引用格式：李兴腾, 冯锋, 黄鹂强. 突破人工智能大模型的“数据瓶颈”——构建国家级语料库运营平台的思考. 中国科学院院刊, 2025, 40(3): 522-529, doi: 10.16418/j.issn.1000-3045.20240510001.
Li X T, Feng F, Huang L Q. Breaking through “data bottleneck” of AI large models—Reflections on building a national corpus operation platform. Bulletin of Chinese Academy of Sciences, 2025, 40(3): 522-529, doi: 10.16418/j.issn.1000-3045.20240510001. (in Chinese)

突破人工智能大模型的“数据瓶颈”

——构建国家级语料库运营平台的思考

李兴腾[1*] 冯 锋[2] 黄鹂强[3]

1 浙江大学 公共管理学院 杭州 310058
2 中国科学技术大学 管理学院 合肥 230026
3 浙江大学 管理学院 杭州 310058

摘要 当前，全球人工智能大模型行业竞争日趋激烈，语料库成为提升人工智能大模型技术性能和应用效果的关键。但是，我国语料库在数量和质量上均存在不足，难以满足快速发展的人工智能大模型训练需求。从全球来看，各国都在加快语料库发展，特别是推动高质量语料库的建设和应用。因此，文章基于国外对标和国内环境分析，从平台定位、总体架构、运营主体、核心内容等维度提出建设国家级语料库运营平台的建议。

关键词 人工智能，大模型，语料库，数据瓶颈

DOI 10.16418/j.issn.1000-3045.20240510001
CSTR 32128.14.CASbulletin.20240510001

习近平总书记强调，人工智能是引领这一轮科技革命和产业变革的战略性技术，具有溢出带动性很强的“头雁”效应。从全球范围来看，人工智能（AI）大模型行业竞争日趋激烈，美国、欧盟、日本等密集出台AI发展战略，全体提升自身科技竞争实力[1]。语料作为AI大模型训练的基础，其范围、数量和质量直接影响到模型的训练效果和性能，高质量语料库已然成为提升系统准确性和泛化能力的核心。因此，构建国家级语料库运营平台显得尤为重要，它不仅是实现高质量数据供给的重要渠道，也是促进我国产业升

*通信作者
资助项目：国家社会科学基金重大项目（24&ZD072），中国博士后基金（2023M731171）
修改稿收到日期：2025年3月7日

级、技术进步的关键力量，更是提升AI国际竞争力的必由之路。

1 数据瓶颈：AI发展面临训练数据枯竭问题

1.1 全球AI大模型行业竞争日益加剧

AI大模型领域呈现前所未有的技术创新活力和全球竞争态势。多个国家投入大模型研发阵营，美国谷歌、OpenAI等机构较早开始大模型技术研发，欧盟、俄罗斯、以色列、韩国等地区和国家也紧跟其后，加入全球AI大模型研发阵营。特别是在ChatGPT发布以来，全球范围内的AI大模型迎来了空前的发展高潮。近年来，我国进入大模型加速发展期，在自然语言处理、机器视觉和多模态等各技术分支上发展迅猛，不仅涌现出“文心一言”“通义千问”“星火认知”等一批具有行业影响力的AI大模型，特别是随着DeepSeek-R1、V3、Coder等系列模型为代表的AI成果不断涌现，国产模型在语言理解、内容生成和逻辑推理等方面展现出强大的能力，初步形成一流的AI大模型技术群。从区域分布来看，当前全球大模型呈现出“美国领跑、中国紧跟、其他区域落后”的态势。2025年，全球AI的竞争将进一步升级为系统性竞争，各国将在基础大模型、行业应用、硬件、产业链等方面展开全面较量。

