华泰证券 HUATAI SECURITIES

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）

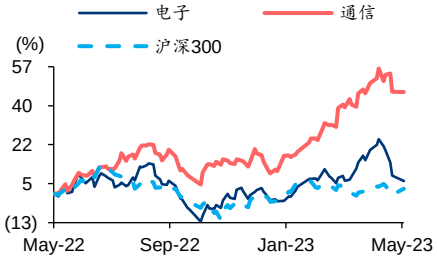
研究员　黄乐平，PhD
SAC No. S0570521050001　leping.huang@htsc.com
SFC No. AUZ066　+(852) 3658 6000

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

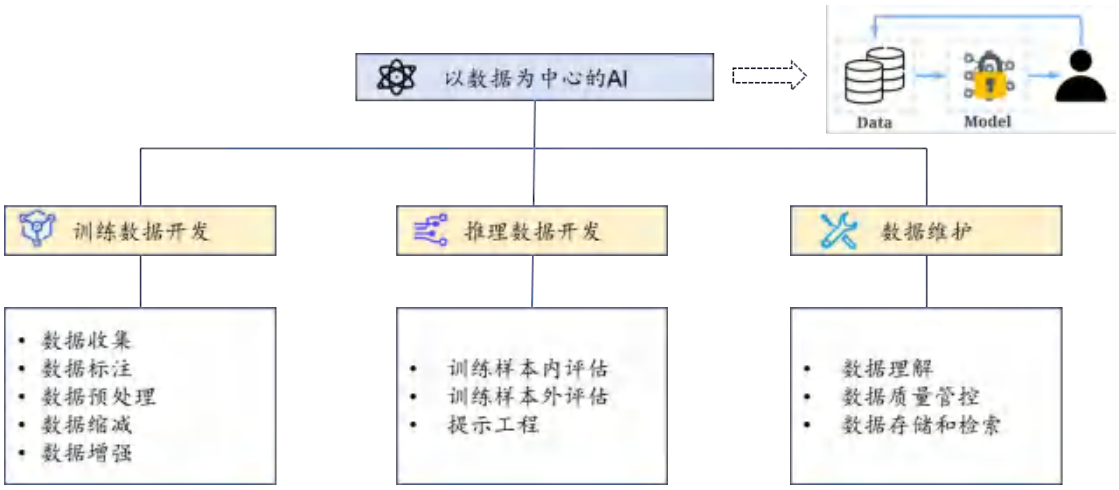
人工智能发展的突破得益于高质量数据的发展。例如，大型语言模型的最新进展依赖于更高质量、更丰富的训练数据集：与 GPT-2 相比，GPT-3 对模型架构只进行了微小的修改，但花费精力收集更大的高质量数据集进行训练。ChatGPT 与 GPT-3 的模型架构类似，并使用 RLHF（来自人工反馈过程的强化学习）来生成用于微调的高质量标记数据。

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

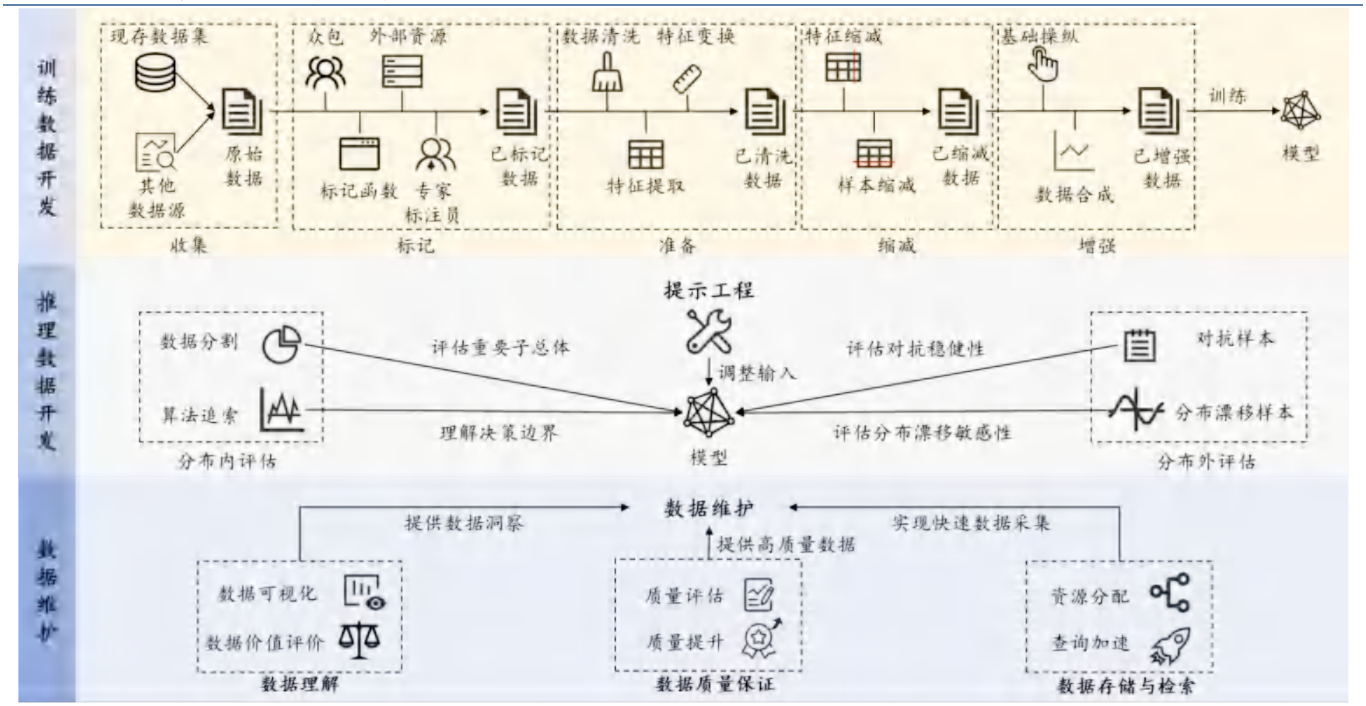
基于此，人工智能领域的权威学者吴承恩发起了“以数据为中心的 AI”运动，即在模型相对固定的前提下，通过提升数据的质量和数量来提升整个模型的训练效果。提升数据集质量的方法主要有：添加数据标记、清洗和转换数据、数据缩减、增加数据多样性、持续监测和维护数据等。**因此，我们认为未来数据成本在大模型开发中的成本占比或将提升，主要包括数据采集，清洗，标注等成本。**

