

人工智能大语言模型数据集 现状和充实对策研究

胡晓女^{1,2}, 李涛³, 李姗姗³

1. 澳门科技大学, 澳门 999078;
2. 中国通信学会, 北京 100846;
3. 中国联合网络通信有限公司研究院, 北京 100037

摘要

人工智能大语言模型训练数据通常具有数据规模大、数据质量高、数据类型丰富等特点。当前, 国内数据资源虽然丰富, 但优质中文大语言模型训练数据仍然稀缺, 中文大语言模型训练数据的数量与质量, 与全球领先国家的仍存在一定差距。基于国内外公开数据集、典型通用大语言模型数据集的情况, 深入研究对比国内外大语言模型数据集的相关情况, 分析我国大语言模型数据集发展面临的挑战和问题, 提出充实人工智能大语言模型数据集供给的对策建议。

关键词

人工智能; 大语言模型; 数据集

中图分类号: TP393

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025085

Research on the status of large language model data set of artificial intelligence and enriching countermeasure

HU Xiaonyu^{1,2}, LI Tao³, LI Shanshan³

1. Macau University of Science and Technology, Macao 999078, China
2. China Institute of Communications, Beijing 100846, China
3. China United Network Communications Co., Ltd. Research Institute, Beijing 100037, China

Abstract

Artificial intelligence large language model training data usually has the characteristics of large data scale, high data quality and rich data types. At present, although domestic data resources are rich, high-quality chinese large language model training data is still scarce, and there is still a certain gap between the quantity and quality of chinese large language model training data and the world's leading countries. Based on the situation of public data sets and typical general large language model data sets at home and abroad, this paper deeply studied and compared the relevant situation of large language model data sets at home and abroad, analyzed the challenges and problems faced by the development of large language model data sets in China, and put forward countermeasures and suggestions to enrich the supply of large language model data sets in artificial intelligence.

Key words
artificial intelligence, large language model, data set

0 引言

截至2024年7月,全球人工智能大语言模型(以下简称大模型)数量已达1 328个,其中国外大模型数量为732个,国内大模型数量为596个。美国拥有的大模型数量位列全球第一,中国拥有的大模型数量占比为36%,位列全球第二。国外主要的大模型包括GPT、Claude、Gemini、LLama等,参数规模从十亿个到万亿个,相应的数据集规模至少为参数规模的20倍,大模型训练成本较高^[1]。例如,GPT-4的模型参数在1.8万亿个左右,一次训练的成本为6 300万美元^[2]。大模型数据集是大模型技术的基础^[3],其发展水平是衡量一个国家科技实力的重要标志。全球各国加速数据战略部署,在技术、政策和市场等多因素驱动下,国内外大模型数据集呈现出不同的发展态势。我国受国内产业环境、数据积累程度、数据处理技术、运营生态等多维因素影响,与领先国家仍存在一定差距,大模型难以拥有足够专业的数据资源进行训练。本文通过对国内外大模型数据集数据基础制度和机制、数据来源和供给渠道、数据处理工程化、数据质量评估、数据应用场景挖掘等方面进行研究梳理,提出充实数据集供给的建设性建议。

1 国外大模型数据集发展情况

1.1 国外典型大模型数据集情况

欧美国家具备深厚的人工智能学术研

究基础,为大模型数据集的开发和应用提供了理论支持,在数据多样性、数据质量和数据开放方面处于领先地位。国外典型大模型数据集情况^[4-21]分析见表1。

1.2 国外典型大模型数据集特点

- (1) 国外大模型数据集来源多样
- 欧美国家的学术机构、企业和公共部门为大模型提供了大量、丰富、广泛的数据集。本文通过对国外通用大模型使用的公开数据集进行不完全统计,得到目前质量比较高的大型公开数据集分布状况:文本类有49个、图片类有35个、语音类有48个、视频类有57个、多模态训练数据集有48个、评测数据集有44个,涵盖从自然语言到科学数据等多个领域,特别是在科技、医学、社交网络等领域形成了高质量的大型公开数据集。
- (2) 欧美国家数据质量普遍较高
- 欧美国家建立了严格的数据采集规范和审核机制。例如,欧盟制定了统一且严格的数据采集框架指令,各成员国的统计机构需按照该指令开展数据收集与统计工作,同时在数据审核方面,要求企业内部建立数据保护官制度,负责监督数据采集并对数据隐私及合规性进行审查。另外,许多企业和科研机构通过设立奖项和资金支持,激励数据集的质量优化和公开共享。
- (3) 欧美国家公共数据开放较充分
- 欧美国家建立了较完备和宽松的公共数据开放政策,颁布相关法律框架确保数据开放和用户隐私保护的平衡。以美国Data.gov网站为例,其是美国政府在2009年推出的开放数据网站,要求联邦机构以标

