推动 AI 高质量数据集建设 促进数据资源开发利用

针对我国人工智能大模型训练数据呈现总体量级不足、质量较低、来源匮乏等特点，本文给出了推动 AI 高质量数据集建设的六点建议。

文丨**王闯** 中国电子信息产业发展研究院　**智佳琦** 中国电子信息产业发展研究院

自 2022 年 ChatGPT 问世以来，人工智能大模型正在以前所未有的速度发展，斯坦福大学以人为本人工智能研究院发布的《2024 年人工智能指数报告》指出，在 AI 世界级的竞赛中，中国与美国是最受瞩目的两个国家。然而，美国在 AI 基础大模型发布数量上优势明显，且是全球顶级 AI 大模型的主要发源地，2023 年美国共发布 109 个基础模型，是中国（20 个）的 5 倍多，美国有 61 个知名 AI 大模型，而中国只有 15 个。业界人士认为，算法、算力与数据，是支撑大模型发展的三大基石，我国大模型与美国之间存在差距，除算法模型建构和芯片算力的原因外，更有训练数据不足的问题。

赛迪网官方微信

数字经济官方微信

一、总体概况

我国人工智能大模型训练数据呈现总体量级不足、质量较低、来源匮乏等特点，具体如下所述：

一是中文语料总量占比较低。研究指出，在全球网站中，英文语料占 59.8%，而中文语料仅占 1.3%。一些主流数据集，如 Common Crawl、BooksCorpus、WiKipedia、ROOT 等都以英文为主，最流行的 Common Crawl 中的中文数据也

只占 4.8%。虽然机器翻译可以在一定程度上弥补这部分缺失，但中文大模型训练以英文开源语料为主，将导致其思维偏西方化。

二是中式价值观类语料较为短缺。在中文语料总量占比较低的情况下，机器翻译尚可弥补客观知识类语料的短缺，但在引入主观意志类语料时，却会带来源语言的价值观问题。我国文言文、古汉语、电子书籍、主流媒体等反映了本土价值观的语料有待开发利用，如我国古籍大概 20 万种、50 万版本，但数字化率不到 30%。

三是中文语料总体质量较低。研究指出，未来模型的表现将有 80% 取决于数据质量。阿里巴巴披露，其“通义千问”的中文语料主要来自知乎、百度百科、百度知道等公开网络数据，其中来源于政府的公共数据较少。而高质量的中文数据集主要集中在政府、知名学术机构、媒体机构中，大模型开发者利用高质量中文数据集面临采集难、获取门槛高等问题。

四是我国公共数据供给不足。美国联邦政府对公共数据秉持“应开尽开”的原则，建设并运营维护专门针对 AI 训练数据的开放平台，其中包括各级政府及政府资助的大学和研究机构的数据。而我国公共数据和科研数据的开发利用程度较低。调查显示，虽然我国公共数据开放平台数量从 2012 年的 3 个上升至 2024 年 7 月的 243 个，但开放数据普遍存在数量少、容量低、颗粒度粗、质量不高、数据更新不及时，甚至无故“断供”、平台运营服务不稳定等问题。

五是我国尚未形成社会力量产出开源数据集的生态。美国开源 / 非盈利组织、互联网公司研究部门、学界等社会力量整合公共数据和海量网络数据，加工处理形成以开源为主的高质量数据集。而我国的社会力量则主要结合海外开源数据集及中文语料，产出训练数据集，且开源数据集较少。据 AI 应用开放社区 Hugging Face 数据统计，中文开源数据集数量仅占英文开源的 11%。

二、原因分析

我国中文数据资源损失、分散、孤岛现象严重，线下数据和版权类数据利用不善，公共数据、领域数据开发利用缺乏制度机制保障，中文合成数据技术和经验储备不足等因素，是阻碍我国 AI 大模型中文训练数据集建设的重要原因。具体如下：

一是中文网站数据资源缺乏有效的保存机制。我国没有长期的网页语料积累工作，中文历史网页数据缺少存量保护，损失严重。而英文网页语料却因为社会力量的存在保存了下来，如美国 Common Crawl 自 17 年前以公益的方式在全球不断爬取网页、积累数据，至今已存有 2500 多亿的网页。

二是中文数据资源缺乏有效的互通机制。据 Web Technology Surveys 网站，自 2013 年到 2024 年全球主要网站网页内容语言使用历史趋势，中文网页的数量从 4.3% 下降至 1.3%，下降高达 70%，而同期英文网站比例则从 50.6% 上

升至60.6%。在这十余年间，我国9亿多网民迁移至各类移动互联网平台，而各家移动平台为了建立“数据护城河”，先后主动切断与传统网站网页的数据联通。很多APP对数据获取设置技术障碍，如不支持Web浏览器、不支持游客身份获取数据、屏蔽第三方搜索引擎等。

三是线下数据电子化进程相对滞后。美国对线下数据进行了高度电子化，主要的学术期刊和论文几乎全部实现了在线获取，而我国电子化程度较低，许多图书、期刊和论文等仍主要以纸质形式存在，许多公开出版物无法上网或没有网络版，线下数据难以被充分利用。

