练成本巨大(例如,GPT-3 的训练据估计花费超过 4 600 万美元^②)。开源创新生态一方面能够促进基础数据资源的自由流动和高速聚集整合^[5],从顶层设计上扩大数据规模、提高数据质量和多样性,加强中文数据的标准化集成和持续积累优化,为大模型算法和技术研发提供数据保障^[6];另一方面可以提供基础的大模型算法技术并

① 北京智源人工智能研究院院长黄铁军:大模型在未来产业生态中处于基础性的核心位置.(2023-07-18)[2024-06-21]. <https://www.cena.com.cn/intelligence/20230718/120856.html>.

② OpenAI's GPT-3 language model: A technical overview. (2023-07-03)[2024-06-24]. <https://lambdalabs.com/blog/demystifying-gpt-3>.

促进算力基础设施的共建共用，以低成本的开放协作模式推动开发者充分探索参数、数据和算力组合情况下的性能表现，推动大模型整体的改进创新。由此，开源创新生态能够通过数据共享、算法开源、算力基础设施共建共用等方式，解决大模型研发和应用中单一机构难以完全满足数据、算法和算力资源要求的问题，从而降低企业乃至全社会商业化大模型的成本。可见，开源创新生态有助于打破垄断、降低大模型技术研发和优化的竞争壁垒，提高大模型数据和算力等基础设施的使用效率，加速推动我国大模型技术的创新发展及快速应用。

1.2 推动技术透明可信，促进技术迭代创新

大模型高昂的研发成本限制了学术界、非营利组织和较小规模工业实验室研究人员对大模型的研究和访问^[7]；不仅如此，闭源的大模型研发过程大幅降低了技术透明度和可信度，难以汇聚社会多方力量深化对大模型技术相关道德伦理风险的认知，进而阻碍大模型技术在各行业中的落地应用。大模型开源创新生态能降低各方潜在参与者参与大模型研究的难度，使得研究者更好理解大模型工作原理，提升社会对大模型应用接受度。同时，大模型的发展具有较强的产业集群效应（图1），开源创新生态有助于数据、算法和算力全方位协同，供应商、从业人员、平台、服务、数据和生产有效结合，加快大模型在各个产业中的应用，促进从模型层、中间层到应用层的多元主体价值共创。开源开放有助于建立社会对大模型技术的信任，推动不同级别大模型在各个行业的应用，而通过广泛应用场景积累的技术需求和技术问题将反哺大模型技术本身，推动大模型技术迭代发展。

1.3 以非对称竞争优势，打破潜在行业垄断

开源开放是全球公认的突破科技垄断或制约的有力手段，推动大模型开源创新生态建设不仅将为我国大模型技术提供新的发展机遇，还有望推动我国大模型产业出海，打破潜在行业垄断，化被动为主动。

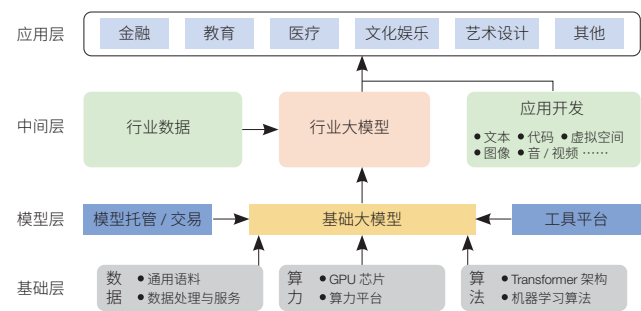


图1 大模型产业创新生态图谱
Figure 1 Innovation ecosystem map of large-scale model industry

资料来源：作者根据中国信息通信研究院《2023 大模型和AIGC 产业图谱》和北京市科学技术委员会、中关村科技园区管理委员会《北京市人工智能行业大模型创新应用白皮书（2023 年）》整理绘制

Data source: The authors have compiled and drawn this based on 2023 Large-Scale Model and AIGC Industry Atlas by the China Academy of Information and Communications Technology and the Beijing Artificial Intelligence Industry Large-Scale Model Innovation Application White Paper (2023) by the Beijing Municipal Science and Technology Commission and the Zhongguancun Science and Technology Park Management Committee

“微软 Windows+OpenAI 大模型+英伟达 GPU”通过强强联合绑定形成新的垄断生态，阻碍我国信创产业发展、威胁我国信创产业的科技安全和产业链安全。大模型开源创新生态能充分发挥我国在开源芯片等领域的技术优势，并通过集中攻关开辟新赛道形成非对称竞争优势。同时，推动我国大模型开源创新生态在全球大模型生态中占据一席之地，可为我国大模型技术在其他国家的应用提供良好契机。这能够打破国外大模型的潜在垄断生态，摆脱对欧美科技基于封闭知识产权的“非对称依赖”。既往发展经验表明，构建开源创新生态不仅能推动上下游相关产业健康有序协同发展，还能掌握一定技术发展路线话语权和主导权，使我国软件产业牢牢嵌套在国际整体生态之中，打破制约垄断。

2 构建开源创新生态的国际经验

开源运动从软件代码的公开协作开始,其开放共享的理念逐步扩散到计算机及相关产业的方方面面。越来越多来自全球的个人开发者和组织积极投身到开源运动中,数十年间国际上围绕开源逐步构建起稳固完善的上游供应生态、丰富多元的下游应用生态和公开有效的治理协调生态,其发展经验值得借鉴以构建我国大模型开源创新生态。

