AI 大模型需要什么样的数据

华泰研究

2023 年 5 月 11 日 | 中国内地　　专题研究

电子　增持（维持）
通信　增持（维持）

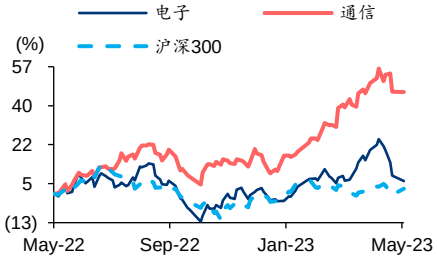
研究员　黄乐平，PhD
SAC No. S0570521050001　leping.huang@htsc.com
SFC No. AUZ066　+(852) 3658 6000

研究员　余熠
SAC No. S0570520090002　yuyi@htsc.com
SFC No. BNC535　+(86) 755 8249 2388

联系人　权鹤阳
SAC No. S0570122070045　quanheyang@htsc.com
+(86) 21 2897 2228

联系人　王珂
SAC No. S0570122080148　wangke020520@htsc.com
+(86) 21 2897 2228

行业走势图



资料来源：Wind，华泰研究

数据是大模型竞争关键要素之一，关注中国 AI 大模型数据发展

AI 的突破得益于高质量数据，我们认为数据是大模型竞争关键要素之一：1）训练大模型需要高质量、大规模、多样性的数据集；2）优质中文数据集稀缺，数字中国战略将促进数据要素市场完善，助力数据集发展。近期欧洲议会议员《人工智能法案》提案、网信办《生成式人工智能服务管理办法（征求意见稿）》对大模型训练数据的版权披露、合法性提出要求，对于数据产业链的投资机会，我们认为：1）数据资产储备公司的商业化进程值得关注；2）行业数据价值高，具有优质数据和一定大模型能力的公司或通过行业大模型赋能业务；3）关注卡位优质客户、技术降低人力成本的数据服务企业。

海外开源数据集积累丰富，合成数据或将缓解高质量数据耗尽隐忧

我们梳理了海外主要的开源语言和多模态数据集，主要的发布方包括高校、互联网巨头研究部门、非盈利研究组织以及政府机构。我们认为海外积累丰富的开源高质量数据集得益于：1）相对较好的开源互联网生态；2）免费线上书籍、期刊的长期资源积累；3）学术界、互联网巨头研究部门、非盈利研究组织及其背后的赞助基金形成了开放数据集、发表论文-被引用的开源氛围。然而，高质量语言数据或于 2026 年耗尽，AI 合成数据有望缓解数据耗尽的隐忧，Gartner 预测 2030 年大模型使用的绝大部分数据或由 AI 合成。

中文开源数据集数量少、规模小，看好数字中国战略激活数据要素产业链

与国外类似，国内大模型的训练数据包括互联网爬取数据、书籍期刊、公司自有数据以及开源数据集等。就开源数据集而言，国内外的发布方都涵盖高校、互联网巨头、非盈利机构等组织。但国内开源数据集数量少、规模小，因此国内大模型训练往往使用多个海外开源数据集。国内缺乏高质量数据集的原因在于：1）高质量数据集需要高资金投入；2）相关公司开源意识较低；3）学术领域中文数据集受重视程度低。看好数字中国战略助力国内数据集发展：1）各地数据交易所设立运营提升数据资源流通；2）数据服务商链接数据要素产业链上下游，激活数据交易流通市场，提供更多样化的数据产品。

数据产业链投资机会：关注数据生产与处理环节

数据产业链包括生产、处理等环节。我们认为数据生产可以分为通用数据和行业数据：1）海外主要数据集的通用数据来自维基、书籍期刊、高质量论坛，国内相关公司包括文本领域的百度百科、中文在线、中国科传、知乎等，以及视觉领域的视觉中国等。2）数据是垂直行业企业的护城河之一，相关公司包括城市治理和 ToB 行业应用领域的中国电信、中国移动、中国联通，CV 领域的海康、大华等。数据处理环节，模型研发企业的外包需求强烈，利好卡位优质客户、技术赋能降低人力成本的数据服务企业，如 Appen、Telus International、Scale AI。