AI大模型领域日益成为中美两国科技竞争的前沿阵地。从全球已发布的AI大模型分布来看，中国和美国大幅领先，合计数量超过全球总数的80%[①]，这充分显示了中美两国在AI大模型领域的领先地位和强大实力。AI大模型的竞争，已经不仅仅是技术层面的竞争，更是国家科技战略的竞争。美国将优先发展AI上升为国家战略，不断向AI领域发展投入大量资源，以实现绝对的优势。而且，美国将中国确定为AI领域的主要竞争对手，出台了一系列法规和政策来限制中国在AI领域的技术获取和合作机会，尤其是针对AI芯片和大模型技术的封锁和限制。例如，美国陆续出台《2020年国家人工智能倡议法案》（*National Artificial Intelligence Initiative Act of 2020*）、《2022年芯片与科学法案》（*CHIPS and Science Act 2022*）等文件，对中国实施AI芯片新限制，试图通过封锁算力抑制中国AI大模型的发展，使美国成为“头号玩家”。细观中国AI大模型产业，得益于政策、技术和市场的共同驱动：一方面，中国政府强有力的政策支持和不断扩大的市场需求为中国AI大模型行业的蓬勃发展提供了有力保障，企业技术创新主体地位更加凸显；另一方面，美国的限制措施和技术封锁，客观刺激和促进了中国技术创新水平的提升，助力中国在全球大模型领域竞争力提升。

1.2 语料库成为大模型竞争的关键要素

AI大模型训练对数据供给要求极高。AI是第四次工业革命的“核心引擎”，数据是AI大模型发展的“燃料”。AI大模型技术的快速迭代，不仅带来对数据的海量需求，也对数据集的构建提出了更多挑战。因为训练AI大模型需要大规模、高质量、多模态的数据集，这些数据通常来自各个领域和多个数据源，包含文本、图像、语音、视频等多种形式。近年来，AI大模型训练所用的数据集规模呈现出显著的增长趋势。以DeepSeek系列模型为例，DeepSeek-LLM（V1）通过数据去重、过滤和混洗（remixing）3个阶段，构建了一个包含约2万亿token[②]的中英双语预训练数据集，以确保数据多样性和高质量；DeepSeek-V2扩展了数据量并提高了数据质量，模型预训练所使用的语料库包含8.1万亿token的多语言数据集；DeepSeek-V3通过提高数学和编程样本的比例来优化预训练语料库，

① 科技部新一代人工智能发展研究中心. 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.

② 在人工智能领域，token指在自然语言处理过程中用来表示处理文本的最小单元或基本元素；token可以是单个字符，也可以是多个字符组成的序列。

模型预训练所使用的语料库提升到14.8万亿token的多语言数据集。

*语料将成为AI时代的下一个竞争焦点。*在AI时代，语料库将成为提升AI大模型技术性能和应用效果的关键。语料数据作为AI大模型优秀输出能力的保证，已经被广泛应用于自然语言处理、机器翻译、智能问答、情感分析等多个领域，成为推动AI技术进步的关键因素。而且，各国都在加快语料库发展，特别是推动高质量语料库的建设和应用。

1.3 训练数据短缺成为全球共性问题

*AI技术的快速迭代，加剧数据供需矛盾。*AI大模型训练所需要的数据集的增速远大于高质量数据生成的速度，将会导致高质量数据逐渐枯竭。专注于AI发展趋势的研究团队EPOCH AI，在研究中预测，最早在2024年人类就可能会陷入训练数据荒，届时全世界的高质量训练数据都将面临枯竭[2]。尽管他们在最新的研究中，将高质量文本数据耗尽的时间推迟到2026—2032年，但是依旧认为训练数据是AI大模型技术发展的主要瓶颈[3]。在此背景下，企业加大了对数据资源的竞争，为了获取更多数据，包括OpenAI、Meta在内的多家企业不断调整数据采集和使用条款，甚至公开讨论如何规避版权保护。*因此，高质量数据短缺将成为制约AI技术发展的重要因素，平衡科技创新与版权保护之间的关系也是不能回避的现实问题。*