表1 国外典型大模型数据集情况

名称	基本情况	功能特点	应用领域	数据集类型、来源	规模与 Token 数量
OpenAI-ChatGPT	一款大型语言模型,于2022年11月正式发布	强大的语言理解能力、出色的文本生成能力、对话式交互风格	日常聊天、知识问答、各类文本内容的创作、智能客服	维基百科、书籍、期刊论文、社交媒体链接数据、CommonCrawl(网站抓取数据集)、其他数据集(代码数据集 Github 等、对话论坛 Stack - Exchange 等和视频字幕数据集等)	作为其基础的 GPT-3 的 CommonCrawl 数据集包含 4 100 亿个 Token(实际应该比这个数大)
Anthropic-Claude	人工智能创业公司 Anthropic 研发的一款大型语言模型,Claude 3 于 2024 年 3 月推出	强大的语言处理能力、多语言支持、视觉处理能力、代码生成与辅助	日常办公、学术研究、软件开发、客户服务	互联网公开抓取数据、来自第三方的非公开数据、数据标注服务、付费承包商提供的数据、Anthropic 内部生成的数据、基准测试数据集、维基百科	Claude 1 上下文窗口约有 75 000 个 Token;Claude 2 有 10 万个 Token;Claude Instant 有 75 000 个 Token; Claude 3 Haiku 每秒可处理 21 000 个 Token;Claude 3 Opus 有 100 万个 Token
Google-Gemini 多模态大模型	谷歌推出的多模态大模型	多模态信息处理能力、强大的推理能力、出色的代码处理能力、高效的移动端适配性	智能办公、科学研究、智能营销、软件开发、内容创作	网络文档、书籍、代码、图像、音频和视频数据	Gemini 接受训练的 Token 数量为 GPT-4 的两倍。 Gemma-2b 的预训练数据量为 2 万亿个 Token, Gemma-7b 的预训练数据量为 6 万亿个 Token。 Gemini 的预训练数据量在数万亿个 Token
Google-PaLM2	谷歌研发的一款大型语言模型	强大的多语言能力、出色的逻辑推理能力、优秀的编程能力、模型高效性、多版本选择	人工智能、教育、医疗、企业、自然语言处理	网络文档、书籍、代码、数学、对话数据	PaLM2 采用了 3.6 万亿个 Token 的数据进行训练。其前一代 PaLM 的训练数据约为 7 800 亿个 token, PaLM2 的训练数据量是其前一代的 5 倍左右
Meta-LLama-2	Meta AI 在 2023 年 7 月发布的大型语言模型	模型架构改进、语境长度扩展、数据与训练优化、安全性提升、语言能力较强	自然语言处理任务(智能聊天机器人、文本生成、语言翻译)、学术研究、软件开发、智能终端应用	来自公开在线资料的数据集, Clue、ThucNews、CNews、Coig、维基百科, 古诗词、散文、文言文等文学作品数据集	LLaMA-2 的预训练数据使用了 2 万亿个 Token。其前身 LLaMA 的预训练数据约为 1 万亿个 Token
Mistral-Mistral Large	法国人工智能初创公司 Mistral AI 推出的一款先进的文本生成模型	强大的推理能力、多语言支持、长文本处理能力、精准的指令执行、函数调用支持	自然语言处理相关应用开发、文本创作与编辑、代码开发、学术研究与知识管理、多语言翻译和本地化	多语言文本数据集(大量的网页文本、学术文献、书籍和出版物)、代码数据集(各种编程语言的开源项目代码、代码库、编程教程等)、社交媒体和对话数据集	Mistral Large 拥有 1 230 亿个参数,使其在代码生成、数学和推理方面比其前身功能更强大
Amazon-Titan	亚马逊云服务中的重要组成部分	多种模型类型(文本生成模型、嵌入模型、高质量的响应、可定制性)	内容创作(媒体、广告、营销等,快速生成文章、广告文案、宣传语等各种文本内容)、智能客服、代码开发、企业搜索与推荐、金融领域	新闻文章(包括政治、经济、科技、文化、体育等)、学术文献(各个学科领域研究成果、理论探讨和专业知识等)、小说文学作品、社交媒体数据、代码数据集、企业内部数据	2023 年,亚马逊使用 13 760 块英伟达 A100 芯片,花费 48 天时间,在 4 000 B 的 Token 数 据上训练了一个 200 B 参数的 Titan 模型

准化格式在线发布开放数据，从数据主题、格式、标签、组织和发布者等视角构建统一数据集资产目录，管理来自近 100 个组织的 30 多万个数据集和数据集集合，并配套发布数据管理与治理的指南、策略，提供统一的数据集处理工具、数据项目孵化和人才技能发展资源。Data.gov 网站进行公共数据开放，为研究人员和企业提供了获取高质量数据集的渠道，网站每月浏览量超 100 万次。

2 我国大模型数据集发展情况

我国庞大的人口基数及发达的互联网应用，为大模型训练提供了海量数据。此外，在政策推动下，智慧城市、智能交通等垂直领域大数据应用场景逐渐丰富。虽然国内数据规模和应用场景优势显著，但大模型数据集相较国外仍存在差距。

2.1 国内通用大模型

据不完全统计，国内具有代表性的通用大模型超过 200 个，包括智谱清言、盘古大模型、阿里巴巴发布的通义千问大模型、百度在线网络技术（北京）有限公司（以下简称百度）发布的文心一言大模型、北京百川智能科技有限公司（以下简称百川智能）发布的百川系列大模型、科大讯飞股份有限公司（以下简称科大讯飞）发布的星火大模型、北京字节跳动科技有限公司（以下简称字节跳动）发布的豆包、腾讯混元大模型、北京市商汤科技开发有限公司（以下简称商汤科技）发布的商量大模型、DeepSeek 发布的大模型等。截至 2024 年年底，我国有 5%~8% 的企业大模型参数实现从千亿级向万亿级的跨越性

跃升。

（1）智谱清言

智谱清言 GLM-4 新一代基座大模型，支持 128 K 上下文，具备理解和规划复杂指令、完成复杂任务的能力。其在不影响任何自然语言处理（natural language processing，NLP）任务性能的情况下，实现了视觉语言特征的深度融合。

（2）盘古大模型

盘古大模型包含五大基础大模型，包括 NLP、计算机视觉、多模态、预测和科学计算大模型，覆盖了结构化数据和非结构化数据的处理。盘古大模型使用的数据集基于海量图像、视频数据以及独特的技术构建，融合了跨模态信息。