2.1 构建稳固完善的开源上游供应生态

上游供应生态的发展为开源项目的技术进步和持续创新奠定了基础。

(1) 支持开发者的开发工具和资源是上游供应生态的关键组成部分。① 开源项目可以为开发者提供友好的协作工具、文档和教育资源,以帮助他们理解和使用项目,提高开发效率并确保代码质量。在国际大模型开源过程中,这些开发工具和资源也被大量采用。例如,开源分布式版本控制系统 Git 为开发者提供了管理代码版本、协作开发和代码审查等功能,其广泛应用使得开发者能够更好地管理和追踪代码的变更,同时也有助于团队间的协作和合作。② 集成开发环境 (IDE) 和编程语言工具链等开发工具为开发者提供了高效的编写环境, Visual Studio Code、Eclipse、PyCharm 等开放的集成开发环境提供了丰富的功能和插件生态系统,使得开发者能够高效地编写、测试和调试代码。

(2) 支持开发者的数据是上游供应生态的关键一环。作为软件开发的重要底座,数据对应用性能训练的提升至关重要。开放的数据集不仅有利于构建公开透明的协作环境,同时能大幅降低技术开发前期成本及开发门槛,推动技术进步。目标检测、自动驾驶、人脸识别、自然语言处理、文本监测、医疗等方向均有大量经典开源数据集,例如人脸识别领域的 YouTube Face Database 包含 1 595 个不同人的 3 425 个

视频,总计 671.41 GB 数据,能够帮助训练优化人脸识别算法,减少开发人员在技术早期开发过程中遇到的困难。这些经典开源数据集也是大模型产生之初可靠的数据来源。

2.2 打造丰富多元的开源下游应用生态

下游应用生态包括开源软件的应用和集成,以及相关的商业生态系统。丰富多元的下游应用生态能吸引更多开发者和企业使用、扩展和创造基于开源项目的应用,促进相关产业的繁荣发展。以往的开源下游应用生态构建经验值得我国在打造大模型开源下游应用生态过程中学习。

(1) 广泛的用户和开发者参与,从不同的角度和需求出发为软件贡献代码、提供反馈并解决问题,从而推动软件本身的发展和改进。例如,Android 移动操作系统的成功很大程度上得益于其拥有丰富多样的下游应用。开发者可以通过使用 Android 开发工具包 (SDK) 创建应用程序,并通过 Google Play 商店这一应用市场将大量涵盖各种领域和需求的应用程序分发给用户。由此,Android 打造的多元下游应用生态为用户提供了广泛的选择,这种繁荣的应用生态系统吸引了全球范围内的开发者和企业,推动了 Android 平台的发展和创新,促进 Android 系统产业整体的发展。又如,OpenAI 也开放其大模型应用程序接口 (API),鼓励其他开发者将其大模型服务集成进其应用产品中,充分开发下游应用生态。

(2) 通过专门的支持机构或社区来提供技术支持、文档、培训和社区管理等服务。这可以帮助用户和开发者更好地理解和使用开源软件,并解决在实际应用中遇到的问题。例如,开源机器学习框架 TensorFlow 和 PyTorch 都有庞大的社区支持和专门的支持机构。这些支持机构提供了官方文档、教程、示例代码等资源,帮助用户和开发者学习和使用这些框架。同时,还通过举办培训课程、开发者大会等活动,促进用户和开发者之间的交流和合作。

(3) 发展基于开源软件的下游商业生态系统。开源软件商业生态系统的核心在于开源软件的产品和服务提供商，他们在开源软件的基础上通过提供定制化的解决方案、额外高级功能、代码托管或整合、搭建并运营插件市场、提供培训和咨询等运维服务等模式（表1）来谋求商业回报。经验表明，开源商业化有助于开源产出成果发挥价值，帮助其实现“价值创造—价值实现—价值分配”的合理闭环。形成有效商业模式的下游开源商业生态系统不仅对开源项目本身的健康可持续发展具有重要作用，还能促进同类技术的持续创新和市场竞争力^[8]。美国大模型领域也积极探索开源商业化模式，意图构建起繁荣可持续的开源大模型下游商业生态。例如，美国 Stability AI 公司通过开发开源大模型 Stable Diffusion 的商用版本，为客户提供定制拓展服务来促进大模型的应用。

2.3 培育公开有效的开源治理协调生态

开源治理协调生态涉及开源项目的决策、管理和

社区参与等方面，开源治理协调生态的健康发展对于项目的长期稳定和社区的繁荣至关重要。主要包括以下3个方面。

(1) 公开透明的决策流程和沟通机制能使所有人了解技术路线决策细节，从而对项目建立长期的信任，促进参与和合作。例如，在美国发布的 Linux 内核社区采用邮件列表作为主要沟通方式，由此使得项目成员能随时了解项目发展方向和最新动态；通过一系列公开的解释文档详细说明了技术开发相关的决策执行机制和协作模式。所有决策流程和相关信息公开可追溯增强了社区的信任感，鼓励更多人参与到开源项目贡献中，从而促进了项目的健康长久发展。

(2) 建立有效的冲突解决机制也是构建成功开源治理协调生态中的关键一环。例如，位于美国的云原生计算基金会（CNCF）下设技术监督委员会来协调组件之间兼容性冲突，其技术监督委员会成员通过选举产生，其成员来自供应商、最终用户等多个方面，