隐私保护：监管与技术手段并举

个人数据的采集、存储和处理引发了对于 AI 时代数据隐私保护的关注。隐私保护可从监管、技术角度着手：1）监管：全球各地区出台相关法律法规，例如《中华人民共和国个人信息保护法》、欧盟《通用数据保护条例》等。2）技术：隐私保护计算在不泄露原始数据的前提下，对数据进行处理和使用。

风险提示：AI 及技术落地不及预期；本研报中涉及到未上市公司或未覆盖个股内容，均系对其客观公开信息的整理，并不代表本研究团队对该公司、该股票的推荐或覆盖。

正文目录

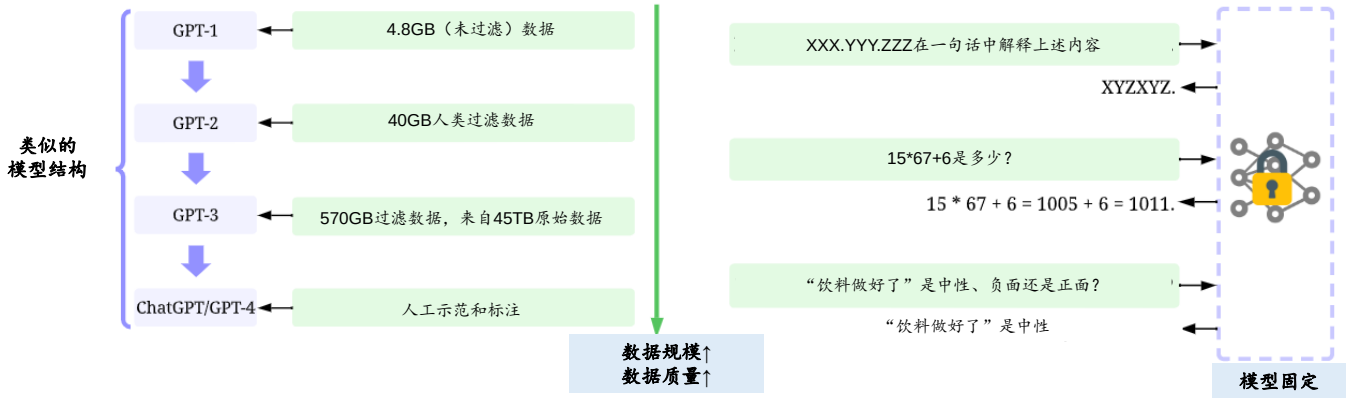
图表目录

AI 大模型需要什么样的数据集

数据将是未来 AI 大模型竞争的关键要素

人工智能发展的突破得益于高质量数据的发展。例如，大型语言模型的最新进展依赖于更高质量、更丰富的训练数据集：与 GPT-2 相比，GPT-3 对模型架构只进行了微小的修改，但花费精力收集更大的高质量数据集进行训练。ChatGPT 与 GPT-3 的模型架构类似，并使用 RLHF（来自人工反馈过程的强化学习）来生成用于微调的高质量标记数据。

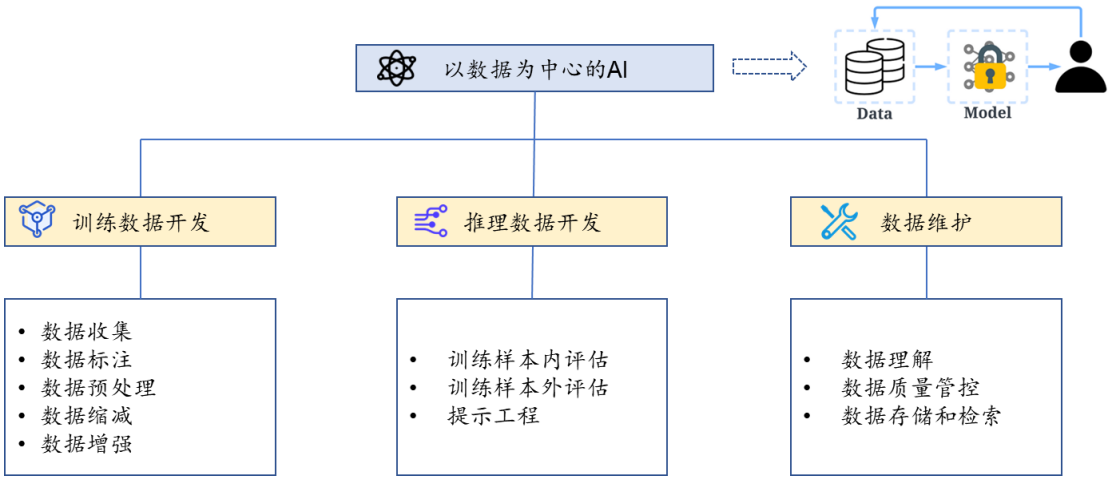
图表1：更高质量、更丰富的训练数据是 GPT 模型成功的驱动力；而除模型权重变化之外，模型架构保持相似