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。

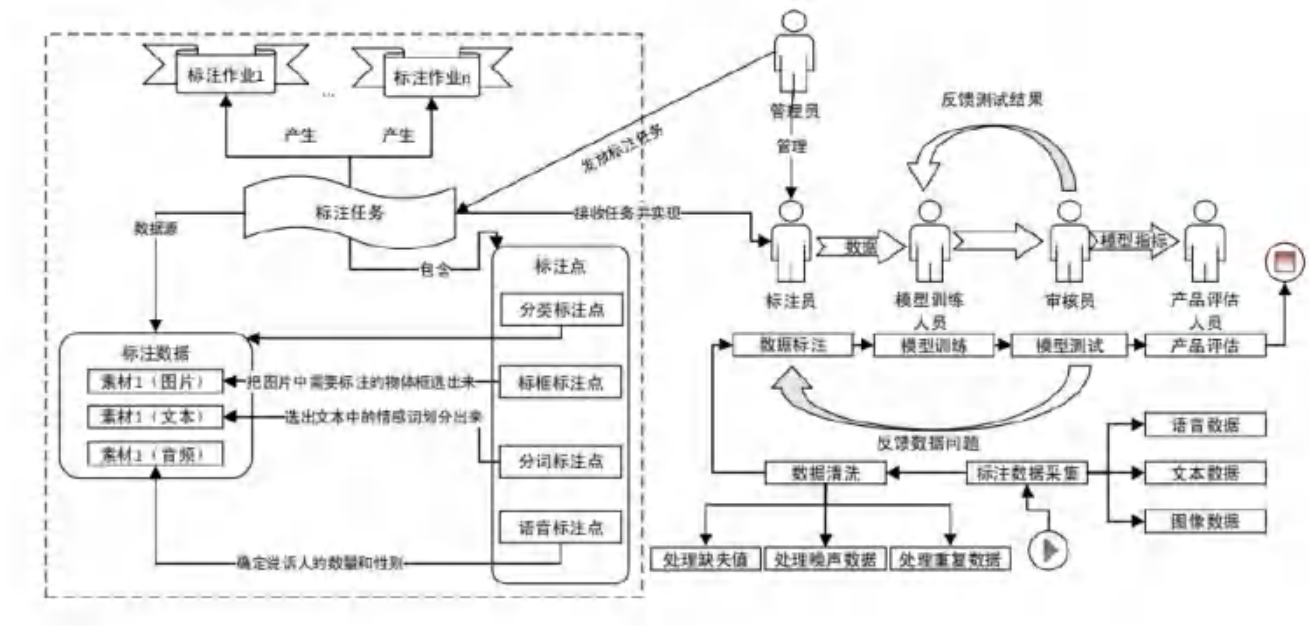
2）大规模：OpenAI 在《Scaling Laws for Neural Language Models》中提出 LLM 模型所遵循的“伸缩法则”（scaling law），即独立增加训练数据量、模型参数规模或者延长模型训练时间，预训练模型的效果会越来越好。