（3）阿里巴巴发布的通义千问大模型

通义千问大模型采用开源大语言模型 LLama 的数据训练方法，具有长文本功能，向所有人免费开放 1 000 万字的长文档处理功能。其在数据训练中使用了 TB 级的训练数据和亿级的参数，以提高预测和推理能力。

（4）百度发布的文心一言大模型

文心一言大模型具有中文领域先进的自然语言处理能力，尤其在中文语言和中国文化上相比其他大模型有更好的表现。其能够正确解释成语、创作诗歌，并且具备多模态生成能力，能够生成文本、图片、音频和视频。

（5）百川智能发布的百川系列大模型

百川系列大模型提供较高性价比的 MoE 模型，基于工程架构的技术创新进行业务落地的优先选择，其针对金融、法律、教育、客服、销售等多个重点行业和领域进行专项优化，形成面向不同行业的大模型。

（6）科大讯飞发布的星火大模型

星火大模型具有超过 1 000 亿个参

数，使用超过1 000亿字的中文文本数据进行训练，涵盖新闻、百科、小说等多个领域，有效地捕捉中文语言的复杂性和多样性。

（7）字节跳动发布的豆包

豆包通过对大量文本等数据进行训练，能够理解问题意图并生成回答。其优势在于能够快速、准确地提供信息，并且可以灵活应对不同类型的NLP任务，为用户在知识获取、文案写作等多个场景提供便利。

（8）腾讯混元大模型

混元大模型由腾讯公司全链路自主研发，在高质量内容创作、数理逻辑、代码生成和多轮对话上的性能表现卓越。其支持人工智能搜索联网插件，整合腾讯优质的内容生态，提供强大的时新、深度内容获取和人工智能问答能力。

（9）商汤科技发布的商量大模型

商量是千亿参数大模型应用平台，拥有语义理解、多轮对话、知识掌握、逻辑推理的综合能力，能理解和处理中文文本，可自主反思及修正错误等。

（10）DeepSeek发布的大模型

DeepSeek 是一家创新型科技公司，专注于开发先进的大模型和相关技术。DeepSeek-R1、V3、Coder 等系列模型，拥有强大的NLP能力，能够理解并回答问题，还能辅助写代码、整理资料 and 解决复杂的数学问题。与 OpenAI 开发的 ChatGPT 相比，DeepSeek 不仅率先实现了等同 OpenAI-o1 模型的效果，还大幅降低了推理模型的成本。其新模型 DeepSeek-R1 以十分之一的成本达到了 GPT-o1 级别的表现，引发海外人工智能圈的广泛讨论。DeepSeek 最大的优势在于算法的改进和优化，节省了算力和数据量。

2.2 国内行业大模型

国内行业大模型的发展正聚焦于技术进步和广泛应用，模型规模和复杂度的增加、多模态整合能力的加强、自监督学习的兴起、可解释性与公平性的关注、部署策略优化及特定领域定制化成为主要发展趋势。

2.2.1 电信行业

中国移动通信集团有限公司（以下简称中国移动）、中国联合网络通信集团有限公司（以下简称中国联通）、中国电信集团有限公司（以下简称中国电信）三大电信运营商在大模型发展上具有数据资源丰富和应用场景广泛的优势，分别发布了在技术架构、行业应用和数据处理方面各有侧重的大模型产品体系。同时，3家电信运营商通过汇聚企业内部、外部开源、行业合作、外部商用等多类来源端数据，按模态、按训练过程、按使用对象进行数据设计分类和清洗标注，形成高质量数据集，共同推动了人工智能技术在通信行业的深入应用和产业数字化转型。

（1）中国移动发布的“九天”人工智能基座大模型

“九天”人工智能基座大模型构建了通/专大模型体系，在网络、客服、营销、知识管理和反诈安全等运营场景驱动业务服务敏捷创新，为客户提供智慧的信息通信服务。同时，其在能源、交通、医疗、政务等行业加快“AI+”深度融合赋能，如九天·网络大模型实现网络规划、建设、运维、优化、运营全流程智能化升级，加快向“AI+网络”转型；九天·客服大模型已在10086在线客服场景规模化应用，显示出其在超大规模客服生产系统中的工程化能力；九天·医疗大模型已在北京协

和医院、中日友好医院等医疗机构落地部署；中国移动与中国石油天然气集团有限公司联合打造昆仑大模型，助力能源化工行业AI+转型。

(2) 中国联通发布的元景大模型

元景大模型构建了“1+1+M”大模型体系，包括一套基础大模型、一个大模型底座和多种行业大模型。中国联通发布的大模型在网络、客服、反诈、工业、政务等行业实现落地应用，如元景热线提供即插即用智能工具包，助力电信客服、政务、企业热线实现降本增效。此外，中国联通发布的鸿湖图文大模型1.0具有文生图、视频剪辑、以图生图等功能，广泛应用于个性化推荐、广告创意、内容创作等领域。

(3) 中国电信发布的星辰大模型平台

星辰大模型平台包括星辰语义大模型、星辰语音大模型等，支持多种方言识别理解。中国电信发布的大模型在经营分析、公文写作、电信客服、云网运营等场景探索智慧应用，如星辰大模型在政务客服、教育、医疗卫生等领域实现行业赋能。中国电信发布的大模型还依托5G消息为政务客户提供政务短信智能化多轮问答服务，显示出其在特定行业的应用潜力。

电信行业是数据密集型行业，其业务覆盖范围广、用户基数大，产生了丰富的多维度数据。典型行业的数据集有通话记录数据，包括用户的通话时间、频率、通话对象等，用于网络优化和用户行为分析；网络流量数据、互联网流量数据，包括流量类型、流量峰值、数据包信息等，支持网络性能优化和流量管理；基站数据，包括基站位置、覆盖范围、信号强度、用户接入情况等，帮助优化基站布局和网络覆盖；客户服务数据，包括客户投诉、故障报告、服务记录等，用于提升客户服务质量和满意度。