资料来源：Daochen Zha et al. ”Data-centric Artificial Intelligence: A Survey” 2023，华泰研究

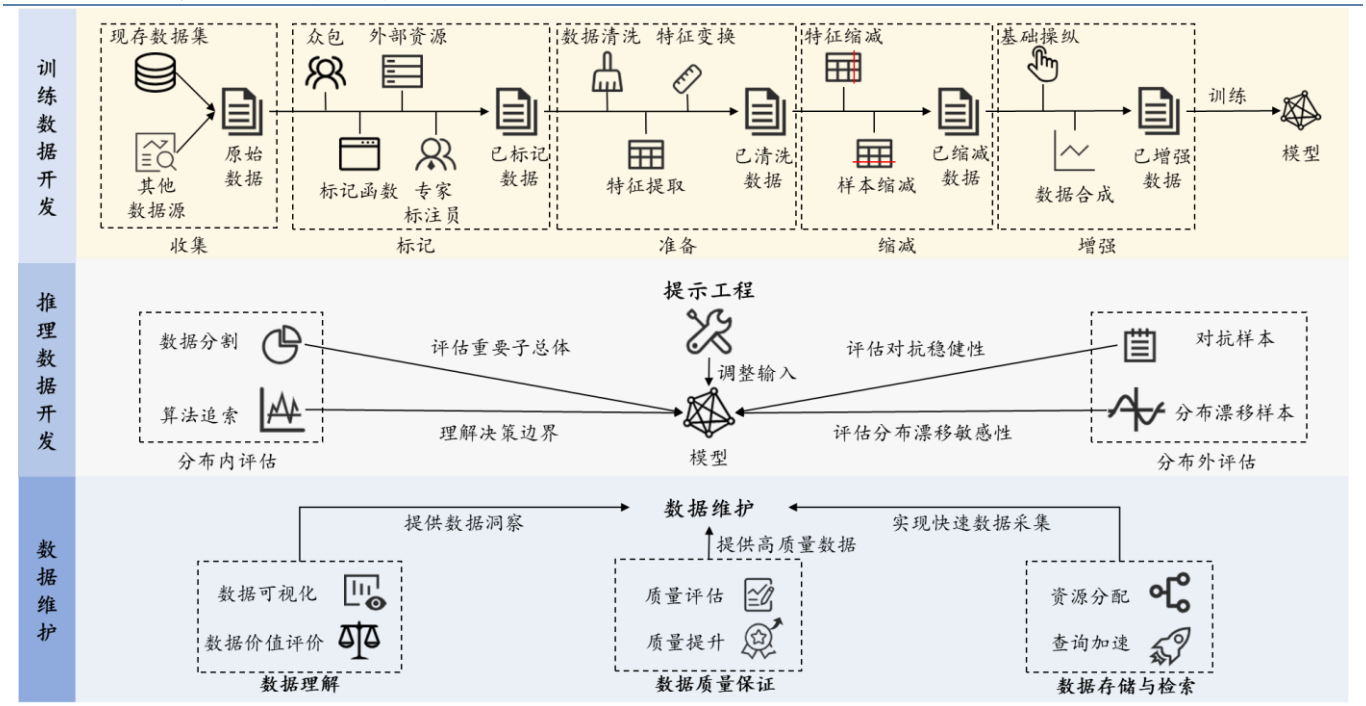
基于此，人工智能领域的权威学者吴承恩发起了“以数据为中心的 AI”运动，即在模型相对固定的前提下，通过提升数据的质量和数量来提升整个模型的训练效果。提升数据集质量的方法主要有：添加数据标记、清洗和转换数据、数据缩减、增加数据多样性、持续监测和维护数据等。**因此，我们认为未来数据成本在大模型开发中的成本占比或将提升，主要包括数据采集，清洗，标注等成本。**