表1 常见开源商业模式
Table 1 Common open source business models

开源商业模式	具体做法	典型案例
开放核心(Open-Core)商业软件模式	将核心或功能受限的软件版本开源,同时将附加功能集成在付费版本的软件产品中;付费版本软件保证了使用的稳定性、易用性和安全性	美国 Red Hat 公司面向企业用户发布的 Linux 操作系统软件 Red Hat Enterprise Linux(RHEL)、美国甲骨文公司的 MySQL 数据库软件、美国 Cloudera 公司的 Cloudera Data Platform 软件
软件即服务(SaaS)模式/托管服务模式	软件开源,向客户销售在线账户和服务器访问权限获得收益;常出现在数据领域	美国 Spark 公司 Databricks 云服务、美国 MongoDB Inc 公司 MongoDB 云服务
基于服务的模式	通过提供基于开源软件的使用培训和咨询等服务获得收益	美国 Red Hat 公司、德国 SUSE 公司
开源硬件模式	硬件设计文件以开源形式发布,通过销售硬件设备、提供硬件相关的增值服务等获得收益	起源于意大利的 Arduino 开源电子平台
广告和赞助模式	在开源软件中嵌入广告或寻求赞助商为项目提供资金支持;适合应用广泛的开源项目	美国谷歌公司 Chrome 浏览器
应用/插件市场模式	软件产品开源,有大量在该软件环境上开发的应用或插件,通过收取用户在使用相关应用及插件过程中的付费抽成获得收益	美国谷歌公司 Android 操作系统、美国 Mozilla 公司 Firefox 浏览器

能充分代表开源社区内各方的利益,有助于维护社区的和谐与稳定,并推动项目的进展。

(3) 良好有效的开源制度设计对开源参与者长期可持续参与到开源项目贡献之中非常重要。其中,开源许可证是开源制度设计中的关键,它决定了如何使用、修改和分发开源软件。选择符合项目目标和社区需求的开源许可证能保护贡献者的权益、推动创新和知识共享。常见的开源许可证包括 MIT 许可证、Apache 许可证和 GNU 通用公共许可证等。阿联酋开发的 Falcon 大模型就采用 Apache-2.0 许可证,其成为第一个可以免费商用的开源大模型,这将促进其模型在科研及商业化中的应用。

3 我国大模型开源创新生态建设面临的挑战

我国开源创新生态尚处于初步探索阶段,社会对开源认知不够,且缺乏建设开源创新生态的经验及配套完善的体制机制。大模型作为新兴技术和产业,其开源创新生态的建设将面临更大的挑战。一方面,我国大模型底层基础研究能力相对薄弱,数据和算力基础制约大模型性能提升;另一方面,大模型产业内各类创新主体间未形成有效协同,产业内无序竞争引发乱象丛生。这些挑战不仅限制了我国大模型进一步的发展应用,更阻碍了我国大模型参与国际竞争,在全球范围内影响力的辐射扩散。

3.1 系统协同政策架构设计缺失

尽管我国在国家层面(表2)及各省级地方政府层面(表3)均高度重视大模型发展,从算力支持、场景开放、技术突破、产品生态等多方面积极出台大模型产业发展措施,鼓励大模型应用落地。然而,我国现有政策系统性不足,主要集中在大模型本身,对大模型产业链条的其他环节关注不够,尤其是数字公

共产品制度、开源商业化制度等适应开源创新生态的体制机制建设尚不健全,导致产业链上下游协同不足,难以满足建设大模型开源创新生态需求。同时,各部门间缺乏有效信息互通、各地政府间技术要素不流动,政策趋同致使无法形成合力推动人工智能大模型产业整体发展,未充分发挥出对实体经济的赋能作用。多个部门同时负有促进大模型应用落地、产业繁荣的职责,部门职能存在重叠导致政策间的协调不足,无法充分发挥政策指导促进的作用。

3.2 技术能力制约生态形成

我国大模型整体技术实力与国外头部企业差距明显,在算法、人才和科研投入方面与国外头部企业差距较大,同时部分关键核心技术尚未突破,尚未形成促进国产大模型发展的支撑基础。根据权威测评榜单 Super CLUE^③的评测^[9],截至2023年10月,GPT-4、Claude2 和 GPT-3.5 在基础模型领域综合排名前3位(图2),我国基础模型在计算、代码、生成与创作、上下文对话、角色扮演、工具使用方面得分与 GPT-4 的相应指标相差10分以上,部分指标接近 GPT-3.5,仅在中文知识题目方面明显优于国际模型。大模型厂商技术上的基本同源导致现阶段较为相似的模型性能,尚未形成显著技术性能优势,同质化严重影响了下游应用生态的构建。同时,我国基础模型缺乏原创性,版本迭代和技术演进高度依赖国外进展^[10]。特别是我国目前广泛应用的主流模型大多基于 Transformer 架构,而非我国自主研发的架构,在一定程度上制约了我国国产大模型自主创新生态的形成。

3.3 数据算力显著限制技术发展

OpenAI、Google 人工智能研究团队相继证明,人工智能模型的性能随着模型规模的指数级上升而线性增长^[11],并在模型规模达到某个阈值时对某些问题的

③ SuperCLUE 聚焦于通用大模型的综合性测评,已形成相对完善的评价大模型能力的测评框架。测评由多轮开放问题 SuperCLUE-OPEN 和三大能力客观题 SuperCLUE-OPT 两部分结果组成。2023年10月测评集共3 754题,其中606道多轮简答题(OPEN),3 148道客观选择题(OPT)。