2.2.2 其他典型行业

国内在医疗、金融、交通、电力和教育等典型行业积极推动大模型技术的应用和发展，显著提升了各领域的智能化和效率。

(1) 医疗行业

在医疗行业，大模型被广泛应用于图像识别、疾病预测与诊断等领域。大模型通过对海量医疗数据进行分析，帮助医生更准确地进行诊断和治疗决策。例如，商汤科技的“大医”和科大讯飞的星火医疗大模型通过分析海量医学数据，提升了医疗问答、辅助诊断和治疗推荐的准确性和效率；百度发布的灵医大模型在互联网医疗、公立医院等领域实现商业化落地。医疗行业的数据集包括医学影像，如数字X成像（X光片）、磁共振成像、计算机断层（computed tomography，CT）扫描图像等；电子健康记录，如病历、诊断记录、治疗方案等；生物信号，如心电图、脑电图、血氧数据等。

(2) 金融行业

金融行业的大模型应用主要集中在风控、智能投顾和客户服务等方面。通过对交易数据和用户行为的分析，大模型能更好地识别潜在风险，防范金融欺诈^[22]。同时，智能投顾系统基于大模型分析为用户提供个性化的投资建议，提升了用户理财体验。金融行业的数据集包括交易数据，如股票交易记录、期货交易数据等；财务报表，如公司年报、季报、财务报告等；市场数据，如市场行情、宏观经济指标等。

(3) 交通行业

在交通领域，大模型促进了智能交通系统的发展。通过交通数据的实时分析与预测，大模型能够优化交通信号控制、规划最优出行路线，从而有效地缓解城市交通拥堵状况，提高整体交通效率。交通行

业的数据集包括车辆数据，如车辆行驶记录、GPS 轨迹数据等；交通流量数据，如道路交通流量、拥堵情况等；公共交通数据，如公交/地铁运营数据、乘客流量等。

(4) 电力行业

电力行业通过大模型实现了电网智能化管理。大模型帮助预测电力需求波动，优化电力调度和分配，提高了电网的可靠性和可持续性，同时支持新能源的高效集成。电力行业的数据集包括电网运行数据，如电网负荷、发电量、输电线路状态等，帮助监控和优化电力系统的运行；用户的用电记录，如家庭、商业和工业用电数据，支持用电模式分析和需求预测；设备状态数据，如电力设备（如变压器、断路器）的运行状态、故障记录等，支持设备维护和故障诊断；与电力生产和传输相关的气象数据，如温度、风速、降雨量等，帮助预测可再生能源发电量（如风电、光伏）。

(5) 教育行业

在教育领域，大模型应用于个性化学习和智能教育管理。通过大模型分析学生的学习行为和表现，教育平台可以提供定制化的学习资源与建议，有力支持因材施教，提升教学效果。教育行业的数据集包括学习行为数据，如在线学习记录、考试成绩、作业提交情况等；课程资源，如教学视频、教材、课件等；师生互动数据，如课堂互动记录、师生交流记录等。

2.3 国内数据集特点

国内大模型数据集的多样化、特定化和丰富场景，严格的隐私安全和多模态支持，构成了我国在大模型领域创新与突破的坚实基础。这些特点不仅提升了大模型的适用性和实用价值，也在全球人工智能竞赛中塑造了我国的独特优势。

(1) 国内数据集主要来源于公众和行业
国内数据集不仅包括来自公有领域的信息，如互联网开源数据、社交媒体内容等，还包括行业应用中的私有数据，如医疗记录、金融交易数据和交通监控数据等。公众领域的信息开源程度较高，行业私有数据一般较封闭。

(2) 国内数据集的语言以中文为主

中文语言具有复杂性，包括丰富的语句结构、同音异义字、成语和谚语等，专注于中文的大模型需要能够处理这些细节。国内大模型数据集在语料采集上更加注重覆盖不同的语言习惯、方言与地域差异，但具有中华文化特色的数据还不够丰富。

(3) 垂直行业大模型的数据集发展速度较快

国内垂直行业大模型应用广泛，针对不同行业的需求，这些数据集被专门设计以应用于通信、医疗、金融、教育、电力、交通等领域。各行业特定的术语、格式和知识体系被仔细整合进这些数据集，以确保大模型能有效地解决实际问题。

(4) 国内大模型对多模态数据需求愈发增长

在智慧城市和自动驾驶汽车等需要综合多种信息来源的领域中，数据集不仅局限于文本数据，还包括图像、音频和视频等多模态数据。多模态数据的使用能够增强大模型理解真实世界的能力，支持更复杂的人机交互和决策任务。

3 我国大模型数据集发展面临的挑战

3.1 中文数据规模总量有待挖掘

据研究公司 Epoch AI 预测，人类生成的公开文本数据的总有效存量约为 300 万亿个 Token。Epoch AI 在 2025 年发布的

论文“Will we run out of data to train large language models?”中预测，未来10年内，数据增长的速度无法支撑起大模型的扩展，大模型会在2028年用完互联网上的所有文本数据。相比英文语料数据，大模型训练数据集的中文数据规模不足问题尤为凸显。阿里巴巴2024年5月发布的《大模型训练数据白皮书》显示，全球网站中文数据仅占1.3%，而英文数据占比高达59.8%。这表明在大模型可获取的基础数据中，中文数据在数量上相比英文数据存在较大差距。

为解决大模型数据缺失的问题，行业认为合成数据可能成为解决该问题的新思路。合成数据通过对现有数据进行深加工，将不能被用于训练的数据转化为可用于训练的数据，提升大模型对数据利用的可能性。但当前大模型仍面临无法完全模拟真实世界的复杂性、缺乏可解释性和可信度、在新场景中的泛化能力不足等问题，需要进一步加强相关技术研究和论证^[23]。