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。

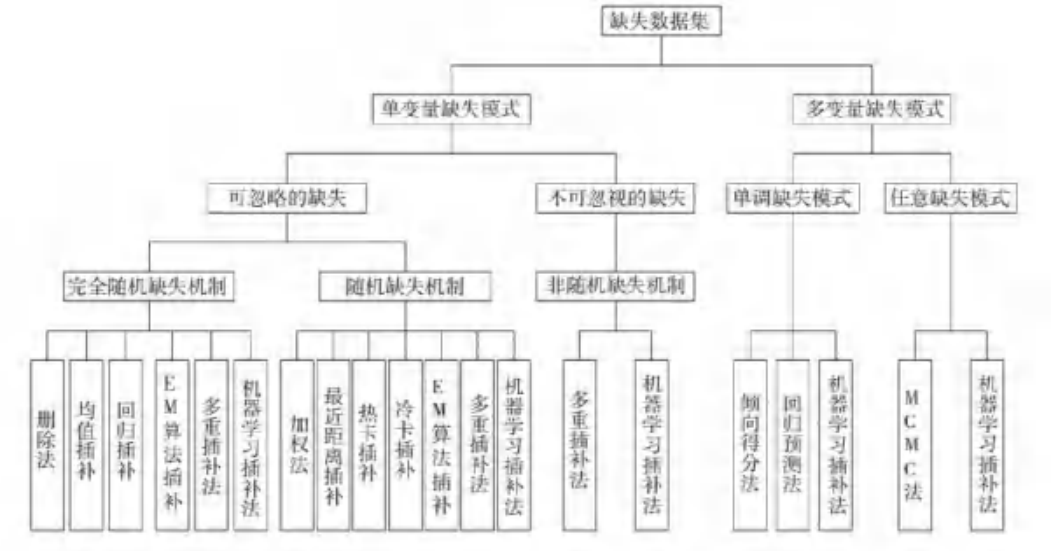
图表5：数据采集三种常见方式

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。

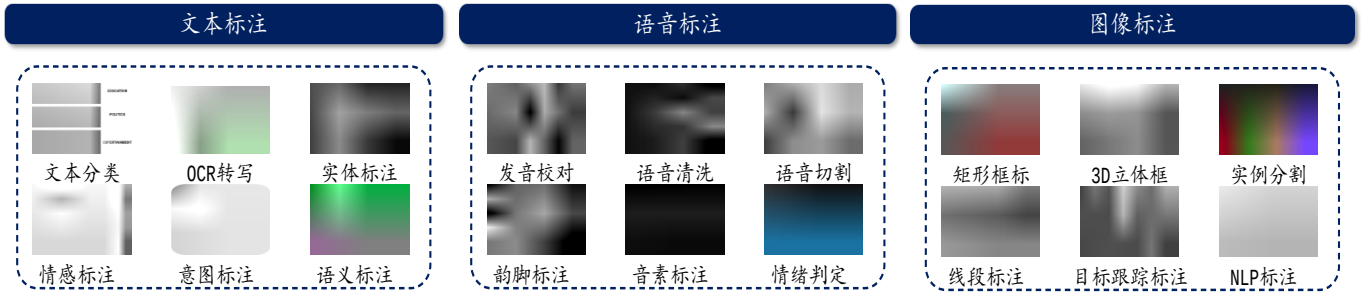
图表6：缺失数据的处理方法



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。

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）

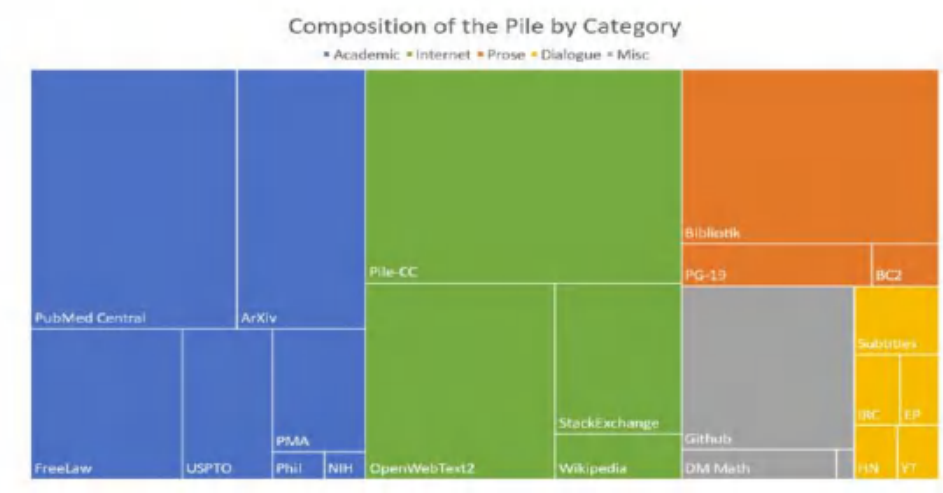
排名	域	Token（百万）	占比
1	Google Patents	750	0.48%
2	The NY Times	100	0.06%
3	Los AngelesTimes	90	0.06%
4	The Guardian	90	0.06%
5	PLoS	90	0.06%
6	Forbes	80	0.05%
7	HuffingtonPost	75	0.05%
8	Patents.com	71	0.05%
9	Scribd	70	0.04%
10	Washington Post	65	0.04%
11	The Motley Fool	61	0.04%
12	IPFS	60	0.04%
13	Frontiers 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。

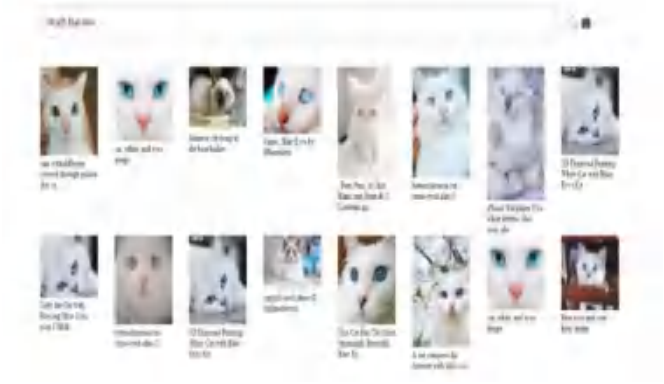
Conceptual Captions 数据集：图像标题注释数据集，其中包含的图像比 MS-COCO 数据集多一个数量级，并代表了更广泛的图像和图像标题风格。通过从数十亿个网页中提取和过滤图像标题注释来实现这一点。

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。

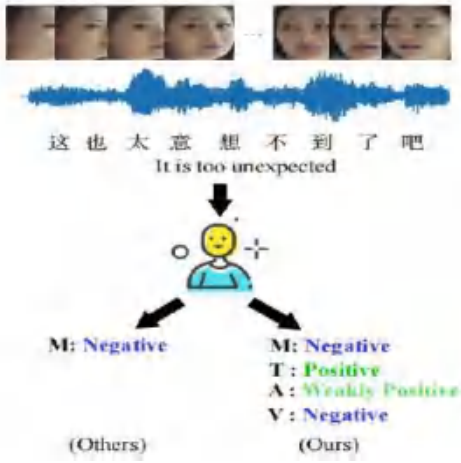
图表23：YFCC100M 数据集中 100 万张照片样本的全球覆盖

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

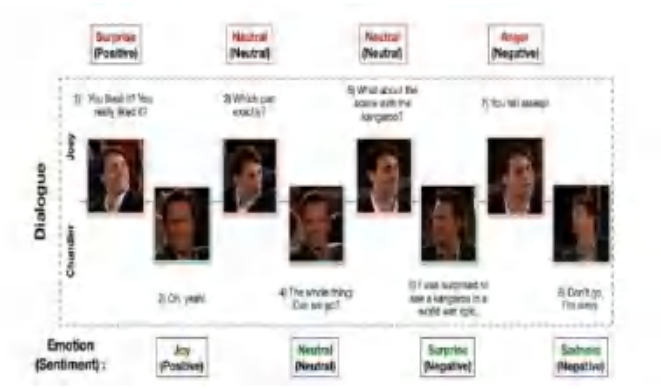
IEMOCAP 数据集：南加州大学语音分析与解释实验室(SAIL)收集的一种新语料库，名为“交互式情感二元动作捕捉数据库”(IEMOCAP)。该数据库记录了 10 位演员在面部、头部和手上的二元会话，这些标记提供了他们在脚本和自发口语交流场景中面部表情和手部动作的详细信息。语料库包含大约 12 小时的数据。详细的动作捕捉信息、激发真实情绪的交互设置以及数据库的大小使这个语料库成为社区中现有数据库的有价值的补充，用于研究和建模多模态和富有表现力的人类交流。

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

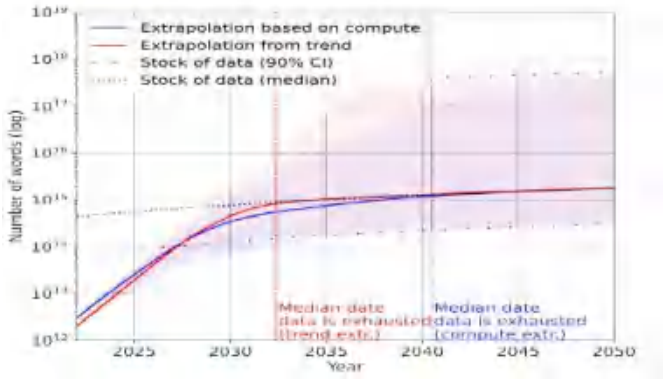
类别	类别	名称	数据来源	发布方
大语言模型数据集	维基百科	Identifying Machine-Paraphrased Plagiarism	维基媒体基金会	德国伍珀塔尔大学、布尔诺孟德尔大学
		Benchmark for Neural Paraphrase Detection	维基媒体基金会	德国伍珀塔尔大学
		Quoref	维基媒体基金会	艾伦人工智能研究所、华盛顿大学
		QuAC (Question Answering in Context)	-	艾伦人工智能研究所、华盛顿大学、斯坦福大学、马萨诸塞大学阿默斯特分校
		TriviaQA	维基媒体基金会	华盛顿大学、艾伦人工智能研究所
		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