图表2：以数据为中心的 AI：模型不变，通过改进数据集质量提升模型效果



资料来源：Daochen Zha et al. "Data-centric Artificial Intelligence: A Survey" 2023，华泰研究

图表3：以数据为中心的 AI：工作流拆解



资料来源：Daochen Zha et al. "Data-centric Artificial Intelligence: A Survey" 2023，华泰研究

我们认为 AI 大模型需要高质量、大规模、多样性的数据集。

1）高质量：高质量数据集能够提高模型精度与可解释性，并且减少收敛到最优解的时间，即减少训练时长。

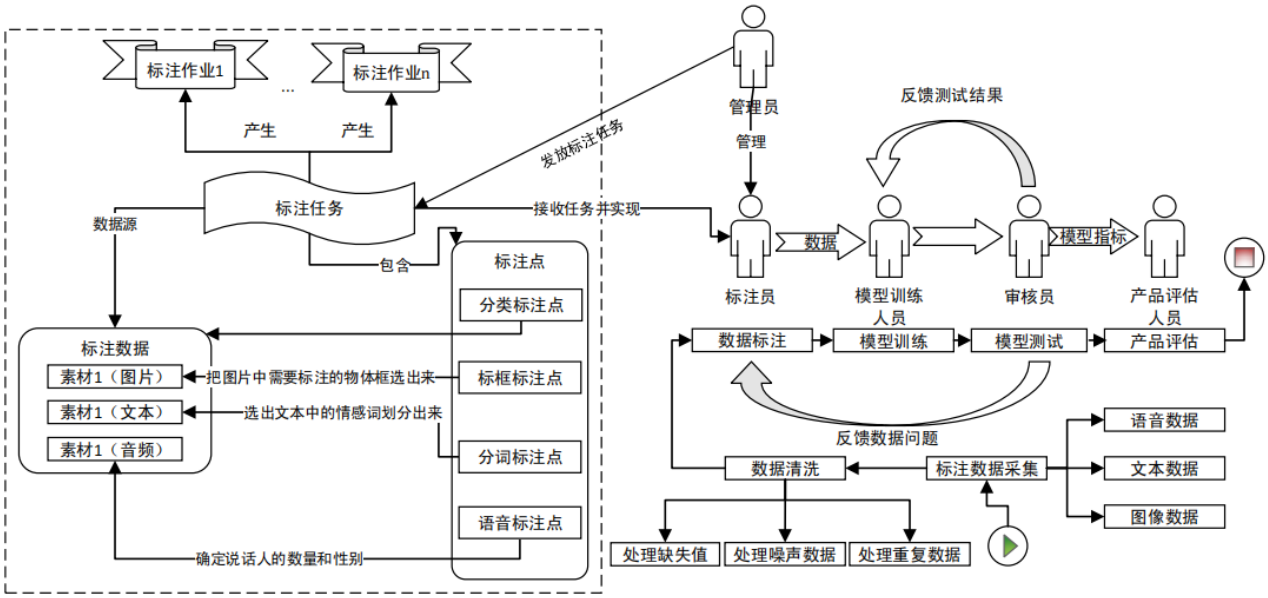
2）大规模：OpenAI 在《Scaling Laws for Neural Language Models》中提出 LLM 模型所遵循的“伸缩法则”（scaling law），即独立增加训练数据量、模型参数规模或者延长模型训练时间，预训练模型的效果会越来越好。

3）丰富性：数据丰富性能够提高模型泛化能力，过于单一的数据会非常容易让模型过于拟合训练数据。

数据集如何产生

建立数据集的流程主要分为 1）数据采集；2）数据清洗：由于采集到的数据可能存在缺失值、噪声数据、重复数据等质量问题；3）数据标注：最重要的一个环节；4）模型训练：模型训练人员会利用标注好的数据训练出需要的算法模型；5）模型测试：审核员进行模型测试并将测试结果反馈给模型训练人员，而模型训练人员通过不断地调整参数，以便获得性能更好的算法模型；6）产品评估：产品评估人员使用并进行上线前的最后评估。