中华文化具有几千年的历史，在几千年的文化传承中，沉淀了很多优秀的中文宝库，如《四库全书》、《康熙字典》、四书五经、《资治通鉴》等。没有证据证明，这些具有中国传统文化特色的中文文本已经被大模型训练语料库收集，并被加工处理成大模型训练数据集。

3.2 体现本土文化和价值的数据集亟待丰富

国内外大模型训练所用数据主要来自互联网、电商、社交、搜索等渠道，中文数据集也主要来自知乎、百度百科、百度知道等公开网络数据，尽管互联网上中文文本存量数据资源丰富，但仍然存在以下问题。

（1）在大模型训练中，承载中华优秀传统文化和价值观的数据集使用比例不够

大模型训练数据集中不同类型数据比例的使用，直接决定了大模型的“性格基因”和“文化基因”。一方面，国内的很多大模型以国外开源大模型为基础进行开发，国外的开源大模型在训练时，中文语料的投喂比例是不够的。另一方面，即便是从头训练的国内大模型，由于数据加工的成本巨大，耗时冗长，大多数公司仅从公开可得的渠道获取可用的中文语料库，很难保证数据足够。

（2）体现本地科学技术和经济特色的数据集缺乏^[24]

由于各种原因，国内科研数据、政府部门数据、反映本地经济运行的数据、大型企业掌握的行业数据等被加工成大模型训练的数据集数量不详。仅从公开可得的少数信息判断，中文语料数据总量规模明显不足。例如，悟道语料库从包括开源开放数据集、Common Crawl网页数据、电子书等的不同渠道获得规模为80 TB的文本、图文和对话数据集，通过数据清洗过滤、去重、质量评估等处理流程，构建了一个规模约为1.1 TB的高质量中文语料数据集；开源中文本数据集CLUECorps的规模为100 GB。相比之下，GPT-3的训练数据以英语为主，数据集规模达45 TB。此外，中文语料还存在本地化数据来源单一、可供大模型训练的有效数据源呈现碎片化分散状态的问题，至少政府部门的公共数据标准不规范，机器直接调用的技术支撑仍不到位。以国家统计局为例，大量数据分布在网站群中，需要人工收集。

大模型训练在语言模型和文化相关数据方面存在差异会影响模型的泛化能力，中文数据语料的缺失会影响大模型对中华文化的理解和表达能力。例如，在ChatGPT训练数据中，中文资料比重不足千分之一，而英文资料占比超过92.6%。

3.3 大模型数据集质量管理体系需要完善

构建高质量数据集通常涉及严格的数据清洗、去重和标注流程，以减少噪声对模型训练的干扰，提升预测准确性^[23]。在大模型数据集质量方面，国内缺乏统一的管理框架和制度规范，且高质量中文数据集缺乏，部分中文大模型采用“英文数据集+翻译软件”的方式生成中文语料库，导致训练结果不佳。此外，在大模型数据集质量评价方面，国内缺乏统一规范的评测体系，导致不同评测结果难以精确对比，阻碍大模型的落地应用。

3.4 公共数据开放机制有待健全

国内公共数据开放起步较晚。在观念上，公共部门工作人员更加偏重保守秘密；在考核上，国内每年均进行保密检查，相对严格，且对公开数据的考核不明确。近年来，随着“数据要素”深入人心，各级政府陆续推动公共数据平台建设，提高数据的开放性和可用性。但标准与规范尚不健全，可操作指南较少，特别是专业人才缺乏。2024年，《中共中央办公厅 国务院办公厅关于加快公共数据资源开发利用的意见》发布，首次对公共数据资源开发利用进行系统部署。公共机构开放数据集的开放利用面临机制体制、标准规范、平台统一、行政手续等挑战，影响了公共数据集的流通和利用。

4 人工智能大模型数据集充实对策建议

4.1 加大数据产业和技术发展支持力度

(1) 加强专项政策支持和研发资金支持

加强在数据要素市场化配置、数据产权制度、数据交易等领域的政策措施来推动数据产业的创新和应用^[25]。同时，设立专项资金，支持大模型数据处理相关技术的研发和应用，鼓励企业和科研机构进行创新。

(2) 优化数据产业发展环境

完善数据流通和交易的法律法规，建立健全数据安全和隐私保护机制，营造公平竞争的市场环境。加大基础设施建设，提升算力和数据存储能力，支持产业链上下游的协同发展。依托国家重点研发计划、国家科技重大专项等，开展数据加密、可信流通、安全治理等关键技术研究 and 攻关。

(3) 支持国内数据开源社区和国家公共数据开放项目

积极资助和鼓励数据开源社区建设，参照支持开源软件项目的政策，支持数据开源项目。制定数据开源标准和评估体系，提升国内开源数据质量和应用水平。推动企业和高校参与开源数据项目，促进技术共享与合作创新。

(4) 支持培育数据标注产业

制定产业扶持政策，建立数据标注职业技能培训体系和职业资格认证制度，建立数据标注行业标准和规范。鼓励和支持数据标注技术创新，研发自动化标注技术、标注工具，提高标注效率。搭建数据标注公共服务平台，整合数据标注资源。

4.2 拓展中文大模型数据集规模

(1) 深入挖掘传统文化数据

收集大量与中华文化相关的图像、音频、视频等多模态数据，深入挖掘中华优秀传统文化数据。整合现有资源，如从图书馆、档案馆、博物馆等文化机构的文献资料、古籍善本、历史档案中提取有价值的文本