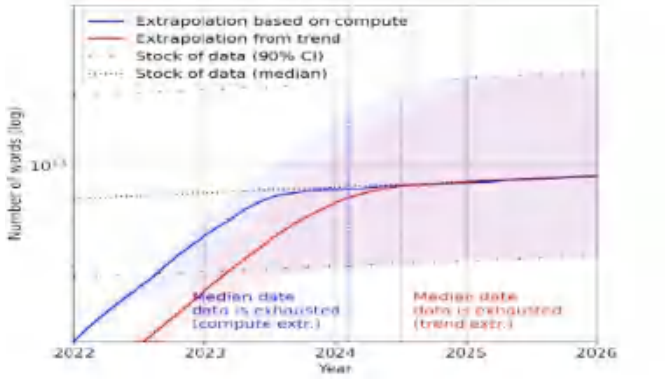
高质量语言数据或将于 2026 年耗尽。数据存量的增长速度远远低于数据集规模的增长速度，如果当前的趋势继续下去，数据集最终将由于数据耗尽而停止增长。在语言模型方面，语言数据的质量有好坏，互联网用户生成的语言数据质量往往低于书籍、科学论文等更专业的语言数据，高质量数据训练出的模型性能更好。根据《Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning》预测，语言数据将于 2030~2040 年耗尽，其中能训练出更好性能的高质量语言数据将于 2026 年耗尽。此外，视觉数据将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



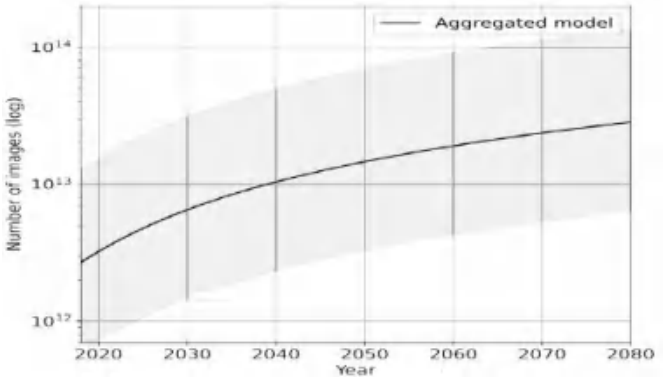
资料来源：Pablo Villalobos et al. “Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning” 2022，华泰研究

图表29：高质量语言数据集数据或将于 2026 年耗尽



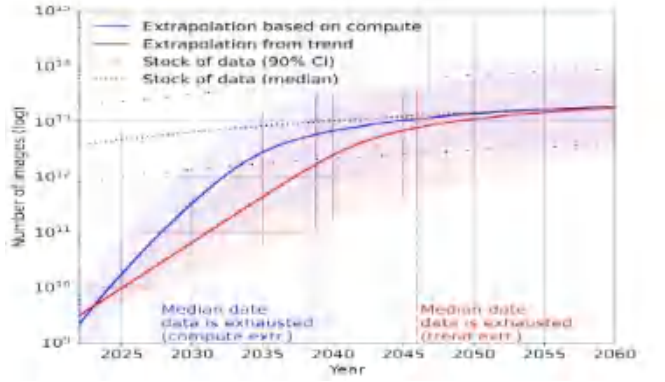
资料来源：Pablo Villalobos et al. “Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning” 2022，华泰研究

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。

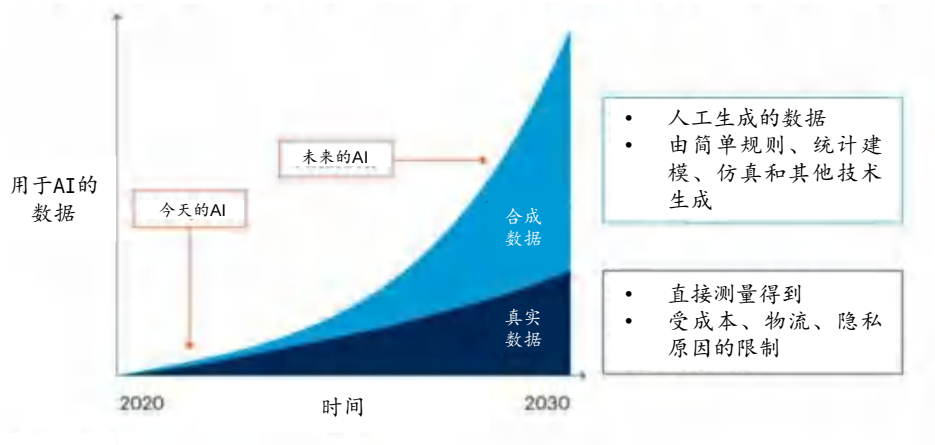
图表32：GPT-4 技术报告中对合成数据应用的探讨

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

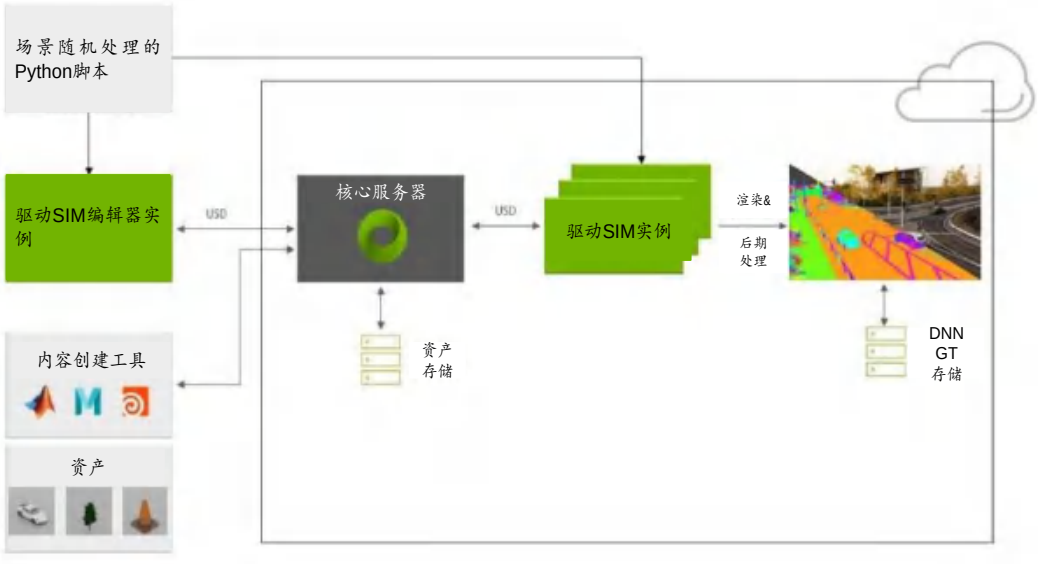
图表33：到 2030 年 AI 模型中的合成数据将完全盖过真实数据



资料来源：Gartner，华泰研究

合成数据有望首先在金融、医疗和汽车等诸多领域落地。在金融行业，金融机构可以在不提供敏感的历史交易信息前提下，通过合成数据集训练量化交易模型提升获利能力，也可以用来训练客服机器人以改善服务体验；在生物医药行业，可以通过合成数据集，在不提供患者隐私信息的条件下训练相关模型完成药物研发工作；在自动驾驶领域，可以通过合成数据集模拟各种驾驶场景，在保障人员和设备安全的条件下提升自动驾驶能力。

