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

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

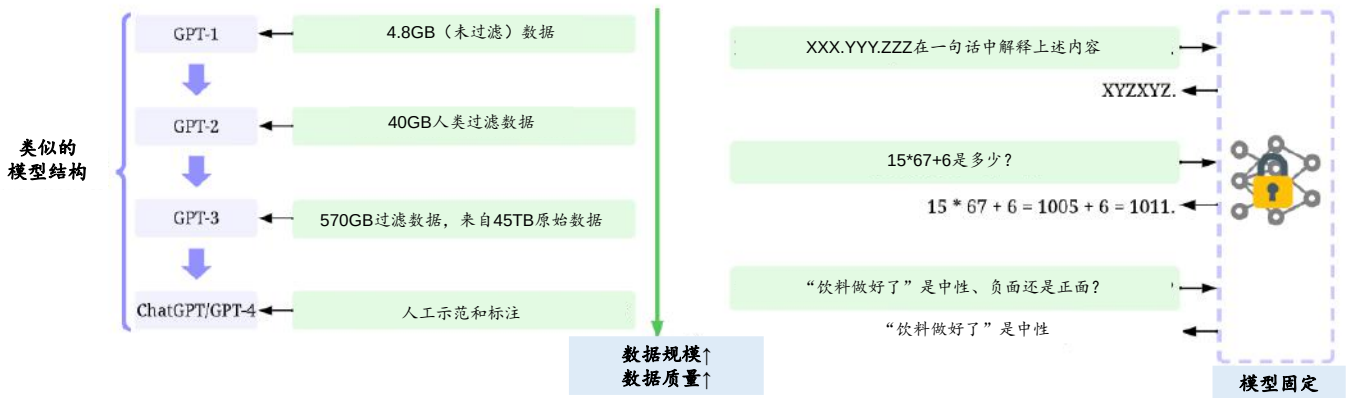
图表目录

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。

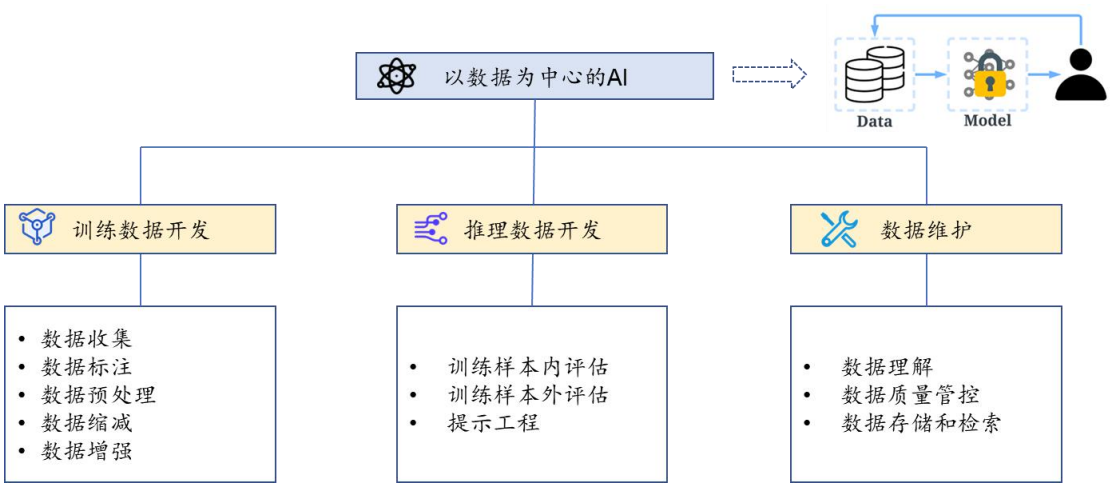
图表1：更高质量、更丰富的训练数据是 GPT 模型成功的驱动力；而除模型权重变化之外，模型架构保持相似



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

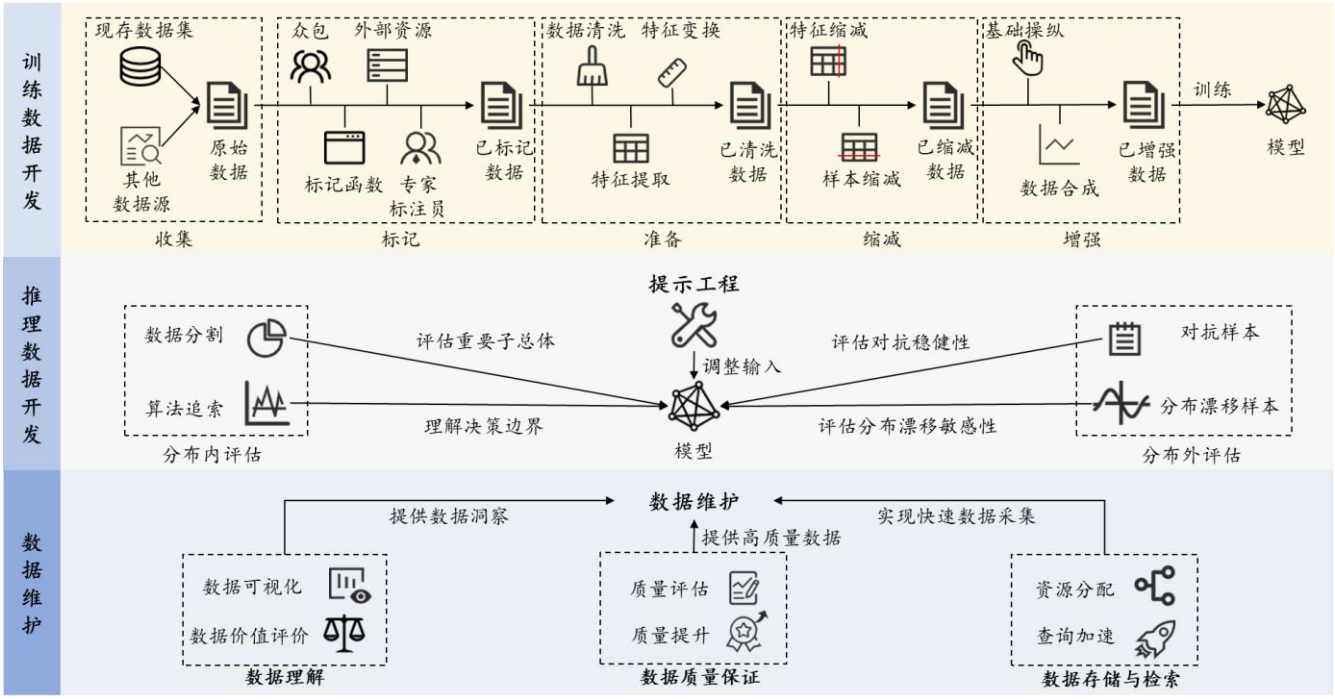
基于此，人工智能领域的权威学者吴承恩发起了“以数据为中心的 AI”运动，即在模型相对固定的前提下，通过提升数据的质量和数量来提升整个模型的训练效果。提升数据集质量的方法主要有：添加数据标记、清洗和转换数据、数据缩减、增加数据多样性、持续监测和维护数据等。**因此，我们认为未来数据成本在大模型开发中的成本占比或将提升，主要包括数据采集，清洗，标注等成本。**