信息，将其数字化后纳入大模型数据集。

（2）促进科研数据共享

加强对英文科技数据的翻译整理，鼓励高校、科研机构将科研过程中产生的实验数据、研究报告、学术文献等数据进行共享，建立科研数据共享平台，制定数据共享的标准和规范，促进科研数据在不同机构和研究团队之间的流通和利用，为大模型训练提供更多专业领域的知识和数据支持。

（3）推动政府、行业、大型企业的公共数据开放

各级政府部门应进一步梳理和整合所掌握的公共数据，包括但不限于人口、地理、气象、交通、医疗、教育、政务服务等领域。在保障数据安全和隐私的前提下，向科研机构、企业和社会公众开放更多高质量的公共数据。鼓励同行业的企业成立数据联盟，共同制定数据共享的规则和标准，将各自分散的数据进行整合和共享。推动大型企业开放公共数据，提升数据治理能力，探索数据流通机制。

（4）鼓励技术融合与创新

在收集处理后的高质量中文种子数据基础上，将数据合成技术与数据增强技术相结合，丰富数据的表现形式和多样性，并探索不同领域数据生成方法的迁移，实现更广泛的数据共享。引入人工审核环节，对生成的数据进行抽样检查和校正；严格遵守相关的法律法规，确保数据合成过程的合法性和合规性；制定相应的伦理准则和规范，引导数据合成技术的合理应用。

4.3 开展大模型中华文化价值度量

（1）构建中华文化价值评测数据集

明确以社会主义核心价值观为核心，涵盖国家、社会、个人层面的价值观，同

时兼顾中华优秀传统文化价值观等，如爱国、敬业、诚信、友善以及尊老爱幼、尊师重道、和谐中庸等。基于收集的文本和实际应用场景，设计多种类型的测试题目，其应涵盖不同的价值观维度和具体情境，以全面考察大模型对中华文化价值观的理解和应用能力^[27]。

（2）构建中华文化价值观评测指标体系并开展评测

从准确性、完整性、一致性、敏感性等方面，构建评测指标体系，采用自动评测与人工评测相结合的方法，从内容生成、问答交互、对抗性等多维度进行评测。可结合对中文大模型的综合评测情况形成达标分值，对于评测结果低于达标分值的大模型不授予认证许可。

（3）研究大模型中文数据集与中华文化价值观的相关性

针对大模型中文数据集比例与中华文化价值观的相关性，组织行业专家学者提出诱导偏见、歧视回答的问题，并构建文化价值评估数据集，探索基于专家原则的对齐研究，制定价值对齐评测基准^[28]。

4.4 推动公共数据的开放和治理

（1）政府机构率先示范，建立高水平公共中文数据集

在国家层面组织专家团队，尽快组织实施政务数据分类分级工作。在《中华人民共和国保守国家秘密法》、《中华人民共和国政府信息公开条例》、保护商业秘密和个人信息的框架下，组织开展政府机关掌握的公共数据的分类分级工作，将政府机关掌握的公共数据分成国家秘密数据不予公开；涉及企业商业秘密和个人隐私的数据不应公开；涉及政府工作的敏感信息，如国家利益相关的战略数据、国家安全数

据、与国家经济发展相关的敏感数据、行业发展数据、与百姓生活密切相关的个人信息数据（涉案调查、刑事侦查）等，暂缓公开。对于上述不作公开的政府机构掌握的数据要明确办法，广而告之。其余信息能公开尽可能公开。

(2) 建立全国统一的数据平台、标准与工具体系

建设国家级公共数据平台，整合各地区、各部门的数据资源，实现数据的集中管理和共享交换，打破“数据孤岛”，提高数据的可用性和价值。制定统一的数据标准，包括数据格式、编码规则、数据质量标准等^[29]，确保不同来源的数据能够相互兼容和整合，提高数据的一致性和准确性，为数据的开放和治理提供基础支撑^[30]。构建全面且分类清晰的数据集目录体系，如按主题、部门、数据格式等多维度分类，方便用户精准定位所需数据^[31]。提供工具来简化数据的发布、共享、查找和使用。依据国家标准《数据安全技术 数据分类分级规则》（GB/T 43697-2024），优先开放数据级别为一般数据的公共数据，逐步扩大数据开放范围^[32]。

(3) 强化监督与评估机制

建立数据开放监督机制，加强对公共数据开放过程的监督，确保数据开放的合法性、公正性和透明度。建立投诉和举报渠道，鼓励社会公众对数据开放中的问题进行监督和反馈。定期对公共数据治理的效果进行评估，评估指标可包括数据开放的程度、数据质量的提升程度、数据应用的创新程度等，根据评估结果及时调整和改进数据治理策略和措施，不断提高数据治理水平。

(4) 完善法律法规与制度建设

建立健全公共数据产权制度，明确公共数据的归属、使用、管理等权利和责任，

为数据开放和治理提供清晰的法律依据，解决数据产权不明晰导致的开放和利用难题。细化数据安全相关法律法规的实施细则，明确公共数据开放过程中不同主体的安全责任和义务，确保数据在开放和治理过程中的安全性。

4.5 建立数据质量评估和激励机制

(1) 加强数据质量管控与审核机制

制定公共数据统一且严格的数据质量管理框架和工具，按数据集质量管理流程，对数据质量进行严格审核。数据集开放机构内部建立数据质量管理体系规范，负责监督数据采集并对数据隐私及合规性进行审查^[33]。

(2) 完善数据质量评估标准体系

从准确性、完整性、一致性、时效性、唯一性、可解释性、简洁性、可访问性、合规性等多个维度构建数据质量评估指标和方法，定期对公共数据质量进行评估，保证数据的真实性和可靠性。同时，探索研究评估指标与数据集质量之间的多变量和单变量相关性，通过挑选特定微调数据集，检验微调后的大模型与基准大模型相比的性能变化^[28]，以进一步验证数据集质量对模型性能的影响。