表2 我国大模型的国家层面相关政策

Table 2 China’s national level policies related to large-scale models

政策	发布时间	大模型相关内容
《“数据要素x”三年行动计划（2024—2026年）（征求意见稿）》	2024年1月	以科学数据支持大模型开发，深入挖掘包含科技文献在内的各类科学数据，通过细粒度的知识抽取，构建科学知识资源底座，建设高质量语料稿库和基础科学数据集，支持开展通用人工智能大模型和垂直领域人工智能大模型训练
《关于加快推进视听电子产业高质量发展的指导意见》	2023年12月	支持骨干企业做大做强，支持人工智能企业研发视听应用大模型
《电子信息制造业2023—2024年稳增长行动方案》	2023年8月	鼓励加大数据基础设施和人工智能基础设施建设，满足人工智能、大模型应用需求
《生成式人工智能服务管理暂行办法》	2023年7月	鼓励生成式人工智能算法、框架、芯片及配套软件平台等基础技术的自主创新，平等互利开展国际交流与合作，参与生成式人工智能相关国际规则制定；推动生成式人工智能基础设施和公共训练数据资源平台建设；促进算力资源协同共享，提升算力资源利用效能；推动公共数据分类分级有序开放，扩展高质量的公共训练数据资源；鼓励采用安全可信的芯片、软件、工具、算力和数据资源
《生成式人工智能服务管理办（征求意见稿）》	2023年4月	首次明确了生成式人工智能“提供者”内容生产、数据保护、隐私安全等方面的法定责任及法律依据，确立人工智能产品的安全评估规定及管理办法。
《关于支持建设新一代人工智能示范应用场景的通知》	2022年8月	坚持面向世界科技前沿、面向经济主战场、面向国家重大需求、面向人民生命健康，充分发挥人工智能赋能经济社会发展的作用，围绕构建全链条、全过程的人工智能行业应用生态，支持一批基础较好的人工智能应用场景，加强研发上下游配合与新技术集成，打造形成一批可复制、可推广的标杆型示范应用场景

资料来源：中国政府网、北京市人民政府网、人民网、《北京市人工智能行业大模型创新应用白皮书（2023年）》
Data source: www.gov.cn; www.beijing.gov.cn; www.people.com.cn; Beijing Artificial Intelligence Industry Large-Scale Model Innovation Application White Paper (2023)

处理性能突增，具备涌现能力^[12]。这一现象凸显数据和算力在提升大模型性能中的重要意义。①在数据方面，尽管我国已有部分中文开源数据集，但从数据规模和语料质量上均与海外有较大差距，且部分内容较为陈旧，高质量全面完整可信的开放中文数据集匮乏。同时，我国尚未建立有效的数据流通规则和数据供需对接机制，企业获取数据资源的成本极高。数据产品供应链尚不完善严重制约了我国大模型的训练表现^[13]。②在算力方面，中国、美国在全球算力规模中的份额分别为33%、34%，其中以图形处理器（GPU）和神经网络处理器（NPU）为主的智能算力规模方面中国高于美国，分别为39%、31%，具备发展大模型产业的有利基础^[14]。然而，现阶段国产GPU性能难以

满足大模型训练要求，与国际主要采用的英伟达A100芯片存在显著差距。例如，国产算力最高的昇腾910芯片计算速度（320 TFLOPS）仅与英伟达A100 PCIe版本持平，与英伟达H100 NVL版本相差10倍以上（表4）。另外，国产人工智能智算芯片配套的编程环境尚不成熟。与英伟达的并行计算平台和编程模型（CUDA）工具包相比，我国相应软件生态建设仍需加强，这是一个投入巨大并且漫长的过程。

3.4 创新主体无序竞争制约整体发展速度

包括：①“百模大战”引发无序竞争，由于数据“孤岛”、赛道重叠、市场竞争等原因企业各自为战，造成资源投入分散、共创共建开源意愿不足等问题。数据显示，截至2023年10月，我国有互联网企业

表3 2023年我国大模型的省级政府层面相关政策

Table 3 China's provincial government level policies related to large-scale models in 2023

省级行政区	政策	发布时间	大模型相关内容
北京市	《北京市人工智能行业大模型创新应用白皮书(2023年)》	2023年11月	从大模型全球发展态势、国内外行业应用概述、北京应用情况和发展建议等方面进行了系统分析和阐述,旨在进一步推动大模型应用落地,展示北京市大模型应用成果,促进大模型价值传播和供需对接
	《北京市加快建设具有全球影响力的人工智能创新策源地实施方案(2023—2025年)》	2023年5月	强调构建高效协同的大模型技术产业生态,建设大模型算法及工具开源开放平台,构建完整大模型技术创新体系,积极争取成为国家人工智能开放生态技术创新中心
	《北京市促进通用人工智能创新发展的若干措施》	2023年5月	强调开展大模型创新算法及关键技术研究、加强大模型训练数据采集及治理工具研发,推动大模型在政务、医疗、科学、金融等领域的示范应用
	《北京市通用人工智能产业创新伙伴计划》	2023年5月	推进人工智能大模型研发和应用,赋能千行百业数智化转型
上海市	《上海市推动人工智能大模型创新发展若干措施(2023—2025年)》	2023年11月	推进大模型创新应用,支持大模型在智能制造、生物医药、集成电路、智能化教育教学等领域构建示范应用场景,打造标杆性大模型产品和服务
广东省	《广东省人民政府关于加快建设通用人工智能产业创新引领地的实施意见》	2023年11月	强调持续加快技术创新场景应用,加强技术与经济、社会、科学领域深度融合,打造一批示范性强、带动性广、显示度高的典型应用场景;通过场景创新促进通用人工智能关键技术迭代升级,形成技术供给和场景需求互动演进的持续创新力,带动提升制造、医疗、教育、金融、科学研究等领域的发展水平
安徽省	《打造通用人工智能产业创新和应用高地若干政策》	2023年11月	加快智算中心建设,推进大模型等关键核心技术攻关,推动全面开放大模型应用场景