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。

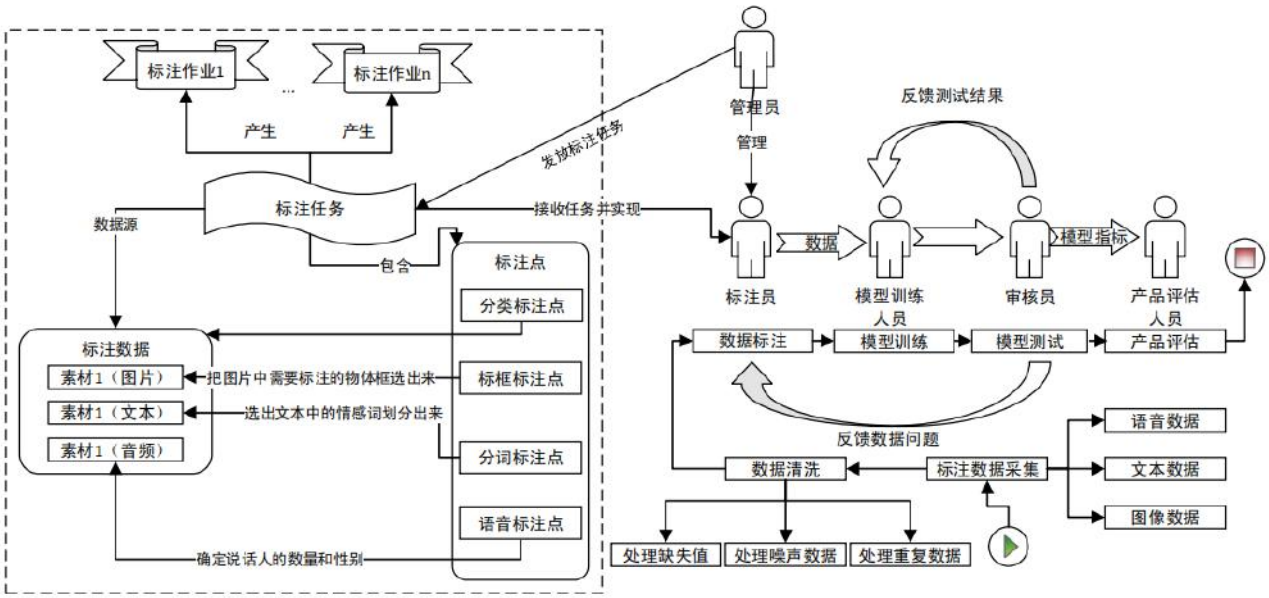
2）大规模：OpenAI 在《Scaling Laws for Neural Language Models》中提出 LLM 模型所遵循的“伸缩法则”（scaling law），即独立增加训练数据量、模型参数规模或者延长模型训练时间，预训练模型的效果会越来越好。