2 高质量语料库：人工智能大模型发展的核心动能

2.1 训练数据直接影响大模型的内容生成

*数据的质量、规模和多样性直接影响AI大模型的性能。*数据规模是AI大模型预训练的基础，数据质量直接影响模型最终生成的内容质量。如果训练数据准确、全面且具备代表性，那么AI大模型在分析和生成自然语言文本方面的能力将得到显著提升，从而更精确地模拟和理解人类语言的复杂性和多样性。此外，通用参数、文本语言、图像、视频音频等不同类别的数据类型直接影响AI大模型的认知边界。而且，AI大模型所需要的数据根据训练阶段有所不同。以ChatGPT为例，在预训练阶段主要关注数据的类型广泛度，需要包括网页、图书、学术论文、新闻报道、社交媒体文本、代码等形式在内的各类数据；在监督微调（SFT）阶段和基于人类反馈的强化学习（RLHF）阶段更关注人类认知的数据，因为这2个阶段是对AI大模型泛化能力和涌现能力的训练，对于数据质量要求较高，强调语料特征与人类价值观的一致。

*数据质量问题对AI大模型生成内容的负面影响不容忽视。*如果训练数据存在错误、偏见或信息稀缺，这些问题将在模型生成的文本中得以体现。*① 准确性问题。*如果训练数据中包含错误或不准确的信息，AI大模型将会学习并重现这些错误，这可能导致模型在生成文本时产生事实性错误或误导性信息。*② 偏见和刻板印象。*数据中的偏见和刻板印象也会被模型学习并反映在其生成的文本中。例如，如果训练数据中存在性别、种族或文化的刻板印象，模型可能会在生成的内容中无意中强化这些偏见。*③ 数据稀缺性。*如果训练数据中某些类型的信息较为稀缺，模型在处理这些信息时可能会表现不佳。总之，不准确的数据可能导致模型产生事实性错误，数据中的偏见会无意识地被模型学习和重现，而数据的稀缺性则可能限制模型在处理特定信息时的表现。

*高质量数据对模型内容生成具有积极影响。*将AI大模型打造成新质生产力工具，建设高质量语料库是关键③。利用高质量数据进行训练，可以显著提升大模型生成内容的准确性、客观性和多样性。*① 提高准*

③ 阿里研究院. 中美大模型的竞争之路：从训练数据讲起. 杭州：阿里巴巴集团，2023.

确性。准确无误的数据集可以帮助模型学习到正确的语言模式和知识，准确模拟真实世界，使模型的预测更贴近实际数据分布。② 增强客观性。经过仔细筛选和清洗数据，并借助优化算法减少训练中的损失函数，可以最大程度地减少数据中的偏见和刻板印象，保证模型生成的文本更加中立和客观。③ 丰富多样性。多样化的训练数据可以使模型在处理不同类型的信息时都能表现出色，无论是通用知识还是专业领域的知识。

2.2 高质量中文语料库建设意义重大

高质量的中文语料数据尤为稀缺。受制于数据集建设的高额成本，以及尚未成熟的开源生态，国内开源数据集在数据规模和语料质量上相比海外仍有较大差距，进而导致数据来源较为单一，且更新频率较低，影响模型的训练效果。据相关数据估算，国内互联网中文语料的质量和规模均大幅低于英文语料，英文文本和数据资料是中文的8倍左右；并且，以公开渠道获取大批量、高质量的中文语料数据的难度较大。而且，中文语料、科研成果等高质量数据集开放程度低，企业用于训练的语料来源不清晰、权属不明确，开源后存在一定的合规隐患，这使得企业更倾向于自采、自用，国内AI大模型数据流通机制尚未形成[3]。

高质量中文语料库建设势在必行，中式价值观类语料更为必要。AI大模型需要依赖现实语料库进行训练，因而可能会延续现实社会中存在的偏见和价值偏差，甚至会因为快速和低成本的应用加剧这些偏见和偏差。当前，中文语料库面临总量不足、分布不均、垂直覆盖有限、质量参差不齐等问题，导致国内许多从事AI大模型开发的机构在进行模型训练时，不得不依赖于外文标注数据集、开源数据集或是爬取网络数据。在国际形势日趋复杂的态势下，意识形态之争正在逐步加剧，而AI大模型很可能被“武器化”，成为进行舆论引导的新工具——经英文语料库训练出来的AI大模型，不可避免地更符合西方主流价值观。因此，需要加大对高质量中文语料库，尤其是反映优秀传统文化和本土价值观的中式价值观类语料的开发[4]，尽快掌控中文语料库的话语权，既是帮助大模型更好地理解和反映我国的文化背景和价值取向，也能在价值引导方面占据主动地位。