资料来源：中国政府网、北京市人民政府网、人民网、《北京市人工智能行业大模型创新应用白皮书（2023年）》

Data source: www.gov.cn; www.beijing.gov.cn; www.people.com.cn; *Beijing Artificial Intelligence Industry Large-Scale Model Innovation Application White Paper (2023)*

(百度、字节跳动、阿里巴巴等)、新兴创业公司（百川智能、MiniMax、月之暗面等）、传统AI企业（科大讯飞、商汤科技等），以及高校科研院所等254家单位开展了通用大模型研发^[15]，导致资源碎片化投入，重复低水平建设^[16]，计算资源竞争加剧。② 国产大模型应用软硬件适配与协同优化尚显不足，软硬件生态有待进一步丰富。对比国内外大模型产品应用流量来源，国外大模型来自移动端的用户流量远高于国产大模型，且国产大模型产品应用在电子邮件、社交应用

程序、自然搜索等外接应用流量上也远低于ChatGPT（表5）。③ 现有国产大模型尚未探索出合适的大模型开源商业模式。我国在开源商业化方面的实践经验不足，采取的开源商业策略单一，企业多面临“技术业务两张皮”的困境，尚未实现诸如微软Office365 Copilot、ChatGPT企业版等对企产品的商业化落地，难以搭建起可持续的大模型下游开源商业生态。目前，按照交易量收取费用、定制开发收取费用是国产大模型产品主要收费模式，这些商业模式难以覆盖大

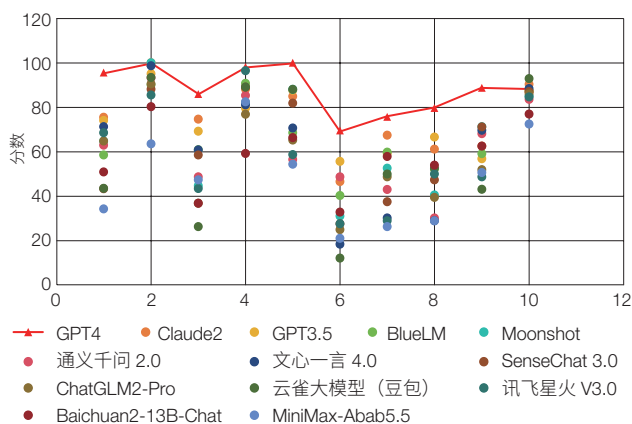


图2 部分大模型 SuperCLUE 总分及
十大基础能力得分 (2023 年 10 月)

Figure 2 Total and top ten basic capabilities scores of
SuperCLUE for some large-scale models (Oct. 2023)

中文通用大模型综合性测评基准 (SuperCLUE) 是针对中文可用的通用大模型的一个测评基准, SuperCLUE 得分旨在评价各类中文大模型的效果情况

SuperCLUE is a comprehensive Chinese large language model (LLM) benchmark, and the score of SuperCLUE is used for evaluating the effectiveness of various Chinese LLMs

模型开发所需的巨大算力及人力成本, 且多为一次性付费, 致使与软硬件生态之间的开源协作受阻。

3.5 开源支持体系建设水平较低

目前, 我国从大模型开发、训练到应用的全链条开源支持体系水平较低, 不利于集中优势力量, 阻碍了技术突破的步伐。①在开源开发平台方面, 我国 Gitee、GitLink、AtomGit 等开源代码托管平台发展尚不完善。例如, 国内 Gitee 等代码托管平台因网络及设备故障而导致用户存储代码丢失的大型故障时有发生, 且维护不透明, 运营稳定性较差, 因此难以维持用户使用黏性; 而国外的美国 Github 专门有网站记录所有故障及修复时间, 稳定的运营机制极大增强了用户信任度, 从而促进了用户的使用量。这一差距充分

反映在访问统计数据上, 我国开源代码托管平台 Gitee 的每月访问量为 800 万次, 美国 Github 平台则高达 4.32 亿次。②在开源测试和训练平台方面, 国际流行的人工智能开源模型库和社区平台 Hugging Face 发展至今已集成了超过 50 万具备图像识别、语音生成、文本生成等多种功能的开源大模型和超过 11 万包含多种数据类型的高质量开源数据集, 有全球超过 5 万家组织使用该平台, 形成了较为成熟的大模型开源工具平台生态。然而, 我国类似的开源平台发展仍处于初级阶段, ModelScope 魔搭开源平台不仅公布的数据集、模型质量参差不齐, 部分有较多漏洞, 难以进一步开发优化或直接应用, 而且开源共建水平较低, 如 ModelScope 魔搭社区开源的 2 158 个模型中接近 60% 的模型由排名前 10 位的贡献者捐出, 超 1/3 模型由阿里巴巴达摩院一家贡献。大模型开源代码托管、训练、测试平台的低水平致使国产大模型往往托管在国外平台上, 造成我国大模型的训练环境 and 应用场景流失在国外, 难以保留在国内, 不利于自主发展。③在开源治理协调平台方面, 我国相关治理机构缺乏与业界的及时深度交流, 导致对开源大模型中涉及的“开源”认定^④、版权归属界定等关键问题认知不足, 难以在负责任开源大模型生态建设过程中发挥引导和平衡作用。同时, 开源基金会等开源促进组织发展尚处于初级阶段, 开源项目运营经验不足, 运营能力欠缺, 难以有效支持大模型开源项目的持续发展。