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。

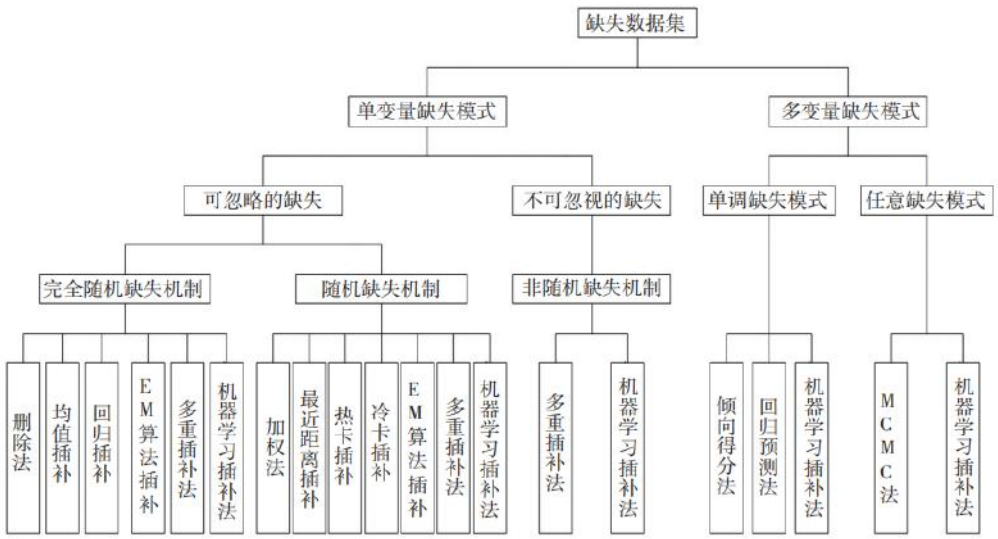
图表5：数据采集三种常见方式

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。

图表12： ArVix 官网

资料来源：AriVix，华泰研究

图表13： 美国国家卫生研究院官网

资料来源：美国国家卫生研究院官网，华泰研究

数据集#4：WebText（来自 Reddit 链接）

Reddit 链接代表流行内容的风向标。Reddit 是一个娱乐、社交及新闻网站，注册用户可以将文字或链接在网站上发布，使它成为了一个电子布告栏系统。WebText 是一个大型数据集，它的数据是从社交媒体平台 Reddit 所有出站链接网络中爬取的，每个链接至少有三个赞，代表了流行内容的风向标，对输出优质链接和后续文本数据具有指导作用。

Reddit 宣布收取数据使用费。2023 年 4 月，Reddit 宣布将向使用其 API 训练 AI 聊天机器人的公司收取数据使用费，其中便包含微软、谷歌、OpenAI 等，目前具体收费标准暂未公布，但可能会根据不同使用者划分不同等级收费标准。许多公司已经意识到数据的价值，如图片托管服务商 Shutterstock 已把图像数据出售给 OpenAI，推特计划针对 API 使用收取几万到几十万美元不等的费用。

图表14：WebText 前 50 个域

排名	域	链接（百万个）	占比	Tokens（百万）
1	Google	1.54	3.4%	514
2	Archive	0.60	1.3%	199
3	Blogspot	0.46	1.0%	152
4	GitHub	0.41	0.9%	138
5	The NY Times	0.33	0.7%	111
6	WordPress	0.32	0.7%	107
7	WashingtonPost	0.32	0.7%	105
8	Wikia	0.31	0.7%	104
9	BBC	0.31	0.7%	104
10	TheGuardian	0.25	0.5%	82
11	eBay	0.21	0.5%	70
12	Pastebin	0.21	0.5%	70
13	CNN	0.20	0.4%	66
14	Yahoo	0.20	0.4%	65
15	HuffingtonPost	0.19	0.4%	62
16	Go	0.19	0.4%	62
17	Reuters	0.18	0.4%	61
18	IMDb	0.18	0.4%	61
19	Goo	0.16	0.4%	54
20	NIH	0.14	0.3%	47
21	CBC	0.14	0.3%	45
22	Apple	0.13	0.3%	43
23	Medium	0.13	0.3%	42
24	DailyMail	0.12	0.3%	40
25	SteamPowered	0.11	0.2%	36
26	Independent	0.11	0.2%	35
27	Etsy	0.11	0.2%	35
28	Craigslist	0.10	0.2%	33
29	BusinessInsider	0.09	0.2%	31
30	Telegraph	0.09	0.2%	31
31	Wizards	0.09	0.2%	30
32	USAtoday	0.08	0.2%	28
33	TheHill	0.08	0.2%	27
34	NHL	0.08	0.2%	27
35	FoxNews	0.08	0.2%	26
36	淘宝	0.08	0.2%	26
37	Bloomberg	0.08	0.2%	26
38	NPR	0.08	0.2%	26
39	MLB	0.08	0.2%	26
40	LA Times	0.08	0.2%	26
41	Megalodon	0.08	0.2%	25
42	ESPN	0.07	0.2%	24
43	KickStarter	0.07	0.2%	24
44	BreitBart	0.07	0.2%	24
45	ABC	0.07	0.2%	23
46	NewEgg	0.07	0.2%	23
47	WWE	0.07	0.1%	22
48	MyAnimeList	0.07	0.1%	22
49	Microsoft	0.07	0.1%	22
50	Buzzfeed	0.06	0.1%	22
总计		9.3	20.7%	

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

数据集#5：Common crawl/C4

Common crawl 是 2008 年至今的一个网站抓取的大型数据集。Common Crawl 是一家非盈利组织，致力于为互联网研究人员、公司和个人免费提供互联网副本，用于研究和分析，它的数据包含原始网页、元数据和文本提取，文本包含 40 多种语言和不同领域。重点研究实验室一般会首先选取它的纯英文过滤版（C4）作为数据集。