(3) 建立公平合理的激励政策

对提供高质量数据的机构、部门或个人给予正向激励，如奖励、表彰等。加强对中文高质量评测数据集的激励。

4.6 完善数据人才培养体系

(1) 优化大数据学科设置，政产学研合作协同育人

鼓励高校优化大数据学科专业设置，推进政产学研合作协同育人。高校应根据市场需求和技术发展动态，调整和优化大

数据相关专业设置，增加跨学科课程。同时，建立政产学研合作机制，鼓励高校与企业、研究机构共同开发课程和项目，推动实践教学，培养具有实际操作能力的人才。

(2) 与企业合作开展专业培训和认证

与企业合作，设计针对性的培训项目，提升企业员工在大数据处理和分析方面的能力。通过与行业领军企业合作，联合开发认证课程，促进企业与人才的良性互动。

(3) 鼓励利用大模型技术提升人才数据能力项目实践

高校和企业应共同探索基于大模型的实践项目，鼓励学生和员工参与实际数据分析和模型训练，提升其数据处理能力和应用水平。通过实习、项目实践和竞赛等形式，培养学生的创新思维和解决实际问题的能力，推动大数据技术在各行业的应用。

5 结束语

随着人工智能大模型技术的飞速发展，其在诸多领域展现出了强大的能力。大模型的性能高度依赖数据集的质量、规模和多样性，数据集作为大模型的“燃料”，是其实现卓越表现的基础。当前，大模型发展进入多模态融合阶段，对大规模、多模态、多领域的高质量数据集的需求愈发迫切。本文深入了解人工智能大模型数据集现状，探讨充实对策及后续研究方向，对于推动人工智能技术持续进步具有重要意义。

人工智能大模型技术领域后续亟须研究的方向有高质量数据集建设研究、高质量数据集评测体系研究、数据隐私保护与安全共享技术拓展研究、面向特定领域的数据集构建与应用研究、数据要素市场与数据集价值实现机制研究等。

参考文献：

[1] 李国杰. 大数据与计算模型[J]. 大数据, 2024, 10(1): 9-16.
LI G J. Big data and computing models [J]. Big Data Research, 2024, 10(1): 9-16.

[2] LIU Y, CAO J H, LIU C Y, et al. Datasets for large language models: a comprehensive survey[EB]. arXiv preprint, 2024, arXiv: 2402.18041.

[3] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[EB]. arXiv preprint, 2021, arXiv: 2108.07258.

[4] ZHANG S S, ROLLER S, GOYAL N, et al. OPT: open pre-trained transformer language models[EB]. arXiv preprint, 2022, arXiv: 2205.01068.

[5] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[EB]. arXiv preprint, 2023, arXiv: 2302.13971.

[6] ZENG W, REN X Z, SU T, et al. Pangu- α : large-scale autoregressive pretrained chinese language models with auto-parallel computation[EB]. arXiv preprint, 2022, arXiv: 2104.12369.

[7] ZENG A H, LIU X, DU Z X, et al. GLM-130B: an open bilingual pre-trained model[EB]. arXiv preprint, 2022, arXiv: 2210.02414.

[8] WORKSHOP B, LE SCAO T, et al. BLOOM: a 176B-parameter open-access multilingual language model[EB]. arXiv preprint, 2022, arXiv: 2211.05100.

[9] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners[EB]. arXiv preprint, 2021, arXiv: 2109.01652.

[10] ZHU D Y, CHEN J, SHEN X Q, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models[EB]. arXiv preprint, 2023,

- arXiv: 2304.10592.
- [11] TAORI R, ISHAAN G, ZHANG T Y, et al. Alpaca: a strong, replicable instruction-following model[EB]. arXiv preprint, 2023, arXiv: 2302.03846.
- [12] HENDRYCKS D, BURNS C, CHEN S, et al. Measuring the alignment of language models with human preferences[EB]. arXiv preprint, 2023, arXiv: 2306.03797.
- [13] BIDERMAN S, SCHOELKOPF H, ANTHONY Q, et al. Pythia: a suite for analyzing large language models across training and scaling[EB]. arXiv preprint arXiv: 2304.01373.
- [14] WANG C, KORDI Y, MISHRA S, et al. Ultrachat: towards a unified chat model with ultra-long context[EB]. arXiv preprint, 2023, arXiv: 2307.15868.
- [15] TAYLOR R, KARDAS M, CUCURULL G, et al. Galactica: a large language model for science[EB]. arXiv preprint, 2023, arXiv: 2211.09085.
- [16] WANG Y Z, MISHRA S, ALIPOORMOLABASHI P, et al. Super-natural instructions: generalization via declarative instructions on 1600+ NLP tasks[EB]. arXiv preprint, 2022, arXiv: 2204.07705.
- [17] KOCETKOV I, FEDOROV A. Less is more: data selection for instruction tuning of large language models[EB]. arXiv preprint, 2023, arXiv: 2402.04333.
- [18] ZHANG Z Y, GU Y X, HAN X, et al. CPM-2: large-scale cost-effective pre-trained language models[EB]. arXiv preprint, 2021, arXiv: 2106.10715.
- [19] SOLAIMAN, I, DENNISON C. Dataset distillation for efficient training of neural networks[EB]. arXiv preprint, 2023, arXiv: 2304.01408.
- [20] GAO L, BIDERMAN S, BLACK S, et al. The pile: an 800GB dataset of diverse text for language modeling[EB]. arXiv preprint, 2021, arXiv: 2101.00027.
- [21] WANG S, YU J, ZHANG Y. A comparative study of domestic and foreign large model datasets for NLP applications[J]. Journal of Natural Language Processing Research, 2024, 3(1), 23-38.
- [22] 程大伟, 贾仁军, 李江彤, 等. 知识增强的中文金融大模型研究[J]. 大数据, 2025, 11(2): 5-18.
- CHENG D W, JIA R J, LI J T, et al. Research on knowledge-augmented Chinese financial large language model[J]. Big Data Research, 2025, 11(2): 5-18.
- [23] 李娇, 张玉清, 吴亚颀. 面向网络安全关系抽取的大语言模型数据增强方法[J]. 信息安全, 2024(10): 1477-1483.
- LI J, ZHANG Y Q, WU Y B. Data augmentation method via large language model for relation extraction in cybersecurity[J]. Netinfo Security, 2024(10): 1477-1483.
- [24] 陈宇, 刘佳, 李明. 中文大模型数据集与英文大模型数据集的对比分析及应用研究[J]. 中文信息学报, 2023, 37 (5): 1-10.
- CHEN Y, LI J, LI M. [J]. Comparative analysis and application research of Chinese large-scale model datasets and English large-scale model datasets, 2023, 37(5): 1-10.
- [25] ZHANG Y F, WANG Z Y, HE Z T, et al. BB-GeoGPT: a framework for learning a large language model for geographic information science[J]. Information Processing & Management, 2024, 61(5): 103808.
- [26] 张辉, 王强, 赵琳. 不同领域大模型数据集的比较与融合策略研究[J]. 计算机科学与探索, 2023, 17(6): 1-12.
- ZHANG H, WANG Q, ZHAO LIN. Research on the comparison and integration strategies of large model datasets in different fields[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(6): 1-12.