4 我国构建大模型开源创新生态的建议

我国应充分吸收开源创新生态构建经验, 秉持开源开放的理念构建大模型开源创新生态, 推动大模型全产业链的繁荣有序发展。一方面, 政府要处理好打

④ 根据大模型应用程序接口、模型参数权重、模型训练代码、模型训练数据是否开放可对大模型开放程度进行区分。目前对大模型的“开源”认定尚不明晰, 存在争议。国际开源标准组织开源促进会 (Open Source Initiative, OSI) 发布的开源人工智能定义最新版草案认定只要应用程序接口、模型参数权重开放的大模型即为开源大模型 (<https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-8>)。

表4 国内外各型号AI芯片算力

Table 4 Computational power table of AI chips from home and abroad suppliers

类别	供应商	型号	数据类型INT8	数据类型FP16
国内	华为技术有限公司	昇腾910	640 TOPS	320 TFLOPS
	中科寒武纪科技股份有限公司	思元290	512 TOPS	—
	百度集团	昆仑芯二代AI芯片	—	128 TFLOPS
国外	美国英伟达公司	A10080GB Pcle	624 TOPS	312 TFLOPS
		A10080GB SXM	1 248 TOPS	624 TFLOPS
		H100 SXM	3 958 TOPS	1 979 TFLOPS
		H100 Pcle	3 026 TOPS	1 513 TFLOPS
	美国AMD公司	H100 NVL	7 916 TOPS	3 958 TFLOPS
		Instrinct MI250	362.1 TOPS	362.1 TFLOPS
		Instrinct MI250X	383 TOPS	383 TFLOPS

注：TOPS和TFLOPS均是计算性能的度量单位，其中TOPS全称为tera operations per second，表示处理器每秒钟可进行的整型运算万亿次数；TFLOPS全称是tera floating-point operations per second的缩写，表示处理器每秒钟可进行的浮点运算万亿次数

Note: TOPS and TFLOPS are both units of measurement for computing performance. To be specific, the TOPS is short for tera operations per second, which indicates the number of trillion integer operations a processor can carry out per second; and the TFLOPS is short for tera floating-point operations per second, representing the number of trillion floating-point operations that can be executed per second

表5 国内主流大模型各渠道流量与ChatGPT对比(2023年10月)

Table 5 Comparison of traffic volume from various channels for mainstream domestic large-scale models and ChatGPT (Oct. 2023)

流量类别	文心一言	通义千问	讯飞星火	字节豆包	智谱清言	百川智能	ChatGPT
移动设备(万次)	143.66	87.37	769.61	26.09	52.31	35.68	240 641.13
移动设备占比(%)	6.50	35.20	19.00	11.10	21.90	46.10	43.80
直接(万次)	1 320.00	133.00	512.00	193.00	164.00	45.40	368 488.32
电子邮件(万次)	7.31	0.94	0.92	0.24	0.76	0.00	561.72
外链(万次)	273.00	69.20	76.60	26.30	22.50	3.37	9 642.86
社交(万次)	6.64	2.11	4.93	2.21	1.38	1.12	90 343.30
自然搜索(万次)	603.00	42.90	175.00	13.20	50.20	27.50	80 372.77

资料来源：<https://similarwebcn.com>
Data source: <https://similarwebcn.com>

造大模型开源生态过程中政府和市场之间的关系，相关部委要明确职责，形成政策合力。另一方面，社会要建立起对开源的合理认知，通过数字公共品制度等探索构建符合大模型产业特性的开源治理体系，推动形成涵盖大模型上下游全产业链的健康开源创新生

态，促进大模型产业创新与可持续发展。具体包括以下4个方面。

4.1 加强顶层设计，明确各个部门职责

(1) 建议效仿中央科技委员会统筹全国科技发展总体部署的机制，国家层面建立统筹大模型发展的组

织或机制。明确中央网络安全和信息化委员会办公室、国家发展和改革委员会、工业和信息化部、科学技术部、教育部、国家数据局等相关部委在大模型及上下游产业链各环节发展中的具体职责，并进行有效统筹。持续关注大模型产业及上下游发展需求，为打造可持续的大模型开源创新生态提供协同有差异的政策支持与资源保障，形成合力促进大模型产业发展。

(2) 以数据、算力和算法为抓手补短板、固底板，推动产学研持续投入大模型开源技术研发。建议由中央网络安全和信息化委员会办公室、工业和信息化部负责大模型产业培育引导，科学技术部、中国科学院、教育部等合作推动大模型底层技术及原理研究，培养产业发展所需的人工智能架构设计方面人才，国家发展和改革委员会牵头地方政府做好算力中心、跨区域算力网络的建设及运营；数据局厘清数据产权、数据资产评估等相关阻碍数据产业链发展的相关问题，推动上游数据产业链繁荣有序健康发展。

4.2 打造共享的大模型研发基础体系

(1) 建设开放国家算力平台支持大模型训练。解决跨数据中心算力协同面临的相关体制机制挑战，提高各地已有智算中心的利用率和使用效率。推动国家实验室算力平台向社会开放，支持组建算力联盟引导算力开放，集中高档GPU算力资源，降低各类大模型研发训练成本。设立国家级开源项目推动头部科技企业搭建公共大模型基础平台、构建低代码开发工具，促进上、中、下游企业间的协同创新。加快落实《算力基础设施高质量发展行动计划》，发挥算力对大模型发展的驱动作用。