图表15：C4 前 23 个域名（不包括维基百科）

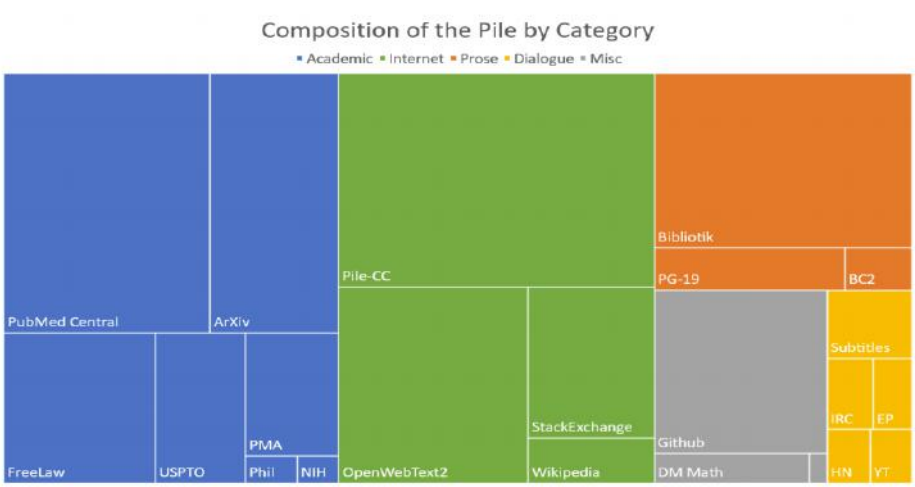
排名	域	Token（百万）	占比
1	Google Patents	750	0.48%
2	The NY Times	100	0.06%
3	Los AngelesTimes	90	0.06%
4	The Guardian	90	0.06%
5	PLoS	90	0.06%
6	Forbes	80	0.05%
7	HuffingtonPost	75	0.05%
8	Patents.com	71	0.05%
9	Scribd	70	0.04%
10	Washington Post	65	0.04%
11	The Motley Fool	61	0.04%
12	IPFS	60	0.04%
13	Frontiers Media	60	0.04%
14	Business Insider	60	0.04%
15	Chicago Tribune	59	0.04%
16	Booking.com	58	0.04%
17	The Atlantic	57	0.04%
18	Springer Link	56	0.04%
19	Al Jazeera	55	0.04%
20	Kickstarter	54	0.03%
21	FindLaw Caselaw	53	0.03%
22	NCBI	53	0.03%
23	NPR	52	0.03%
总计		2219	1.42%

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

其他数据集

The Pile 数据集：一个 825.18 GB 的英语文本数据集，用于训练大规模语言模型。The Pile 由上文提到的 ArXiv、WebText、Wikipedia 等在内的 22 个不同的高质量数据集组成，包括已经建立的自然语言处理数据集和几个新引入的数据集。除了训练大型语言模型外，The Pile 还可以作为语言模型跨领域知识和泛化能力的广泛覆盖基准。

图表16：按有效尺寸划分的 The Pile 组成树状图



资料来源：Leo Gao et al. “The Pile: An 800GB Dataset of Diverse Text for Language Modeling” 2020，华泰研究

其他数据集包含了 GitHub 等代码数据集、StackExchange 等对话论坛和视频字幕数据集等。

图表17：其他常见 NLP 数据集

数据集分类	数据集	简介
代码数据集	Github	一个大型的开源代码库，在多年以前的预训练语言模型例如 BERT、GPT 里几乎没有人用，该代码数据的加入对语言模型的逻辑推理能力有极大的帮助
	CodeSearchNet	一个大型函数数据集，其中包含来自 GitHub 上的开源项目的用 Go、Java、JavaScript、PHP、Python 和 Ruby 编写的相关文档
	StaQC	是迄今为止最大的数据集，大约有 148K Python 和 120K SQL 域问题代码对，它们是使用 Bi-View Hierarchical Neural Network 从 Stack Overflow 中自动挖掘出来的
	CodeExp	其中包含 (1) 2.3 的大分区 百万原始代码-docstring 对，(2) 一个介质 158,000 对的分区从 使用学习的过滤器的原始语料库，以及 (3) 具有严格的人类 13,000 对的分区注释
	ETH Py150 Open	来自 GitHub 的 740 万个 Python 文件的大规模去重语料库
论坛数据集	StackExchange	StackOverflow 的超集，包含有不限于计算机的各种各样不同领域的高质量问答
	Federated Stack Overflow	数据由所有问题和答案的正文组成。Body 被解析成句子，任何少于 100 个句子的用户都会从数据中删除。最少的预处理如下进行：小写文本，对 HTML 符号进行转义，删除非 ASCII 符号，单独的标点符号作为单独的标记（撇号和连字符除外），去除多余的空白，用特殊标记替换 URLS
	QUASAR	一个由 QUASAR-S 和 QUASAR-T 组成的大规模数据集。这些数据集中的每一个都旨在专注于评估旨在理解自然语言查询、大量文本语料库并从语料库中提取问题答案的系统。具体来说，QUASAR-S 包含 37,012 个填空题，这些问题是使用实体标签从流行的网站 Stack Overflow 收集的
	GIF Reply Dataset	发布的 GIF 回复数据集包含 1,562,701 次 Twitter 上的真实文本 - GIF 对话。在这些对话中，使用了 115,586 个独特的 GIF。元数据包括 OCR 提取的文本、带注释的标签和对象名称，也可用于该数据集中的一些 GIF
视频字幕数据集	TVC (TV show Captions)	电视节目 Caption 是一个大规模的多模态字幕数据集，包含 261,490 个字幕描述和 108,965 个短视频片段。TVC 是独一无二的，因为它的字幕也可以描述对话/字幕，而其他数据集中的字幕仅描述视觉内容

资料来源：Hugo Touvron et al. “LLaMA: Open and Efficient Foundation Language Models” 2023，OpenDataLab，华泰研究

他山之石#2：海外主要多模态数据集

模态是事物的一种表现形式，多模态通常包含两个或者两个以上的模态形式，包括文本、图像、视频、音频等。多模态大模型需要更深层次的网络和更大的数据集进行预训练。过去数年中，多模态大模性参数量及数据量持续提升。例如，2022 年 Stability AI 发布的 Stable Diffusion 数据集包含 58.4 亿图文对/图像，是 2021 年 OpenAI 发布的 DALL-E 数据集的 23 倍。