[27] 黄佳, 周峰, 吴迪. 大模型数据集的质量评估与比较研究[J]. 软件学报, 2024, 35(2): 1-15.
HUANG J, ZHOU F, WU D. Quality assessment and comparative study of large model datasets[J]. Journal of Software, 2024, 35(2): 1-15.

[28] 王浩, 李华, 张宇. 大模型数据集在知识图谱构建中的应用比较[J]. 计算机工程与应用, 2024, 60(1): 1-10.
WANG H, LI H, ZHANG Y. A comparative study on the application of large model datasets in knowledge graph construction[J]. Computer Engineering and Applications, 2024, 60(1): 1-10.

[29] 刘思远, 孙悦, 陈晨. 基于大模型数据集的语义理解能力对比研究[J]. 模式识别与人工智能, 2023, 36(4): 1-9.
LIU S Y, SUN Y, CHEN C. A comparative study on the semantic understanding capabilities of large model datasets[J]. Pattern Recognition and Artificial Intelligence, 2023, 36(4): 1-9.

[30] 杨明, 刘静, 陈亮. 大模型数据集对语言生成任务的影响对比[J]. 小型微型计算机系统, 2023, 44(7): 1-8.
YANG M, LIU J, CHEN L. Comparison of

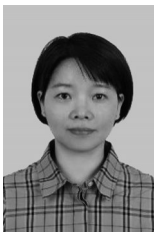
the impact of large model datasets on language generation tasks[J]. Journal of Chinese Computer Systems, 2023, 44(7): 1-8.

[31] 赵鑫, 孙浩, 李阳. 大模型数据集的跨文化适应性比较研究[J]. 情报学报, 2024, 43 (1): 1-10.
ZHAO X, SUN H, LI Y. A comparative study on the cross-cultural adaptability of large model datasets[J]. Journal of the China Society for Scientific and Technical Information, 2024, 43(1):1-10.

[32] 郑纬民. 构建支持大模型训练的计算机系统需要考虑的4个问题[J]. 大数据, 2024, 10(1): 1-8.
ZHENG W M. Four issues to consider in building a computer system supporting large model training[J]. Big data research, 2024, 10(1): 1-8.

[33] 周涛, 吴艳, 郑伟. 大模型数据集的标注策略与效果比较[J]. 数据分析与知识发现, 2024, 8 (1): 1-10.
ZHOU T, WUB Y, ZHENG W. Comparison of Annotation Strategies and Effects of Large Model Datasets[J]. Data Analysis and Knowledge Discovery, 2024, 8(1): 1-10.

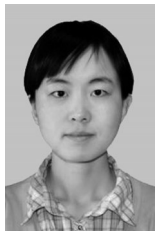
作者简介



胡晓女（1980-），女，澳门科技大学博士生、中国通信学会业务主管，主要研究方向为城市与区域经济可持续发展、人工智能、大数据、算力网络、数字经济、工业互联网、信息通信等。



李涛（1978-），男，教授级高级工程师，中国联合网络通信有限公司研究院团队负责人、中国科协科技人才奖项评审专家、中国通信学会人工智能技术与应用专业委员会委员、中国人工智能学会深度学习专业委员会委员，主要研究方向为网络智能化、AI大模型、智能算力等。



李姗姗（1986- ），女，中国联合网络通信有限公司研究院高级工程师，主要研究方向为数字化转型规划咨询、数字化评价、数字孪生网络、自智网络等。

收稿日期: 2025-02-21

通信作者: 胡晓女, huxiaonv@china-cic.cn

基金项目: 工业和信息化部2024年度指导性软课题(No.GXZK2024-49); 2024年度全国学会服务国家战略专项-面向AI的算力网关键技术路线图(No.(2024)837)

Foundation Items: Ministry of Industry and Information Technology of the People's Republic of China 2024 Guidance Soft Topic (No. GXZK2024-49), The 2024 National Academic Society Service National Strategy Special Project – Key Technology Roadmap for AI Computing Power Network(No.(2024)837)