图表4：数据标注基本流程

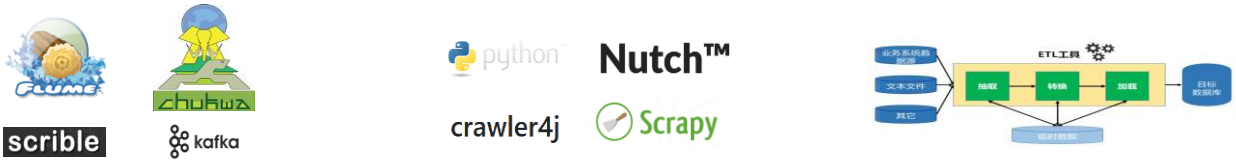


资料来源：蔡莉等《数据标注研究综述》2020，华泰研究

流程#1：数据采集。采集的对象包括视频、图片、音频和文本等多种类型和多种格式的数据。数据采集目前常用的有三种方式，分别为：1）系统日志采集方法；2）网络数据采集方法；3）ETL。

图表5：数据采集三种常见方式

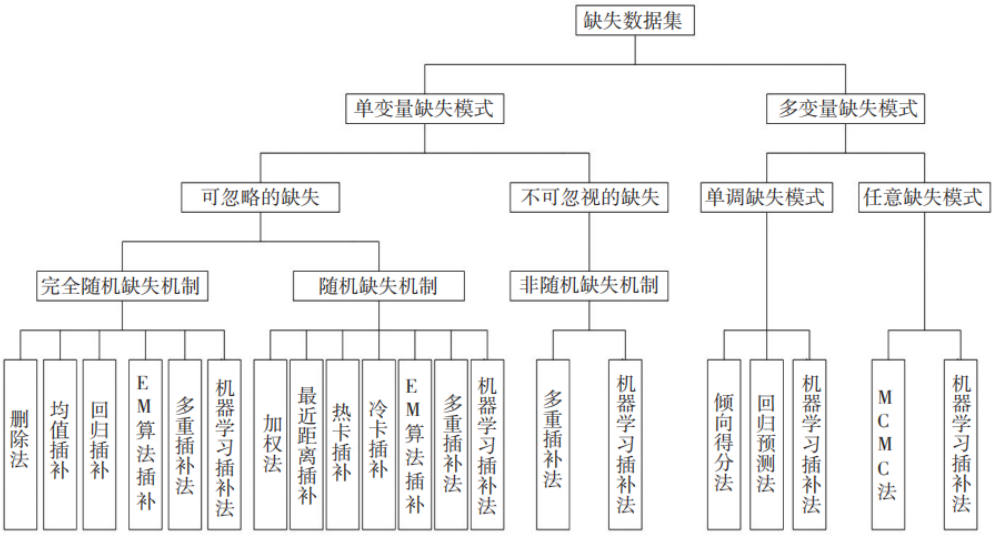
数据采集三种常见方式		
系统日志采集方法	网络数据采集	ETL
• 构建应用系统和分析系统的桥梁，并将它们之间的关联解耦； • 支持近实时的在线分析系统和分布式并发的离线分析系统； • 具有高可扩展性，也就是说，当数据量增加时，可以通过增加节点进行水平扩展； • 目前为止，运用较为广泛的有Flume、Chukwa、Scrible和Kafka。	• 通过网络爬虫或网站公开API方式获取大数据信息； • 网络爬虫工具包括 python爬虫、分布式网络爬虫工具、Java网络爬虫工具、非Java网络爬虫工具。分布式网络爬虫工具，如Nutch。	• 即Extract-Transform-Load，描述将数据从来源端经过抽取(extract)、转换(transform)、加载(load)至目的端的过程； • 它是一个数据集成过程，将来自多个数据源的数据组合到一个单一的、一致的数据存储中，该数据存储被加载到数据仓库或其他目标系统中。



资料来源：CSDN，Apache，Scrible，Python，GitHub，Scrapy，IBM，搜狗百科，华泰研究

流程#2：数据清洗是提高数据质量的有效方法。由于采集到的数据可能存在缺失值、噪声数据、重复数据等质量问题，故需要执行数据清洗任务，数据清洗作为数据预处理中至关重要的环节，清洗后数据的质量很大程度上决定了 AI 算法的有效性。