2.3 “扩源提质”打造高质量语料库

“扩源提质”是建设高质量语料库的有效策略。“扩源”意味着要不断扩大数据的来源和多样性，通过收集、汇聚社交媒体文本、学术论文、新闻报道等多种来源的数据，覆盖文本、图像、视频、音频等多种数据类型，为大模型提供丰富的语言环境和知识背景。“提质”则强调的是提升数据的质量和准确性，对数据进行去重、格式化、迭代更新、标注、内容监督等深入挖掘和精细化处理，形成包含预训练数据集、指令微调数据集、测试数据集等内容的、高效可用的多模态语料库，以支持后续数据的深度分析、模型训练，以及数据应用与服务需求。

高质量合成数据获将是普通数据的有效补充。基于各类原始数据，运用模数学模型创建生成新的合成数据，能够为模型提供训练材料。例如，专攻棋类的AlphaZero就是使用合成数据训练出来的。合成数据既可以基于真实数据构建，也可以通过现有模型或者人类专业知识创建；合成数据在丰富数据多样性的同时，能够更快地生成多模态数据，帮助模型预训练。但是，由于合成数据生成过程可能存在偏差或噪声，其质量和真实性无法完全模拟客观世界，在数据可信度、泛化能力及伦理方面面临更多的挑战。因此，基于当前数据现状，以及合成数据的发展实践来看，合成数据为丰富模型训练数据提供了一种解决方案，但

④ 阿里研究院. 大模型训练数据白皮书. 杭州: 阿里巴巴集团, 2024.

是要想让合成数据成为有效的训练数据，必须保证合成数据的质量。

3 语料库运营平台：提升人工智能国际竞争力的必由之路

3.1 对标国外：欧美国家积极建设语料库运营平台

美国、欧盟积极建设语料库运营平台以实现各类语料库的汇聚、开发、利用。例如，美国最全面的公共数据平台Data.Gov、欧盟“共同数据空间”（Common European Data Spaces）等。通过对国外语料库运营平台架构分析发现，这些平台建设内容主要包括数据汇聚共享、数据治理，以及安全监管等方面。具体来看，各国主要基于数据处理不同的阶段进行平台的设计和建设。

数据汇聚阶段，各国不断扩大数据来源，并选取合理方式实现数据汇聚。各国加大对公共、企业、个人数据汇聚的同时，注重对科研数据的收集、汇聚。例如，欧盟“共同数据空间”汇聚了法律、气象、安全执法等公共数据，制造业、绿色节能、交通、健康等17类行业数据，以及姓名、邮箱等个人数据。在数据汇聚方式上，大多采用物理汇聚和逻辑接入的方式。例如，欧盟出于对数据安全的考量，更倾向于逻辑接入，而非物理汇聚方式进行集中存储。

数据治理阶段，国内外普遍通过数据清洗、数据标准化、数据标注、数据质量评价等方式实现数据高效治理。具体实践中，数据清洗更多侧重明确清洗规则、使用自动化技术和工具；数据标准化旨在统一数据格式、数据类型、数据命名等规范；数据标注环节关注标注技术和工具研发、人才培养和生态培育等内容；数据质量评价更多侧重数据质量评价指标体系打造、反馈机制及优化等内容。例如，美国Data.gov主要采取包括人工评价、系统自动评估、第三方评价在内的综合数据质量评价体系。此外，国外倡导政府、行业协会、非营利性平台、企业等主体共同参与数据治理，营造良好的数据治理生态。

数据服务阶段，主要通过公共数据平台和社会数据平台提供各类数据服务。具体方式包括：建立检索下载平台、开发数据工具服务、组建语料库联盟、构建开源生态等。例如，大模型训练数据库Common Crawl以API接口服务形式为GPT-3、腾讯WeLM等AI大模型提供语料。而且，国外积极引入数据中介、数据经纪商等多方力量，构建多元服务生态。