图表18：多模态大模型数据集介绍

公司	多模态大模型	发布时间	最大参数量（B）	数据集（M_图文对/图像）	数据集类别
OpenAI	DALL-E	2021.1	12	250	Conceptual Captions、YFCC100M、Wikipedia
Meta	Make-a-scene	2022.3	4	35	-
谷歌、Hugging Face	DALL-E mini	2022.4	0.4	15	-
OpenAI	DALL-E 2	2022.4	6.5	650	AVA
谷歌	Imagen	2022.5	7.6	860	内部数据、LAION-400M
谷歌	Parti	2022.6	20	4800	MS-COCO、LAION-400M、FIT400M、JFT-4B
Stability AI	Stable Diffusion	2022.8	na	5840	LAION-5B
谷歌	PaLM-E	2023.3	562	na	Language-Table

资料来源：Aditya Ramesh et al. “Zero-Shot Text-to-Image Generation” 2021，Oran Gafni et al. “Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors” 2022，Aditya Ramesh “Hierarchical Text-Conditional Image Generation with CLIP Latents” 2022，Chitwan Saharia et al. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding” 2022，Jiahui Yu et al. “Scaling Autoregressive Models for Content-Rich Text-to-Image Generation” 2022，Jay Alammar ”The Illustrated Stable Diffusion” 2022，Danny Driess et al. “PaLM-E: An Embodied Multimodal Language Model” 2023，华泰研究

类别#1：语音+文本

SEMAINE 数据集：创建了一个大型视听数据库，作为构建敏感人工侦听器(SAL)代理的迭代方法的一部分，该代理可以使人参与持续的、情绪化的对话。高质量的录音由五台高分辨率、高帧率摄像机和四个同步录制的麦克风提供。录音共有 150 个参与者，总共有 959 个与单个 SAL 角色的对话，每个对话大约持续 5 分钟。固体 SAL 录音被转录和广泛注释：每个剪辑 6-8 个评分者追踪 5 个情感维度和 27 个相关类别。

图表19： SEMAINE——四个 SAL 角色化身

资料来源：Gary McKeown et al. “The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent” 2011，华泰研究

类别#2：图像+文本

COCO 数据集：MS COCO 的全称是 Microsoft Common Objects in Context，起源于微软于 2014 年出资标注的 Microsoft COCO 数据集，与 ImageNet 竞赛一样，被视为是计算机视觉领域最受关注和最权威的比赛之一。 COCO 数据集是一个大型的、丰富的物体检测，分割和字幕数据集。图像包括 91 类目标，328,000 张图像和 2,500,000 个 label。

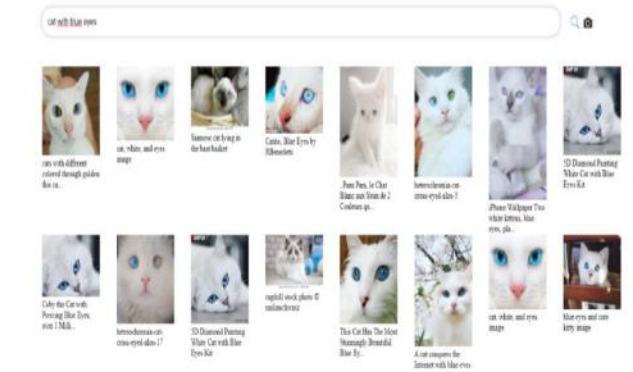
Conceptual Captions 数据集：图像标题注释数据集，其中包含的图像比 MS-COCO 数据集多一个数量级，并代表了更广泛的图像和图像标题风格。通过从数十亿个网页中提取和过滤图像标题注释来实现这一点。

ImageNet 数据集：建立在 WordNet 结构主干之上的大规模图像本体。ImageNet 的目标是用平均 5,001,000 张干净的全分辨率图像填充 WordNet 的 80,000 个同义词集中的大多数。这将产生数千万个由 WordNet 语义层次结构组织的注释图像。ImageNet 的当前状态有 12 个子树，5247 个同义词集，总共 320 万张图像。

LAION-400M 数据集：LAION-400M 通过 CommonCrawl 提取出随机抓取 2014-2021 年的网页中的图片、文本内容。通过 OpenAI 的 Clip 计算，去除了原始数据集中文本和图片嵌入之间预先相似度低于 0.3 的内容和文本，提供了 4 亿个初筛后的图像文本对样本。

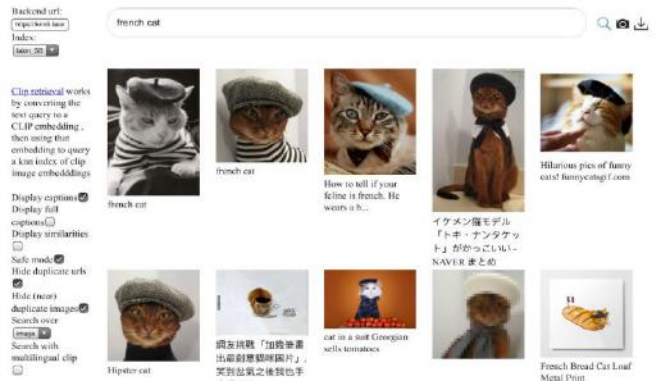
LAION-5B 数据集：其包含 58.5 亿个 CLIP 过滤的图像-文本对的数据集，比 LAION-400M 大 14 倍，是世界第一大规模、多模态的文本图像数据集，共 80T 数据，并提供了色情图片过滤、水印图片过滤、高分辨率图片、美学图片等子集和模型，供不同方向研究。

图表20： LAION-400M 搜索“蓝眼睛的猫”得出的结果示例



资料来源：Christoph Schuhmann et al “LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs” 2021，华泰研究