(2) 推动建立国产智算芯片开源编译生态。统一各国产智算芯片编译环境接口，构建类CUDA平台打通硬件和AI训练之间的中间软件层，加大对适应人工智能计算所具有的计算密度高、需要大量低精度计算等特点的软硬件协同设计研发^[17]。这能够降低采用不同GPU进行大模型训练时额外的学习成本，有利于大

模型发展。同时开源所汇聚的合力能降低芯片厂家的开发成本，促进算力领域技术研发，加快国产GPU芯片发展。注重与国内硬件生态连接，形成软硬件有效协同，提升产业创新体系整体效能。通过设立大模型开源大基金等方式，推动国产大模型开源软硬件生态发展，形成基础软硬件与大模型有效协同。

(3) 促进开放数据体系建设。发挥国家数据局的统一协调作用构建高质量数据集，扩大政府开放数据范围并通过建立多层次数据开放体系加强数据交换共享，形成大模型发展的开放数据支撑。加快构建有利于促进大模型产业发展的数据版权制度，借鉴国外大模型训练版权责任豁免机制，探索实现更为逻辑周密和利益平衡的数据版权规则设计。

4.3 强化全产业链开源开放体系建设

加强大模型相关全产业链生态布局，推动大模型开发、训练、应用全链条支撑平台有组织地建设，由中立的组织机构主导、科技企业参与大模型产业创新生态基础层和模型层的开源，由科技企业主导大模型产业创新生态中间层和应用层的开源。

(1) 从产业生态的角度引导推动大模型产业应用落地。全面调研和布局大模型相关的产业链，促进开源大模型在行业核心应用场景如生物医药、智能化教育教学、智能制造等领域进行应用示范，推动开发各类新型应用场景，支持AI创新企业采用公共算力开发行业智能应用，引导行业用户与大模型厂商合作，推动各行业智能化升级。

(2) 加强面向开源代码的计算和训练型大模型平台的设计开发和推广。对标GitHub和Hugging Face等建设利于大模型开发、测试和训练的开源平台，开展我国开源平台建设工作，助力大模型的利用和推广。发挥开源基金会或新型研发机构作用，引导企业依托国内代码托管平台开源一批具有行业影响力的软件项目，积极培育我国开源生态环境。

(3) 探索新型大模型商业开源运营机制。借鉴

OpenAI的“非营利性机构+有限入股营利回报”模式，加强市场主导和产业政策支撑共同推进基础大模型市场建设，构建可持续的开源创新成果商业模式。

(4) **鼓励社会资本参与开源大模型技术的产业投资。**推动社会资本参与大模型产业的风险投资和产业投资，探索建立线下孵化器空间，联合开源社区及代码托管平台共同打造线上线下融合、极具活力的开发者社区，促进开源大模型下游商业生态繁荣发展。

4.4 完善开源创新治理体系鼓励发展

(1) **推动商业开源政策研究。**研究制定有利于开源商业化实施的相关政策，推动建成公众贡献数据和使用数据行业规范等数字公共产品制度，强化开源许可证的法律效力，有力保护开源成果知识产权，将“开源不等于免费”的开源理念贯彻到大模型产学研用全过程。研究制定实验室开源大模型开源许可机制，针对开源社区上不同类型下游开发者和用户，打造不同开源层级的许可协议，授权开源使用。推动开源产业发展，以税收优惠等方式鼓励企业积极探索开源，参与开源生态建设，深入了解开源回馈方式，寻找有效的基于开源的商业反馈模式。

(2) **推动开源社区治理水平提升。**持续支持国内开源基金会、开源社区等开源力量发展，推动开源文化理念在社会的广泛传播。提高开源社区运营水平，运用大数据分析手段精确评估社区内参与合作者的贡献情况，精准识别社区内核心开源贡献者并予以奖励，形成良好的“贡献-承认”正向反馈循环。完善大模型开源评价、安全评估框架等监测机制，以推动大模型产业良性健康发展。

(3) **推动大模型开源国际交流合作。**打造具有国际先进技术水平的大模型开源开放平台，并加强与国际沟通大模型伦理治理，参与探讨制定国际标准。鼓励企业融入国际顶尖开源社区、参与开源规则制定等，通过开源争取全球智慧。依托开源社区，加强大模型技术人才自主培养和国际交流，推动高校、科研

院所与企业培育更多有热情做开源贡献的人才。

参考文献

- 1 Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models. arXiv, 2022, doi: 10.48550/arXiv.2108.07258.
- 2 Bonaccorsi A, Rossi C. Why open source software can succeed. Research Policy, 2003, 32(7): 1243-1258.
- 3 Xie, X M, Wang H W. How can open innovation ecosystem modes push product innovation forward? An fsQCA analysis. Journal of Business Research, 2020, 108: 29-41.
- 4 中国信息通信研究院. 开源生态白皮书. 北京: 中国信息通信研究院, 2020.
China Academy of Information and Communications Technology. Open Source Ecosystem White Paper. Beijing: China Academy of Information and Communications Technology, 2020. (in Chinese)
- 5 余江, 丁禹民, 刘嘉琪, 等. 深度数字化背景下开源创新的开放机理、治理机制与启示分析. 创新科技, 2021, 21(11): 13-20.
Yu J, Ding Y M, Liu J Q, et al. Analysis of the open mechanism, governance mechanism, and enlightenment of open source innovation under the background of deep digitalization. Innovation Science and Technology, 2021, 21(11): 13-20. (in Chinese)
- 6 隆云滔, 王小明, 顾荣, 等. 国际开源发展经验及其对我国开源创新体系建设的启示. 中国科学院院刊, 2021, 36(12): 1497-1505.
Long Y T, Wang X M, Gu R, et al. International open source development experience and its enlightenment to the construction of China's open source innovation system. Bulletin of Chinese Academy of Sciences, 2021, 36(12): 1497-1505. (in Chinese)
- 7 Le S T, Fan A, Akiki C, et al. Bloom: A 176b-parameter open-access multilingual language model. arXiv, 2022, doi: 10.48550/arXiv.2211.05100.
- 8 Feller J, Fitzgerald B, Hissam S A, et al. An analysis of open source business models// Perspectives on Free and Open