四是大模型使用版权类训练语料成本较高。大模型的用数模式与传统的版权类使用方式有所区别，并非是“以欣赏作品原有价值为目的”的利用，也不是对作品内容进行复制、传播，而是为了培训大模型掌握基础智能知识，而相关版权方要求大模型开发者需按版权使用付费，这对大模型训练而言是较重的负担。

五是公共数据开发利用不充分、不持续。各省市的数据开放情况不均衡；开放数据总量规模较小；开放数据以政府数据为主，行业数据较少；政府、企事业单位间的数据标准尚未统一，数据接口错综复杂，开放、共享的数据质量有待提升；数据时效性不强，缺乏常态化的质量反馈和监督机制。

六是公共数据统筹管理体系和能力不适应客观需求。由于数据技术、条块分割体制等限制，我国各级政府部门的数据呈现“孤岛”形态，数据大都处于割裂和休眠状态。各级政府之间尚未形成公共数据开发利用的协同体系和生态，如跨地域、跨层级的公共数据开放平台之间，缺乏有效的互联互通，数据资源的整合度普遍偏低，不同平台数据的开放字段、颗粒度等标准不完全一致，严重制约了公共数据开发利用深度和广度的拓展。

七是领域数据开发利用缺乏完善的机制。领域数据通常由专业部门在从事专门知识劳动中长期积累而来，集中在科研机构、医院、大型网络平台等企事业单位，呈现专业门槛高、积累周期长、数据分散、数据治理不足等特点。例如，科研数据主要分散在各个科学家手中，尚未形成专门的大模型训练科研数据库，也没有专业人士负责科研数据的治理。另外，出于数据安全、数据权属、商业利益、知识产权等多种因素考虑，我国相关权利主体缺乏共享领域数据的积极性。

八是中文合成数据储备不足。据Gartner预测，到2030年合成数据将彻底取代真实数据，成为AI大模型所使用数据的主要来源。美国英伟达、微软、亚马逊等科技巨头纷纷推出合成数据生成工具。与之相比，合成数据在我国发展时间较短，国内大模型企业在合成数据方面的储备不足，缺乏足够的经验和技术积累。

三、工作建议

基于以上分析，我们提出以下建议：

一是促进公共数据的开发利用。出台

国家层面的公共数据开发利用的法律法规和国家标准，鼓励用于大模型训练的公共数据“应开尽开”，明确收益分配机制；依靠制度规定、考核评价、激励机制等推动各级政府部门对数据进行开放、共享或授权运营，构建各级政府部门之间的协同生态，保障数据供给的数量、质量和通畅度，形成具有公共或准公共属性的高质量数据集，建立面向市场的简便的获取程序和条件，推动我国在大模型数据集建设方面的标准化和规范化。

二是鼓励领域数据集的高质量建设。鼓励高校、科研院所和企业间的数据合作与共享，加强数据标注和预处理技术研发，开展数据标注基地建设，提高数据处理效率和质量。加大对领域数据的前期治理、融合开发的人才、资金投入，创新数据治理人才激励机制，建设融合领域知识的垂直大模型高质量训练数据集，促进垂直语料开放共享，探索收益机制，形成“原始数据－高质量数据集－高价值数据集”的发展模式。

三是探索训练数据积累、共享的市场机制。发动社会力量对中文网页数据进行保存，如行业协会、数据基金组织等；鼓励企事业单位共同建设大模型数据空间，研制数据标准，形成可信认证的大模型训练数据流通机制和收益分配机制；利用法律或行业自律的方式推动企业之间数据的互联互通，为大模型发展提供亟须的数据源，实现多方共赢。

四是挖掘线下数据和移动数据。中国国家图书馆是亚洲最大的图书馆，藏书3700万册，主要是中文图书；中文期刊全文数据库收录各类期刊7400种；中文报纸出版种类丰富，仅2019年出版种类就达1851种；截至2023年8月，国内市场上监测到活跃的APP数量有260万个。这些载体上不乏真知灼见。因此加快线下数据电子化进程，推动移动数据挖掘利用和互联互通，是丰富中文语料尤其是中式价值观语料的重要方式。

五是构建版权类训练数据合理使用的制度。目前，已有法律实践在模型训练使用版权作品方面做出突破，如欧盟《单一数字市场版权指令》为符合条件的“文本和数据挖掘”设置了豁免例外，日本对《著作权法》的修订将“不以欣赏作品原有价值为目的”的大模型数据训练纳入到合理使用的范畴。在大模型预训练阶段，我国也可考虑认定利用版权作品进行训练，原则上构成合理使用。

六是鼓励合成数据的发展。合成数据是解决高质量训练数据供给不足的新方案，具有提升数据多样性、加强模型安全性和可靠性、有助于隐私保护等优势。尽管现在对合成数据还存在很多质疑，但是总体上应在设置安全管控策略的前提下鼓励合成数据的发展，加强对合成数据质量的评估检测，为合成数据设置备用的真实世界数据集，对用于模型优化、对齐的合成数据在适当环节引入人类参与。

责任编辑：金烨 jinye@staff.ccidnet.com