图表21： LAION-5B 搜索“法国猫”得出的结果示例



资料来源：LAION-5B 官网，华泰研究

Language Table 数据集：Language-Table 是一套人类收集的数据集，是开放词汇视觉运动学习的多任务连续控制基准。

IAPR TC-12 数据集：IAPR TC-12 基准的图像集合包括从世界各地拍摄的 2 万张静态自然图像，包括各种静态自然图像的横截面。这包括不同运动和动作的照片，人物、动物、城市、风景和当代生活的许多其他方面的照片。示例图像可以在第 2 节中找到。每张图片都配有最多三种不同语言（英语、德语和西班牙语）的文本标题。

AVA 数据集：AVA 是美学质量评估的数据库，包括 25 万张照片。每一张照片都有一系列的评分、语义级别的 60 类标签和 14 类照片风格。

OpenViDial 数据集：当人们交谈时，说话者接下来要说什么在很大程度上取决于他看到了什么。OpenViDal 一个用于此目的的大型多模块对话数据集。这些对话回合和视觉环境都是从电影和电视剧中提取出来的，其中每个对话回合都与发生的相应视觉环境相匹配。版本 1 包含 110 万个对话回合以及存储在图像中的 110 万个视觉上下文。版本 2 要大得多，包含 560 万个对话回合以及存储在图像中的 560 万个视觉上下文。

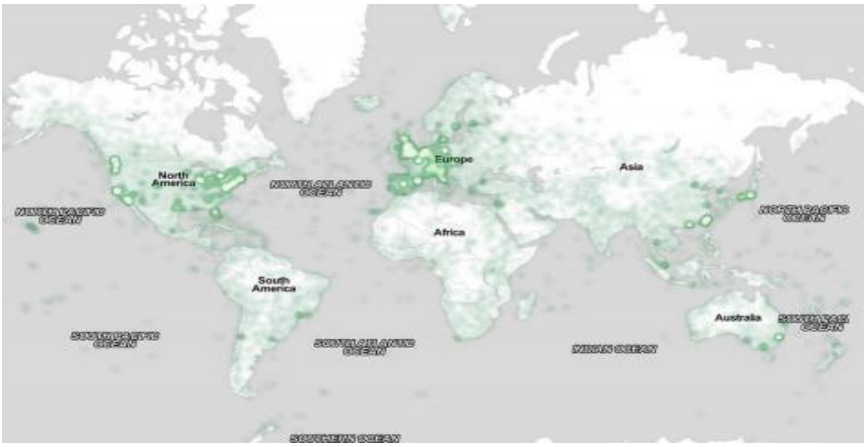
图表22： OpenViDial——两个简短对话中的视觉环境

资料来源：GitHub，华泰研究

类别#3：视频+图像+文本

YFCC100 数据集：YFCC100M 是一个包含 1 亿媒体对象的数据集，其中大约 9920 万是照片，80 万是视频，所有这些都带有创作共用许可。数据集中的每个媒体对象都由几块元数据表示，例如 Flickr 标识符、所有者名称、相机、标题、标签、地理位置、媒体源。从 2004 年 Flickr 成立到 2014 年初，这些照片和视频是如何被拍摄、描述和分享的，这个集合提供了一个全面的快照。

图表23：YFCC100M 数据集中 100 万张照片样本的全球覆盖

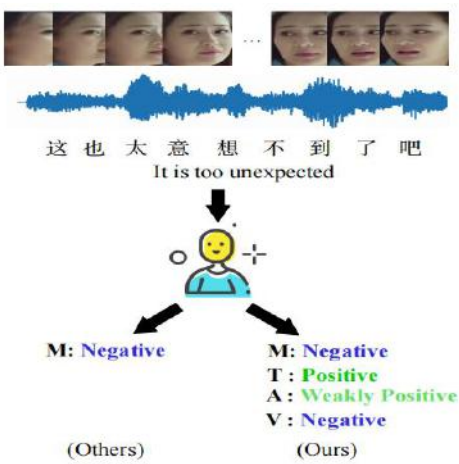


资料来源：Bart Thomee et al. “YFCC100M: The New Data in Multimedia Research” 2016，华泰研究

类别#4：图像+语音+文本

CH-SIMS 数据集：CH-SIMS 是中文单模态和多模态情感分析数据集，包含 2,281 个精细化的野外视频片段，既有多模态注释，也有独立单模态注释。它允许研究人员研究模态之间的相互作用，或使用独立的单模态注释进行单模态情感分析。

图表24：CH-SIMS 与其他数据集之间注释差异的示例



资料来源：Wenmeng Yu et al. “CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotations of Modality” 2020，华泰研究

类别#5：视频+语音+文本

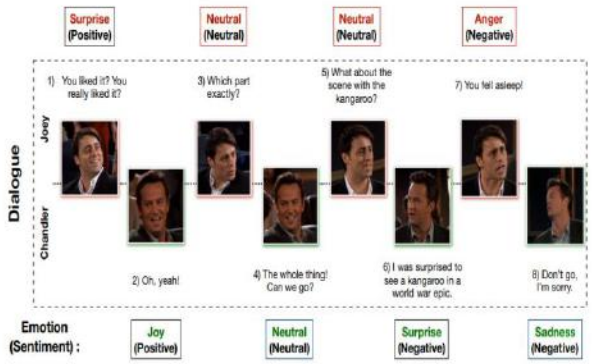
IEMOCAP 数据集：南加州大学语音分析与解释实验室(SAIL)收集的一种新语料库，名为“交互式情感二元动作捕捉数据库”(IEMOCAP)。该数据库记录了 10 位演员在面部、头部和手上的二元会话，这些标记提供了他们在脚本和自发口语交流场景中面部表情和手部动作的详细信息。语料库包含大约 12 小时的数据。详细的动作捕捉信息、激发真实情绪的交互设置以及数据库的大小使这个语料库成为社区中现有数据库的有价值的补充，用于研究和建模多模态和富有表现力的人类交流。