图表34：NVIDIA Omniverse——用户可使用 Python 为自动驾驶车辆生成合成数据



资料来源：英伟达官网，华泰研究

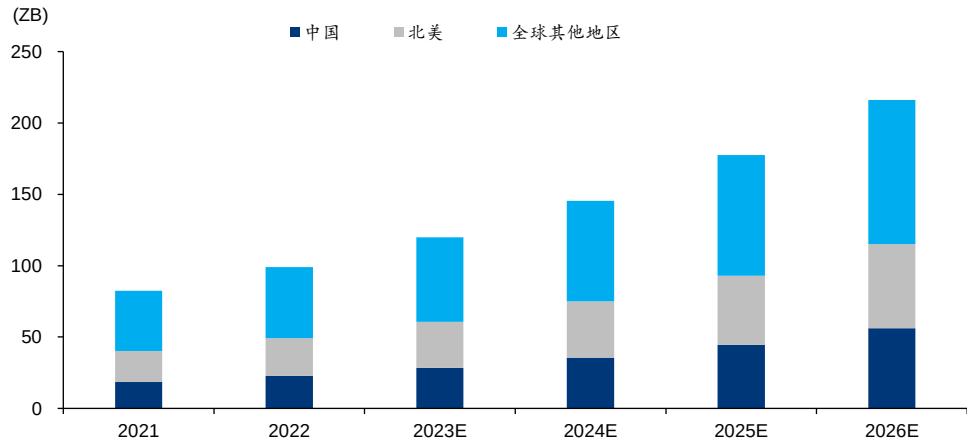
数字中国战略助力中国 AI 大模型数据基础发展

数据将是 AI 大模型的关键竞争要素之一，数字中国战略有望助力我国 AI 大模型训练数据集的发展。近日,中共中央、国务院印发了《数字中国建设整体布局规划》，数据要素为数字中国建设战略中的关键一环。我们认为当前国内虽然数据资源丰富，但优质的中文大模型训练语料仍然稀缺。数字中国战略将极大促进我国数据要素市场的完善，从数量和质量两个维度助力中文大模型数据集的发展：1）数量方面，各地数据交易所设立并运营后，数据资源将能够在各行业、各企业之间自由流通，缓解大模型训练数据数量不足的问题；2）质量方面，国内数据服务产业有望蓬勃发展，未来数据服务商将提供数据标注、清洗、维护等服务，大数据产业专业化分工将助力大模型训练数据集质量提升。

中国 AI 大模型数据集从哪里来

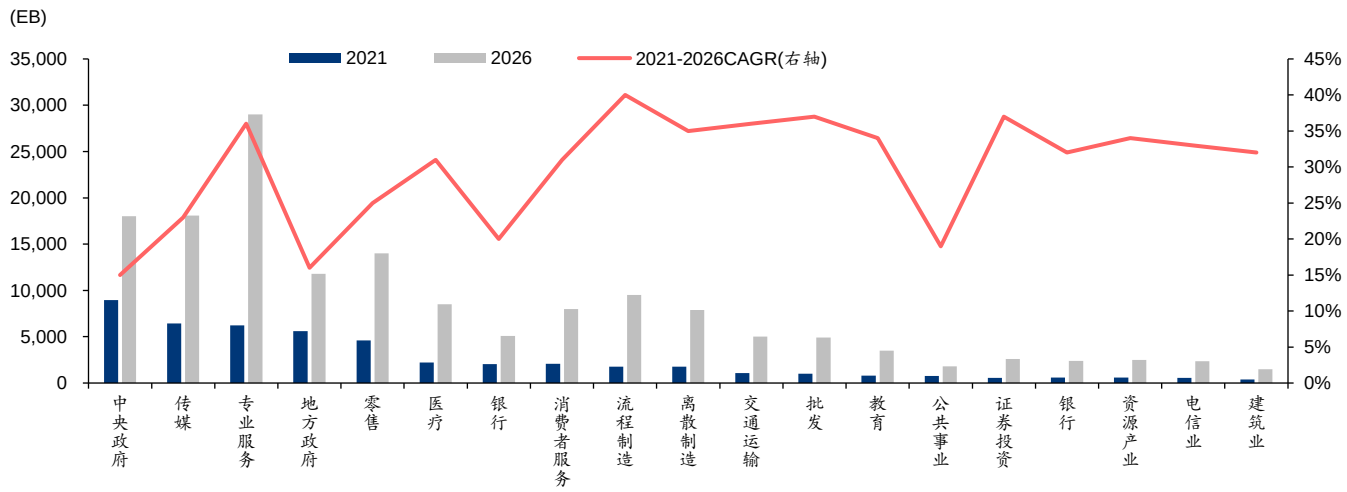
国内各行业数据资源丰富，2021-2026 年数据量规模 CAGR 高于全球，数据主要来源于政府/传媒/服务/零售等行业。据 IDC，2021-2026 年中国数据量规模将由 18.51ZB 增长至 56.16ZB，CAGR 达到 24.9%，高于全球平均 CAGR。从数据来源看，国内各行业数据差异化发展，2021 年政府、传媒、专业服务、零售等行业数据量占比较高，分别为 30.4%、13.4%、13.0%、9.6%，其中接近 90%的数据为非结构化数据，这也要求了海量数据采集设备和软件的互联互通以及互动互控。另外随着智能化转型的深入，制造、交通运输、批发、教育等行业数据规模在未来也拥有较大的增长潜力，2021-2026 年数据量增长 CAGR 将分别达到 37.6%、36.1%、37.1%、34.0%。

图表35：2021-2026 中国数据量规模 CAGR 达到 24.9%，位居全球第一



资料来源：IDC Global DataSphere, 2022，华泰研究

图表36：国内各行业数据量分布及增长预测



资料来源：IDC，华泰研究

尽管国内数据资源丰富，但由于数据挖掘不足，数据无法自由在市场上流通等现状，优质中文优质数据集仍然稀缺。目前中文优质数据仍然稀缺，如 ChatGPT 训练数据中中文资料比重不足千分之一，为 0.0991%，而英文资料占比超过 92.6%。据加利福尼亚大学和 Google 研究机构发现，机器学习和自然语言处理模型使用的数据集 50%由 12 家 Top 机构提供，其中 10 家为美国机构，1 家为德国机构，仅 1 家机构来自中国，为香港中文大学。值得一提的是，数据集与数据机构的基尼系数有升高的趋势，即数据集被少数 Top 机构或特定数据库掌控的集中有所增加。

图表37：数据集分布及发展趋势

注：左：截至 2021 年 6 月，每个机构的数据集使用情况图。网点大小表示使用次数。蓝点表示营利机构，橙点表示非营利机构。机构占使用量的 50%以上。右图：机构和数据集在整个 Papers With Code 数据集上使用集中度的基尼系数。圆点大小表示当年的使用次数。