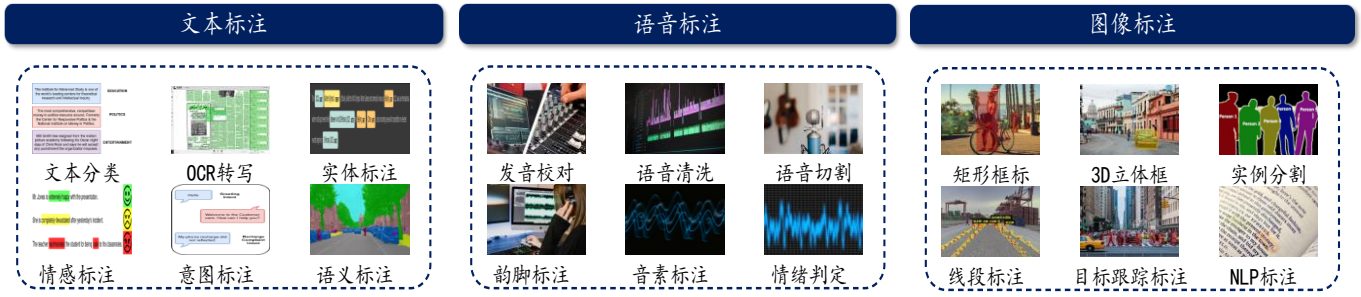
图表6：缺失数据的处理方法



资料来源：邓建新等《缺失数据的处理方法及其发展趋势》2019，华泰研究

流程#3：数据标注是流程中最重要的一个环节。管理员会根据不同的标注需求，将待标注的数据划分为不同的标注任务。每一个标注任务都有不同的规范和标注点要求，一个标注任务将会分配给多个标注员完成。

图表7：三大类数据标注



资料来源：Devol Shah “A Step-by-Step Guide to Text Annotation” 2022，CSDN，景联文科技，华泰研究

流程#4：最终通过产品评估环节的数据才算是真正过关。产品评估人员需要反复验证模型的标注效果，并对模型是否满足上线目标进行评估。

图表8：各数据标注质量评估算法对比

分类	算法名称	优点	缺点
图像标注质量评估算法	MV 算法	简单易用，常用作其他众包质量控制算法的基准算法	没有考虑到每个标注任务、 标注者的不同可靠性
	EM 算法	在一定意义下可以收敛到局部最大化	数据缺失比例较大时，收敛速度比较缓慢
	RY 算法	将分类器与 Ground-truth 结合起来进行学习	需要对标注专家的特异性和敏感性强加先验
文本标注质量评估算法	BLEU 算法	方便、快速、结果有参考价值	测评精度易受常用词干扰
	ROUGE 算法	参考标注越多，待评估数据的相关性就越高	无法评价标注数据的流畅度
	METEOR 算法	评估时考虑了同义词匹配， 提高了评估的准确率	长度惩罚，当被评估的数据量小时，测量精度较高
	CIDEr 算法	从文本标注质量评估的相关性上升到质量评估的相似性进	对所有匹配上的词都同等对待会导致部分词的重要性被削弱
	SPICE 算法	从图的语义层面对图像标注进行评估	图的语义解析方面还有待进一步完善
	ZenCrowd 算法	将算法匹配和人工匹配结合，在一定程度上实现了标注质量和效率的共同提高	无法自动为定实体选择最佳数据集
语音标注质量评估算法	WER 算法	可以分数字、英文、中文等情况分别来看	当数据量大时，性能会特别差
	SER 算法	对句子的整体性评估要优于 WER 算法	句错误率较高，一般是词错误率的 2 倍~3 倍

资料来源：蔡莉等《数据标注研究综述》2020，华泰研究

他山之石#1：海外主要大语言模型数据集

参数量和数据量是判断大模型的重要参数。2018 年以来，大语言模型训练使用的数据集规模持续增长。2018 年的 GPT-1 数据集约 4.6GB，2020 年的 GPT-3 数据集达到了 753GB，而到了 2021 年的 Gopher，数据集规模已经达到了 10,550GB。总结来说，从 GPT-1 到 LLaMA 的大语言模型数据集主要包含六类：维基百科、书籍、期刊、Reddit 链接、Common Crawl 和其他数据集。