数据运营阶段，当前语料库运营平台运营主体主要包括政府、高校和科研机构、非营利（开源）组织，以及大型互联网公司和专业机构。不同类型的运营主体根据对语料库的定位不同，采取不同的建设运营模式，也对应不同收费模式。例如，美国政府基于公私合营打通数据运营全链条，形成以“开放共享数据集+高质量语料库+全生命周期的语料处理+灵活多样的配套运营保障”为核心的全链服务矩阵。此外，语料库运营平台的安全监管和运营生态建设也是各国关注的重点内容。

3.2 国内环境：建设语料库运营平台是科技竞争的必然

发展AI语料库不仅是科技竞争的关键所在，也是落实国家战略、推动产业升级、优化资源配置的重要举措。从国家战略要求看，建设国家级语料库运营平台是落实国家AI战略，发挥平台经济作用，推动高质量发展的重要载体。《新一代人工智能发展规划》的推出，将AI发展放在国家战略层面系统布局、主动谋划。建设国家级语料库运营平台是基于AI大模型发展对高质量、大规模、安全可信语料数据资源需求的现实考量，是加快推进发展AI，促进新质生产力发展的重要引擎[4]。此外，推动平台经济发展是国家立足新发展阶段、贯彻新发展理念、构建新发展格局、推动高质量发展的战略布局。建设国家级语料库运营平台，以数据基础设施为重要支撑，以促进数据关键生产要素价值发挥为目标，能够充分凸显平台建设的价

值和优势。

从产业发展的角度来看，实施“AI+”行动已经成为推动现代化产业体系建设和经济高质量发展的重中之重。AI与实体经济的深度融合，不仅促进传统产业的智能化改造和转型升级，还可以催生出一批新兴产业。数据是AI发展的催化剂，大模型驱动的AI发展对于高质量数据供提出了更高要求。在AI领域，无论是算法的优化、模型的改进还是新技术的应用，都需要大量的数据进行实验和验证。推动语料库运营平台建设，加大高质量语料库供给，才能充分发挥数据的基础资源作用和创新引擎作用。

从资源配置的角度来看，数据资源的集约配置是提高AI技术应用效率的关键。通过建设集中、统一的国家级语料库运营平台，能够避免数据的重复采集和浪费，提高数据资源的利用效率。语料库运营平台还可以通过集成和整合国家AI“五大”训练基地的数据资源，以实现数据资源的互通共享。这不仅可以降低数据获取和处理成本，也能够为企业和个人提供更便捷、高效的AI服务。

3.3 建设策略：积极打造国家级语料库运营平台

3.3.1 明晰平台定位，打造国家语料库汇聚与运营平台

国家级语料库运营平台是抢抓AI发展战略机遇，构筑我国AI竞争优势的重要突破口。① 平台的建设应定位为“国家语料库集聚与运营服务平台”，致力于打造全国范围内最权威、最全面、最精准的语料数据和服务提供载体。因此，平台建设应当突出国家战略部署和基础服务功能，强化其公共属性和公益定位；同时，考虑大规模语料汇聚、治理、开发等工作所需要的巨大资源投入，平台可以通过语料产品的开发来获取运营收益，反哺平台的建设运营。② 平台应兼顾汇聚和运营，不仅能够采集、汇聚和存储海量的语料数据，还应通过数据治理，形成对外提供语料检索、分析和应用的服务能力，以支持自然语言处理、机器学习、AI等领域的研究与应用。③ 平台应以需求为导向，面向AI企业、AI训练基地等具有高质量语料的需求方提供数据服务或产品。④ 平台应着眼于产业发展和生态构建，在数据治理和数据服务等环节，发挥平台优势，充分链接更多市场参与主体，通过专业化、链接型、前瞻性的战略布局，推动市场构建语料生态。