资料来源：Bernard Koch et al. “Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research” 2021，华泰研究

我们认为国内缺乏高质量数据集主要有以下几方面的原因：1）高质量数据集需要巨大资金投入，当前国内数据挖掘和数据治理的力度不足；2）国内相关公司不具有开源意识，数据无法在市场上自由流通；3）国内相关公司成立较晚，数据积累相对国外公司更少；4）学术领域中文数据集受重视程度低；5）国产数据集市场影响力及普及度较低等。

图表38：国内缺乏高质量数据集的主要原因

原因	解释
高质量数据集需要巨大资金投入	一个好的数据集应该从顶层设计、标注规范、标注质量把控以及发布后更新升级各个方面严格把关，这个过程是需要经费、人力等综合因素的投入，一般是长期投入的企业或者自然基金项目才有动力开展这样的工作
国内相关公司不具有开源意识	互联网公司拥有服务自身应用的数据集因竞争原因不愿公开，工业界里一些公司因保密原因不愿公开数据，传统金融领域公司重视合规保护；同时开源政策及社区不活跃，开源支持不到位，后期服务跟不上
国内相关公司数据积累相比国外更少	国内互联网公司成立普遍晚于国外（亚马逊 1994 年，阿里巴巴 1999 年成立；谷歌 1998 年，百度 2000 年成立），早年中国互联网公司以模仿国外的业务为主，在数据上的沉淀和积累较少，特别是特有中文数据
学术领域中文数据集受重视程度低	使用中文数据集的论文往往不好发表，而高质量、受认可的中文会议期刊的数量不够多。从而使得学术界在发展中文数据集上的动力不足
国产数据集市场影响力及普及度较低	目前国内大部分数据集产品仅限于企业内部使用，未经过市场检验，导致数据库创新能力不足。使得企业开发数据集的意愿较低，同时大模型训练普遍产学结合，学术界对中文数据集的忽视也会影响到业界

资料来源：Datawhale，华泰研究

目前国内科技互联网头部企业主要基于公开数据及自身特有数据差异化训练大模型。具体而言，百度“文心”大模型训练特有数据主要包括万亿级的网页数据，数十亿的搜索数据和图片数据等。阿里“通义”大模型的训练数据主要来自阿里达摩院。腾讯“混元”大模型特有的训练数据主要来自微信公众号，微信搜索等优质数据。华为“盘古”大模型的训练数据出公开数据外，还有 B 端行业数据加持，包括气象，矿山，铁路等行业数据。商汤“日日新”模型的训练数据中包括了自行生成的 Omni Objects 3D 多模态数据集。

图表39：国内科技互联网厂商训练大模型基于的数据基础

厂商	模型	客户	模型基于的数据集
百度	文心一言	C 端为主	万亿级网页数据，数十亿的搜索数据和图片数据，百亿级的语音日均调用数据以及 5500 亿事实的知识图谱等
阿里	通义	C 端+B 端	训练数据来自于阿里达摩院，包含了大量的语言和文本数据，以及各类专业知识和技术文档等
腾讯	混元	C 端+B 端	公开数据集 + 腾讯内部数据，如：微信公众号内容（优质数据）、微信搜索、腾讯广告数据等
华为	盘古	B 端为主	B 端行业数据：气象，矿山，铁路等行业数据 + 公开数据集：4400 万个多回合对话会话，1 亿个话语和 13 亿个 token（PANGU-BOT 基于的数据）
商汤	日日新	C 端+B 端	基于多模态数据集，自行开发了 Omni Objects 3D 数据集（包含了 190 个类别，超过 6000 个物体，有大量的真实物体的扫描的数据）

资料来源：各公司官网，华泰研究

未来专业及垂直内容平台有望成为国内优质中文数据集的重要来源：1）专业内容平台：知乎作为问答知识平台，拥有超过 4300 万创作者产生的超过 3.5 亿条优质中文问答内容，涉及政治，经济，文化，历史，科技等几乎全部门类。其问答的数据形式天然适合作为大语言类模型训练使用。微信公众号作为内容分享平台，背靠国民级应用微信生态链，2022 年公众号产出超 3.9 亿篇文章，其中既有专业领域内容分析，也有时事热点分析，这些内容对语言模型的训练迭代有重要作用。**2）垂类内容平台：**参考彭博基于金融垂类数据发布 BloombergGPT 案例，国内在金融，医疗，生物等行业公司的数据可以作为细分领域大模型精确训练的基础。

中国大模型如何构建数据集#1：LLM

我们选取了在其论文中详细阐述如何构建预训练数据集的三个大语言模型，研究中国大模型预训练数据集的来源。我们发现：1）类似海外大语言模型，中国大语言模型的预训练数据集也主要来自互联网抓取数据（如 Common Crawl、中文公共社交媒体抓取等）、网络百科全书（如百度百科、搜狗百科）、书籍等等；2）充分借助已有的高质量开源数据集，例如对 Common Crawl 等进行二次处理。

图表40：中国大语言模型数据集构成

公司	NLP大模型	发布时间	最大参数量（B）	数据集（TB_文本）	数据来源
百度	Plato-XL	2021.9	11	-	中文：公共领域的社交媒体、英文：Reddit评论
华为	盘古	2021.4	200	10	开源数据集、百科全书、电子书、Common Crawl、新闻
腾讯	WeLM	2022.9	10	80	Common Crawl、新闻、书籍、流行的在线论坛以及学术著作

资料来源：Siqi Bao et al. “PLATO-XL: Exploring the Large-scale Pre-training of Dialogue Generation” 2021，Wei Zeng et al. “PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation” 2021，Hui Su et al. “WeLM: A Well-Read Pre-trained Language Model for Chinese” 2022，华泰研究

百度 Plato-XL 大模型：百度于 2021 年发布 PLATO-XL，包括中英文 2 个对话模型，预训练语料规模达千亿级 token，模型规模高达 110 亿参数。预训练语料库分为：1)英语：会话样本从 Reddit 评论中提取，这些评论由第三方收集，并在 pushshift.io 上公开发布，遵循了 PLATO-2 的精心清洗过程；2）中文：预训练数据来自公共领域的社交媒体，过滤后训练集中有 1.2 亿个样本。

华为盘古大模型：华为于 2021 年发布盘古，是业界首个 2000 亿参数以中文为核心的预训练生成语言模型，目前开源了盘古 α 和盘古 α 增强版两个版本，并支持 NPU 和 GPU 两个版本，支持丰富的场景应用，在知识问答、知识检索、知识推理、阅读理解等文本生成领域表现突出，具备较强的少样本学习的能力。