MELD 数据集：MELD 收录了《老友记》电视剧 1,433 个对话中的 13,708 个话语。MELD 优于其他对话式情绪识别数据集 SEMAINE 和 IEMOCAP，因为它由多方对话组成，并且 MELD 中的话语数量几乎是这两个数据集的两倍。MELD 中的话语是多模态的，包括音频和视觉形式以及文本。

图表25：IEMOCAP——有 8 个摄像头的 VICON 运动捕捉系统

资料来源：Carlos Busso et al. “IEMOCAP: interactive emotional dyadic motion capture database. Lang Resources & Evaluation” 2008，华泰研究

图表26：MELD 数据集——对话中和对话前说话人情绪变化对比



资料来源：Soujanya Poria et al. “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations” 2018，华泰研究

他山之石#3：海外主要大模型数据集由何方发布

海外主要开源大模型数据集发布方主要分为：

1) 非营利组织/开源组织：古腾堡文学档案基金会发布的 Project Gutenberg 截至 2018 年已收录 57,000 部书籍，平均每周新增 50 部。Common Crawl 抓取网络并免费向公众提供其档案和数据集，一般每个月完成一次抓取。艾伦人工智能研究所分别于 2017 年、2018 年和 2019 年发布了基于维基百科的 TriviaQA、QuAC、Quoref。Eleuther AI 发布了 825GB 多样化文本数据集 The Pile。LAION 2021 年发布包含 4 亿图文对的 LAION-400M 数据集，2022 年发布包含 58.5 亿图文对的 LAION-5B 数据集；
2) 学术界：例如多伦多大学和麻省理工学院联合发布了 BookCorpus；
3) 互联网巨头研究部门：例如 Google Research 发布了 C4 文本数据集、AVA 和 Conceptual Captions 等等图像数据集等；
4) 政府机构：政府机构是一些常见的数据集发布方，通常包含关于经济和医学等方面的数据，美国国家卫生研究院发布的 MedQuAD 包括从 12 个 NIH 网站创建的 47,457 个医学问答对；
5) 多种类型机构合作：尤其是学术界与互联网巨头研究部门、开源组织之间的合作。例如 Facebook、伦敦大学学院和 DeepMind 联合发布了 ArxivPaper 数据集。卡内基梅隆大学、雅虎研究院和 International Computer Science Institute 联合发布了 YFCC100M。

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基金形成了开放数据集、发表论文-被引用的开源氛围。

图表27：常见大模型数据集发布方总结

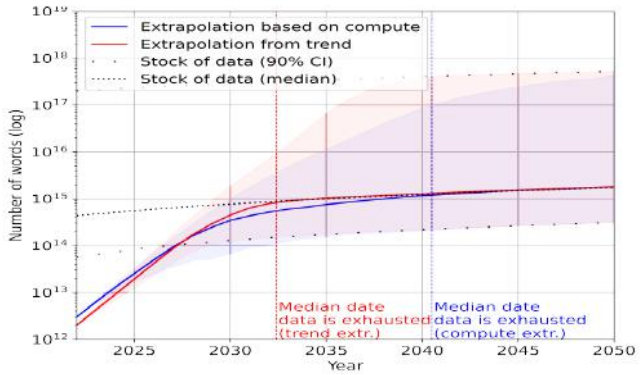
类别	类别	名称	数据来源	发布方
大语言模型数据集	维基百科	Identifying Machine-Paraphrased Plagiarism	维基媒体基金会	德国伍珀塔尔大学、布尔诺孟德尔大学
		Benchmark for Neural Paraphrase Detection	维基媒体基金会	德国伍珀塔尔大学
		Quoref	维基媒体基金会	艾伦人工智能研究所、华盛顿大学
		QuAC (Question Answering in Context)	-	艾伦人工智能研究所、华盛顿大学、斯坦福大学、马萨诸塞大学阿默斯特分校
		TriviaQA	维基媒体基金会	华盛顿大学、艾伦人工智能研究所
		WikiQA	维基媒体基金会	微软研究院
	书籍	BookCorpus	Smashwords	多伦多大学、麻省理工学院
		Project Gutenberg	古腾堡文学档案基金会	古腾堡文学档案基金会
	期刊	ArxivPapers	arXiv	Facebook、伦敦大学学院、DeepMind
		MedQuAD	美国国家卫生研究院	美国国家卫生研究院
		Pubmed	PubMed	马里兰大学
		PubMed Paper Reading Dataset	PubMed	伊利诺伊大学厄巴纳香槟分校、滴滴实验室、伦斯勒理工学院、北卡罗来纳大学教堂山分校、华盛顿大学
		PubMed RCT (PubMed 200k RCT)	PubMed	Adobe Research、麻省理工学院
		MedHop	PubMed	伦敦大学学院、Bloomsbury AI
		unarXive	arXiv	Karlsruhe Institute of Technology
		arXiv Summarization Dataset	arXiv	Georgetown University、Adobe Research
		SCICAP	arXiv	宾夕法尼亚州立大学
	Reddit 链接	OpenWebText	Reddit	华盛顿大学、Facebook AI Research
	Commom Crawl	C4 (Colossal Clean Crawled Corpus)	Common Crawl	Google Research
		Common Crawl	Common Crawl	法国国家信息与自动化研究所、索邦大学
	综合	The Pile	-	EleutherAI
多模态数据集		Conceptual Captions	网络	Google Research
		YFCC100M	Flickr	卡内基梅隆大学、雅虎研究院、International Computer Science Institute
		AVA	-	Google Research
		LAION-400M	Common Crawl	慕尼黑工业大学、EleutherAI、LAION
		COCO	微软	微软
		LAION-5B	Common Crawl	LAION
		Language-Table	-	-