- Source Software. Cambridge: MIT Press, 2007: 279-296.
- 9 Liang X, Li A Q, Zhu L, et al. SuperCLUE: A comprehensive Chinese large language model benchmark. arXiv, 2023, doi: 10.48550/arXiv.2307.15020.
 - 10 黄鹏, 李宏宽. 中国开源软件生态构建的风险及对策. 科技导报, 2021, 39(2): 83-95.
Huang P, Li H K. Risks and Countermeasures in the construction of China's open source software ecosystem. Science & Technology Review, 2021, 39(2): 83-95. (in Chinese)
 - 11 Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv, 2020, doi: 10.48550/arXiv.2001.08361.
 - 12 Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. arXiv, 2022, doi: 10.48550/arXiv.2206.07682.
 - 13 中国信息通信研究院. 数据要素白皮书. 北京: 中国信息通信研究院, 2023.
China Academy of Information and Communications Technology. Data Element White Paper. Beijing: China Academy of Information and Communications Technology, 2023. (in Chinese)
 - 14 中国信息通信研究院. 中国算力指数发展白皮书. 北京: 中国信息通信研究院, 2023.
China Academy of Information and Communications Technology. China Computing Power Index Development White Paper. Beijing: China Academy of Information and Communications Technology, 2023. (in Chinese)
 - 15 北京市科学技术委员会, 中关村科技园区管理委员会. 北京市人工智能行业大模型创新应用白皮书(2023年). 北京: 北京市科学技术委员会, 2023.
Beijing Municipal Science and Technology Commission, Zhongguancun Science and Technology Park Management Committee. Beijing Artificial Intelligence Industry Large-Scale Model Innovation Application White Paper (2023). Beijing: Beijing Municipal Science and Technology Commission, 2023. (in Chinese)
 - 16 中国科学技术信息研究所. 中国人工智能大模型地图研究报告. 北京: 中国科学技术信息研究所, 2023.
Institute of Scientific and Technical Information of China. Research Report on the Map of China's Artificial Intelligence Large-Scale Models. Beijing: Institute of Scientific and Technical Information of China, 2023. (in Chinese)

Challenges and recommendations for building open source innovation ecosystem for large-models in China

WEN Xin¹ ZHANG Chao^{2*} GUO Rui³ CHEN Kaihua³ FENG Ze¹ ZHU Qigang^{4,5}

(1 Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190, China;

2 Institute of Strategic Studies, Shanghai Jiao Tong University, Shanghai 200052, China;

3 School of Public Policy and Management, University of Chinese Academy of Sciences, Beijing 100049, China;

4 Shanghai Open Source Information Technology Association, Shanghai 201620, China;

5 Institute of Global Open Source Collaboration, University of International Business and Economics,
Beijing 100029, China)

Abstract Addressing the current technological development issues that constrain the development of China's large-scale model industry is needed to promote the continuous prosperity and development of the industry and enhance its international competitiveness. The study analyzes the significance of the open-source innovation ecosystem for the development of large-scale models in China. Based on reviewing the international experience of constructing the open-source innovation ecosystem, it further dissects the problems and challenges faced by the construction of the open-source innovation ecosystem for large-scale models in China and puts forward targeted suggestions. The study finds that the open-source innovation ecosystem for large-scale models in China is faced with problems such as the restriction of ecological formation by technical capabilities, the significant limitation of technological development by data computing power, the disorderly competition among innovation subjects restricting the overall development speed, the low level of construction of the open-source support system, and the lack of the design of the system collaborative policy architecture, which hinder its rapid development and the improvement of competitiveness. There is a need to strengthen the top-level design and coordinated development, to build a shared basic system for large-scale model research and development, to enhance the construction of open-source and open systems throughout the entire industry chain, and to improve the governance of the open-source innovation system for large-scale models.

Keywords large-scale model development, open-source innovation ecosystem, policy recommendations

温 馨 中国科学院科技战略咨询研究院博士研究生。主要研究领域为开源创新、创新管理与创新政策等。

E-mail: wenxin212@mailsucas.ac.cn

WEN Xin Ph.D. student at Institutes of Science and Development, Chinese Academy of Sciences (CAS). Her main research interests include open source innovation, innovation management and innovation policy, etc. E-mail: wenxin212@mailsucas.ac.cn

张 超 上海交通大学国家战略研究院助理研究员。主要研究领域为创新生态系统、创新发展政策、创新投融资等。

E-mail: 18061291419@163.com

ZHANG Chao Assistant Professor at Institute of Strategic Studies, Shanghai Jiao Tong University. His research interests focus on innovation ecosystem, innovation development policy, investment and financing of innovation, etc.

E-mail: 18061291419@163.com

■责任编辑：文彦杰

*Corresponding author