图表41：华为盘古大模型 1.1TB 中文文本语料库数据组成

数据集	大小（GB）	数据来源	处理步骤
开源数据集	27.9	15 个开源数据集：DuReader、百度 QA、CAIL2018、搜狗 CA 等	格式转换和文本重复删除
百科全书	22	百度百科、搜狗百科等	文本重复删除
电子书	299	各种主题的电子书(如小说、历史、诗歌和古文等)	基于敏感词和模型垃圾邮件过滤
Common Crawl	714.9	来自 Common Crawl 的 2018 年 1 月至 2020 年 12 月的网络数据	所有步骤
新闻	35.5	1992 年至 2011 年的新闻数据	文本重复删除

资料来源：Wei Zeng et al. “PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation” 2021，华泰研究

腾讯 WeLM 大模型：腾讯于 2022 年发布 WeLM，数据来源主要分为三部分：1）Common Crawl：Common Crawl 于 2020.08 至 2022.01 期间的文本内容，使用 langdetect 工具过滤掉非中文的文本；2）特定领域语料库：混合了来自各种来源的数据，包括新闻、书籍、流行在线论坛以及学术著作，仅中文数据。3）英文数据：添加了从上述来源收集到的约 750GB 的英语数据。数据中有大量的噪音如胡言乱语或模板文本、冒犯性语言、占位符文本和源代码等，首先应用一组基于规则的过滤器，再在剩余的数据上手动构建好坏数据分类器提升数据清理泛化能力。

图表42：WeLM 大模型训练语料库统计

来源	%过滤	#剩余 Tokens	预训练比例
Common Crawl	92%	198.5B	50.6%
书籍	40.9%	61.9B	38.7%
新闻	7.5%	1.91B	6.7%
论坛	6.7%	1.0B	3.5%
学术著作	2.5%	0.39B	0.5%

资料来源：Hui Su et al. “WeLM: A Well-Read Pre-trained Language Model for Chinese” 2022，华泰研究

中国大模型如何构建数据集#2：多模态大模型

我们选取了在其论文中详细阐述如何构建预训练数据集的三个多模态模型，研究中国大模型预训练数据集的来源。我们发现网页抓取、自有数据和开源数据集是多模态大模型数据集的重要来源：1）网页抓取图文对：例如阿里 M6 大模型、百度 ERNIE-ViLG 大模型都从网页中抓取文本-图片对，然后经过一定过滤，形成最终数据集的一部分；2）自有数据：例如阿里 M6 大模型有来自电商的图文数据，百度 ERNIE-ViLG 大模型从内部图像搜索引擎中收集查询文本和对应的用户点击图像；3）开源数据集：例如百度 ERNIE-ViLG 大模型的部分图文对数据来自开源的 CC 和 CC12M，并通过百度翻译 API 翻译。

图表43：中国多模态模型数据集构成

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）	数据来源
阿里	M6	2021.3	100	-	百科全书、社区QA、论坛讨论、Common Crawl、抓取网页、电商数据
百度	ERNIE-ViLG	2021.12	10	145	中文网页、内部图像搜索引擎、CC、CC12M
上海人工智能实验室等	InternVideo	2022.12	1.3	-	Kinetics-400、WebVid2M、WebVid10M、HowTo100M、AVA、Something-Something V2、Kinetics-710、自采视频

资料来源：Junyang Lin et al. “M6: A Chinese Multimodal Pretrainer” 2021，Han Zhang et al. “ERNIE-ViLG: Unified Generative Pre-training for Bidirectional Vision-Language Generation” 2021，Yi Wang et al. “InternVideo: General Video Foundation Models via Generative and Discriminative Learning” 2022，华泰研究

阿里 M6 大模型：于 2021 年发布，参数规模达到 1000 亿。阿里构建了最大的中文多模态预训练数据集 M6-Corpus，包含超过 1.9 TB 图像和 292GB 文本，涵盖了百科全书、问答、论坛讨论、产品说明等类型的数据集。研究人员设计了完善的清洁程序：1）文本数据：删除 HTML 标记和重复的标点符号，只保留中文和英文的字符和标点符号。删除短于 5 个字符的标题和短于 15 个字符的文本内容。使用“内部垃圾邮件检测器”筛选包含某些政治问题、色情或脏话等不合适的句子。建立一个语言模型进行评估文本的困惑程度，去掉困惑程度高的句子；2）图片数据：只有超过 5000 像素的图像才有资格被保留用于预训练。

图表44：M6 预训练数据集构成

来源	模态	图像（M）	Tokens(B)	段落（M）	平均长度图像大小（TB）	文本大小(GB)
百科全书	纯文本	-	31.4	34.0	923.5	-	65.1
社区 QA	纯文本	-	13.9	113.0	123.0	-	28.8
论坛讨论	纯文本	-	8.7	39.0	223.1	-	18.0
Common Crawl	纯文本	-	40.3	108.7	370.7	-	83.3
百科全书	图像&文本	6.5	7.9	10.4	759.6	0.1	15.0
抓取网页	图像&文本	46.0	9.1	106.0	85.8	1.5	70.0
电商	图像&文本	8.0	0.5	8.5	62.1	0.3	12.2
总计		60.5	111.8	419.6	266.4	1.9	292.4

注：电商数据包含 260k 来自淘宝的成对产品描述和产品图片
资料来源：Junyang Lin et al. “M6: A Chinese Multimodal Pretrainer” 2021，华泰研究

百度 ERNIE-ViLG 大模型：于 2021 年发布，参数规模达到 100 亿。百度构建了一个由超过 1.45 亿对高质量中文图像-文本对组成的大规模图像-文本数据集，数据来源如下：1）中文网页。从各种中文网页中抓取了 8 亿对原始的中文替代文字描述和图片，进行了几个步骤的过滤，总共收获了 7000 万对文本-图片，过滤规则主要包括文本长度、文本内容和图像-文本相似度；2）图片搜索引擎：从内部图像搜索引擎中收集了大约 6000 万个查询文本和相应的用户点击图像；3）开源图像-文本数据集：从 CC 和 CC12M 中共收集了 1500 万文本图像对，这些数据集中的字幕通过百度翻译 API 翻译成中文。

InternVideo 大模型：由上海人工智能实验室等、南大、港大、复旦、中科院深圳先进技术研究院等于 2022 年发布，使用了 6 个来自各个领域的开源数据集和自采视频片段。

图表45：InternVideo 预训练过程中使用的数据集统计

预训练数据集	域	样本剪辑	帧数×采样率
Kinetics-400	Youtube 视频	240k	16 × 4
WebVid2M	网络视频	250k	16 × 4
WebVid10M	网络视频	10M	16 × 4
HowTo100M	Youtube 视频	1.2M	16 × 4
AVA	电影	21k	16 × 4
Something-Something V2	剧本镜头	169k	16 × 2
自采视频	Youtube, Instagram	250k	16 × 4
Kinetics-710	Youtube 视频	680k	16 × 4