资料来源：OpenDataLab，CSDN，华泰研究

高质量语言数据和图像数据或将耗尽，合成数据有望生成大模型数据

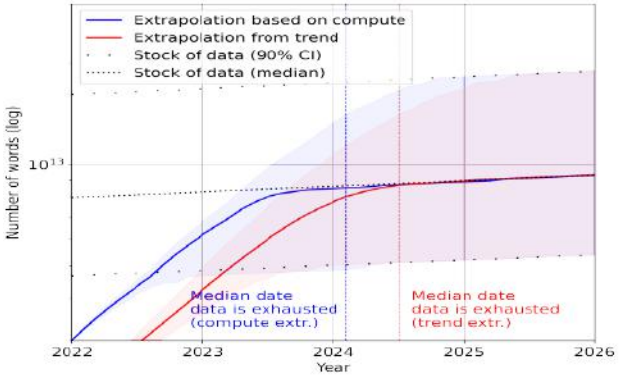
高质量语言数据或将于 2026 年耗尽。数据存量的增长速度远远低于数据集规模的增长速度，如果当前的趋势继续下去，数据集最终将由于数据耗尽而停止增长。在语言模型方面，语言数据的质量有好坏，互联网用户生成的语言数据质量往往低于书籍、科学论文等更专业的语言数据，高质量数据训练出的模型性能更好。根据《Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning》预测，语言数据将于 2030~2040 年耗尽，其中能训练出更好性能的高质量语言数据将于 2026 年耗尽。此外，视觉数据将于 2030~2060 年耗尽。

图表28：低质量语言数据集数据或将于 2030 年耗尽



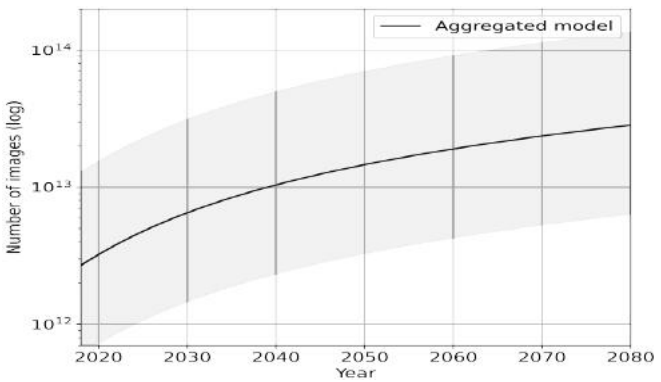
资料来源：Pablo Villalobos et al. “Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning” 2022，华泰研究

图表29：高质量语言数据集数据或将于 2026 年耗尽



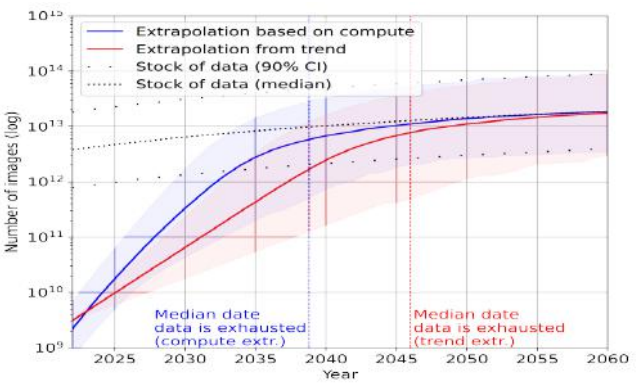
资料来源：Pablo Villalobos et al. “Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning” 2022，华泰研究

图表30：图像数据存量为 $8.11e^{12}$ ~ $2.3e^{13}$



资料来源：Pablo Villalobos et al. “Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning” 2022，华泰研究

图表31：图像数据集数据趋势或将于 2030~2060 年耗尽



资料来源：Pablo Villalobos et al. “Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning” 2022，华泰研究

合成数据或将弥补未来数据的不足。合成数据是计算机模拟或算法生成的带有注释的信息，可以替代真实数据。它可以用于模拟实际情况，补充真实数据的不足，提高数据质量和数量，以及降低数据采集和处理的成本。OpenAI 在 GPT-4 的技术文档中重点提到了合成数据的应用，可见其对该领域的重视。根据 Gartner 的预测，2024 年用于训练大模型的数据中有 60%将是合成数据，到 2030 年大模型使用的绝大部分数据将由人工智能合成。

图表32：GPT-4 技术报告中对合成数据应用的探讨

For closed-domain hallucinations, we are able to use GPT-4 itself to generate synthetic data. Specifically, we design a multi-step process to generate comparison data:

1. Pass a prompt through GPT-4 model and get a response
2. Pass prompt + response through GPT-4 with an instruction to list all hallucinations
   (a) If no hallucinations are found, continue
3. Pass prompt + response + hallucinations through GPT-4 with an instruction to rewrite the response without hallucinations
4. Pass prompt + new response through GPT-4 with an instruction to list all hallucinations
   (a) If none are found, keep (original response, new response) comparison pair
   (b) Otherwise, repeat up to 5x

资料来源：OpenAI “GPT-4 Technical Report” 2023，华泰研究