3.3.2 设计总体架构，实现业务和技术的深度融合

业务架构上，国家级语料库运营平台采用“三横三纵”的总体架构（图1）。横向维度，平台贯通数据汇聚、数据治理和数据服务三大环节。数据汇聚模块，以全国一体化政务大数据平台和各省市政务大数据平台为抓手实现公共数据、企业数据、专项数据等各类数据的采集、汇聚；数据治理模块，通过数据清洗、数据标准化、数据标注和数据质量评价的治理手段，形成直接可用于AI大模型训练的预训练数据集、指令微调数据集、监督测试数据集；数据服务模块，提供数据检索、数据共享、数据流通交易等配套服务，着力于开源数据生态打造。纵向维度，平台覆盖技术工具、安全监管、生态创新等“三大能力”的全流程支撑。技术工具方面，通过隐私保护、数据互操作、跨域数据交换等技术的更新迭代，助力语料库打通多主体、跨层级数据流通壁垒；安全监管方面，强调对数据安全、隐私保护和合规性的全面监管，构建“技术+运维+管理”三元语料库安全防护体系，以保证平台平稳运行的基础；生态创新方面，通过数据标准生态、行业多元主体参与生态的打造，增强语料库运营平台价值发挥，向市场传递重构语料生态的顶层设计理念。

技术架构上，建议国家级语料库运营平台采用“1+N”一体化架构设计。国家级语料库运营平台设计必须考虑当前我国数据资源现状，以数据安全为底线，综合考虑国家统筹管理与区域现状特点相结合，注重资源高效利用，推动建立全国数据要素统一大市场。因此，借鉴全国一体化在线政务服务平台建设和

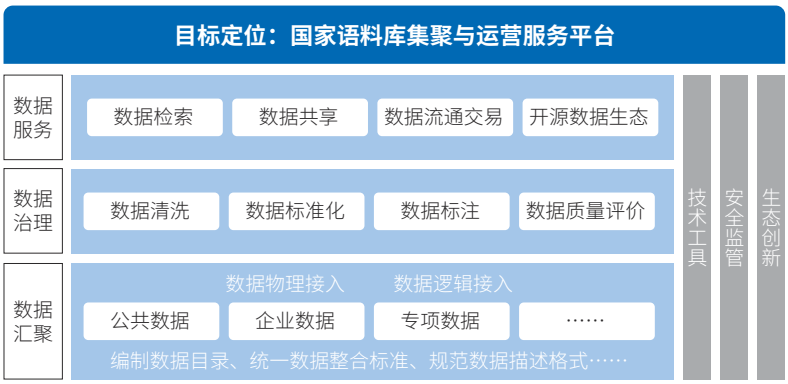

图1　国家级运料库运营平台架构图设想

Figure 1　Conceptual architecture of national material warehouse operation platform

数据汇聚的思路，建议国家级语料库运营平台采用“1+*N*”的一体化架构设计。其中，“1”，指国家语料库运营平台，即中心平台。中心平台负责国家级语料库运营平台的全国统筹管理，建立中心编目系统管理分布式数据平台的元数据，但不直接进行数据治理和数据运营；具体通过制定标准、开源系统工具支撑、开放接口建设等，实现所有平台之间的整体联动和协同共享。此外，中心平台还需负责国家电子政务数据、部委、央企等单位数据的汇聚。“*N*”，指选取部分区域建设*N*个国家级语料库运营平台。例如，支持以国家AI“五大”训练基地所在区域为试点，建设国家级语料库运营平台，负责各区域内的语料汇聚和存储。在“1+*N*”的一体化架构下，基于全国数据互联、服务互通的统一数据门户，中心平台在收到用户请求时，根据元数据描述从分布系统实时调用对应的数据集，形成全国语料库服务“一张网”。

3.3.3 确定运营主体，高效推动平台建设与运营

国家级语料库运营平台的建设运营主体，是影响平台建设进度和成效的关键要素。初步设想，有4种路径：① 由国家数据局统一规划建设统一运营管理，因为在国家数据局等部门印发《“数据要素×”三年行动计划（2024—2026年）》中明确提出建设高质量语料库和基础科学数据集，支持开展AI大模型开发和训练。② 由国家数据局委托国家信息中心、中国信息通信研究院等具有国家信息化项目建设经验的单位开展建设运营，国家数据发展研究院协助建设。③ 以国家数据局为总牵头，协调“东数西算”八大枢纽节点或国家AI“五大”训练基地所在地区发展和改革委员会、经济和信息化厅等相关部门，联合组建国家级语料库运营主体。④ 由国家数据局指导中国移动、中国联通、中国电信等电信运营商进行建设与运营，发挥运营商在数字基础设施、数字化能力及大型信息化项目建设方面所具备的较强优势。