图表9：大语言模型数据集综合分析

大模型	维基百科	书籍	期刊	Reddit链接	Common Crawl	其他	合计
GPT-1		4.6					4.6
GPT-2				**40**			40
GPT-3	11.4	21	101	50	**570**		753
The Pile v1	**6**	**118**	**244**	**63**	**227**	**167**	825
Megatron-11B	**11.4**	4.6		**38**	**107**		161
MT-NLG	6.4	118	77	63	983	127	1374
Gopher	**12.5**	**2100**	164.4		3450	4823	10550
LLaMA	**83**	**85**	**92**		**4162.2**	**406**	**4828.2**

注：以 GB 为单位，公开的数据以粗体表示，仅原始训练数据集大小
资料来源：Alan D. Thompson “What’s in My AI” 2023，Hugo Touvron et al. “LLaMA: Open and Efficient Foundation Language Models” 2023，华泰研究

数据集#1：维基百科

维基百科是一个免费的多语言协作在线百科全书。维基百科致力于打造包含全世界所有语言的自由的百科全书，由超三十万名志愿者组成的社区编写和维护。截至 2023 年 3 月，维基百科拥有 332 种语言版本，总计 60,814,920 条目。其中，英文版维基百科中有超过 664 万篇文章，拥有超 4,533 万个用户。维基百科中的文本很有价值，因为它被严格引用，以说明性文字形式写成，并且跨越多种语言和领域。一般来说，重点研究实验室会首先选取它的纯英文过滤版作为数据集。

图表10：英文维基百科数据集分类

排名	类别	占比	大小（GB）	Tokens（百万）
1	生物	27.80%	3.1	834
2	地理	17.70%	1.9	531
3	文化和艺术	15.80%	1.7	474
4	历史	9.90%	1.1	297
5	生物、健康和医学	7.80%	0.9	234
6	体育	6.50%	0.7	195
7	商业	4.80%	0.5	144
8	其他社会	4.40%	0.5	132
9	科学 & 数学	3.50%	0.4	105
10	教育	1.80%	0.2	54
总计		100%	11.4	3000

资料来源：Alan D. Thompson “What’s in My AI” 2023，华泰研究

数据集#2：书籍

书籍主要用于训练模型的故事讲述能力和反应能力，包括小说和非小说两大类。数据集包括 Project Gutenberg 和 Smashwords (Toronto BookCorpus/BookCorpus)等。Project Gutenberg 是一个拥有 7 万多本免费电子书的图书馆，包括世界上最伟大的文学作品，尤其是美国版权已经过期的老作品。BookCorpus 以作家未出版的免费书籍为基础，这些书籍来自于世界上最大的独立电子书分销商之一的 Smashwords。

图表11：BookCorpus 分类

序号	类别	书籍数量	占比（书籍数量 / 11038）
1	浪漫	2880	26.10%
2	幻想	1502	13.60%
3	科技小说	823	7.50%
4	新成人	766	6.90%
5	年轻成人	748	6.80%
6	惊悚	646	5.90%
7	神秘	621	5.60%
8	吸血鬼	600	5.40%
9	恐怖	448	4.10%
10	青少年	430	3.90%
11	冒险	390	3.50%
12	其他	360	3.30%
13	文学	330	3.00%
14	幽默	265	2.40%
15	历史	178	1.60%
16	主题	51	0.50%
总计		11038	100.0%

资料来源：Alan D. Thompson “What’s in My AI” 2023，华泰研究

数据集#3：期刊

期刊可以从 ArXiv 和美国国家卫生研究院等官网获取。预印本和已发表期刊中的论文为数据集提供了坚实而严谨的基础，因为学术写作通常来说更有条理、理性和细致。ArXiv 是一个免费的分发服务和开放获取的档案，包含物理、数学、计算机科学、定量生物学、定量金融学、统计学、电气工程和系统科学以及经济学等领域的 2,235,447 篇学术文章。美国国家卫生研究院是美国政府负责生物医学和公共卫生研究的主要机构，支持各种生物医学和行为研究领域的研究，从其官网的“研究&培训”板块能够获取最新的医学研究论文。