3.3.4 聚焦核心内容，覆盖语料生产应用全生命周期

国家级语料库运营平台覆盖了语料获取、清洗、加工、治理、应用和管理的全生命周期，具有多种灵活的采集、汇聚方式；能分布式高效处理海量语料，有效提升语料开发利用效率，赋能企业或更多机构建设大模型、增强大模型能力。在数据汇聚环节，一方面，保证数据来源，关注公共数据、企业数据等数据来源和获取渠道，兼顾数据在时间和领域维度的融合，建立数据长期更新机制；另一方面，选取合理的数据汇聚方式——公共数据可以考虑以逻辑接入为主，企业数据视情况选择不同汇聚方式。在数据治理环节，既要考虑数据汇聚之后的治理，也要基于不同的场景需求，服务于数据运营需求；考虑采用先进审核技术、动态策略管理等中间层技术，对“有毒”数据进行拦截与修改。在数据服务环节，一方面，积极探索服务内容，平台除主要提供数据目录、数据共享、数据交换、数据工具等服务内容外，还应加强探索合成数据的建设和应用；另一方面，要建立合理的数据运营机制，在明确平台运营主体之后，基于服务内容，科学设定数据定价机制和收益分配机制。

参考文献

1　王文. 全球科技竞争进入“高科技冷战时代”. 中国科学院

院刊, 2024, 39(1): 112-120.

Wang W. Global technological competition enters high-tech cold war era. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 112-120. (in Chinese)

2 Villalobos P, Ho A, Sevilla J, et al. Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv, 2022, doi: 10.48550/arXiv.2211.04325.

3 Villalobos P, Ho A, Sevilla J, et al. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. (2024-03-02) [2025-03-06] https://openreview. net/forum?id=ViZcgDQjyG.

4 《求是》杂志评论员 . 深刻认识和加快发展新质生产力 . 求是, 2024, (5): 39-41.

Commentator for Qiushi Magazine. Deeply understanding and accelerating the development of New Qualitative Productivity. Qiushi, 2024, (5): 39-41. (in Chinese)

Breaking through “data bottleneck” of AI large models

—Reflections on building a national corpus operation platform

LI Xingteng[1*] FENG Feng[2] HUANG Liqiang[3]

（1 School of Public Affairs, Zhejiang University, Hangzhou 310058, China;

2 School of Management, University of Science and Technology of China, Hefei 230026, China;

3 School of Management, Zhejiang University, Hangzhou 310058, China）

Abstract At present, the competition within the global artificial intelligence (AI) large model industry is intensifying, and corpus resources emerging as a critical determinant for enhancing the technical performance and practical efficacy of AI systems. Nevertheless, China’s corpus development faces dual challenges in both quantity and quality, struggling to meet the escalating training demands of the rapidly evolving AI large model sector. Internationally, nations are ramping up efforts to develop their corpus infrastructures, particularly prioritizing the creation and deployment of high-quality linguistic datasets. In this context, through comparative analysis of international benchmarks and domestic conditions, this study proposes a strategic framework for establishing a national corpus management platform. The proposal encompasses four pivotal dimensions: platform orientation, architectural design, governing entities, and key functional components.

Keywords artificial intelligence, large models, corpus, data bottleneck

李兴腾 浙江大学公共管理学院博士后。主要研究领域：数据治理，数字政府。E-mail: lxt23@zju.edu.cn

LI Xingteng Post-doctoral Researcher, School of Public Affairs, Zhejiang University. His research focuses on data governance and digital government. E-mail: lxt23@zju.edu.cn

■责任编辑：岳凌生

*Corresponding author