资料来源：Yi Wang et al. “InternVideo: General Video Foundation Models via Generative and Discriminative Learning” 2022，华泰研究

中国开源数据集#1：大语言模型数据集

DuReader 数据集：于 2018 年由百度发布。DuReader 是一个大规模的开放域中文机器阅读理解数据集。该数据集由 200K 问题、420K 答案和 1M 文档组成，是迄今为止最大的中文 MRC 数据集。问题和文档基于百度搜索和百度知道，答案是手动生成的。该数据集还提供了问题类型注释——每个问题都被手动注释为实体、描述或是否以及事实或意见之一。

图表46：DuReader 汉语六种题型示例(附英文注释)

	Fact	Opinion
Entity	iphone哪天发布 On which day will iphone be released	2017最好看的十部电影 Top 10 movies of 2017
Description	消防车为什么是红的 Why are firetrucks red	丰田卡罗拉怎么样 How is Toyota Carola
YesNo	39.5度算高烧吗 Is 39.5 degree a high fever	学围棋能开发智力吗 Does learning to play go improve intelligence

资料来源：Wei He et al. “DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications” 2017，华泰研究

WuDaoCorpora 数据集：于 2021 年由北京人工智能研究院、清华大学和循环智能联合发布。WuDaoCorpora 是北京智源研究院最新构建的高质量数据集，悟道文本数据集采用 20 多种规则从 100TB 原始网页数据中清洗得出最终数据集，注重隐私数据信息的去除，从源头上避免 GPT-3 存在的隐私泄露风险；包含教育、科技等 50+个行业数据标签，可以支持多领域预训练模型的训练。

图表47： WuDaoCorpora 示例


Example
{
"id": "200023",
"url":"[illegible]
html",
"dataType": news,
"title": "3D微视频-共同家园 (3D Micro Video - Common Homeland)",
"content":" 黄海之滨，青青岛城。八方宾客齐聚，上合组织青岛峰会今天举行。人民日报推出3D微视频，共同家园。让我们跟随习近平主席的同期声，以全新方式讲述上合故事。诞生17年，上海合作组织日益壮大。从地区安全到经贸合作，从人文交流到开放包容，上合组织合作硕果累累。千帆过尽，不忘初心。上海精神始终贯穿上合组织发展历程。携手建设共同家园，中国智慧为上合组织注入新的发展动力。新的发展方位，新的历史起点，上合组织发展迎来历史性机遇。从青岛再度扬帆出发，上合组织未来发展新蓝图正在绘就。青青之岛，和合共生。(On the shore of the Yellow Sea, Green Island City. The Shanghai Cooperation Organization (SCO) Qingdao Summit opens today. People's Daily launches 3D micro video, Common Home. Let's follow the voice of President Xi Jinping and tell the story of Shanghai Ho in a new way. Since its founding 17 years ago, the Shanghai Cooperation Organization has been going from strength to strength. From regional security to economic and trade cooperation, from people-to-people exchanges to openness and inclusiveness, the SCO cooperation has yielded fruitful results. After a thousand sails have been completed, we will never forget our original aspiration. The Shanghai Spirit has been throughout the development of the SCO. Joining hands to build a common home, the Chinese wisdom has injected new impetus into the development of the SCO. A new development juncture and a new historical starting point represent a historic opportunity for the SCO's development. Starting from Qingdao, we are drawing up a new blueprint for the SCO's future development. Green island, harmonious coexistence.)"
"dataCleanTime": "2021-01-09 00:59:15"
}

资料来源：Sha Yuan et al. “WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models” 2021，华泰研究

CLUECorpus2020 数据集：于 2020 年由 CLUE 发布。CLUECorpus2020 是一个可以直接用于语言模型预训练或语言生成等自监督学习的大型语料库，它有 100G 的原始语料库，包含 350 亿个汉字，这些语料库来自 Common crawl。

CAIL2018 数据集：于 2018 年由清华大学、北京大学、中国科学院软件研究所和中国司法大数据研究院联合发布。CAIL2018 是第一个用于判决预测的大规模中国法律数据集，收录了中国最高人民法院公布 260 万件刑事案件，是现有判决预测工作中其他数据集的数倍。对判断结果的注释也更加详细和丰富。它由适用的法律条款、指控和刑期组成，根据案件的事实描述而定。

图表48： CAIL2018 示例

Fact	Relevant Law Article	Charge	Prison Term	Defendant
被告人胡某...	刑法第234条	故意伤害	12个月	胡某
The Defendant Hu...	234th article of criminal law	intentional injury	12 months	Miss./Mr. Hu

资料来源：Chaojun Xiao et al. “CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction” 2018，华泰研究

Math23K 数据集：于 2017 年由腾讯人工智能实验室发布。Math23K 是为解决数学问题而创建的数据集，数据包含从在线教育网站上抓取的 6 万多个中文数学单词问题，都是小学生真正的数学应用题，有 23,161 个标有结构化方程和答案的问题。

图表49： Math23K 和其他几个公开数据集对比

数据集	问题	模板	句子	单词	问题类型
Alg514	514	28	1.62k	19.3k	代数、线性
Dolphin1878	1,878	1,183	3.30k	41.4k	数字应用题
DRAW-1K	1,000	-	6.23k	81.5k	代数、线性、一元
Math23K	23,161	2,187	70.1k	822k	代数、线性、一元

资料来源：Yan Wang et al. “Deep Neural Solver for Math Word Problems” 2017，华泰研究

Ape210K 数据集：于 2020 年由猿辅导 AI Lab 和西北大学联合发布。Ape210K 是一个新的大规模和模板丰富的数学单词问题数据集，包含 210K 个中国小学水平的数学问题，是 Math23K 的 9 倍。每个问题都包含黄金答案和得出答案所需的方程式，有 56K 个模板，是 Math23K 的 25 倍。

图表50： Ape210K 与现有数学应用题数据集的比较

数据集	问题	模板	w/ EC(%)
Alg514 (Kushman et al., 2014)	514	28	-
Dolphin1878 (Shi et al., 2015)	1,878	1,183	-
AllArith (Roy and Roth, 2017)	831	-	-
MAWPS (Koncel-Kedziorski et al., 2016)	2,373	-	-
Dolphin18K (Huang et al., 2016)	18,460	5,871	-
Math23K (Wang et al., 2017)	23,160	2,187	0.80%
Ape210K	**210,488**	**56,532**	**37.70%**

注：“w/ EC(%)”指具有除 1 和 π 以外的外部常数的方程的百分比。

资料来源：Wei Zhao et al. “Ape210K: A Large-Scale and Template-Rich Dataset of Math Word Problems” 2020，华泰研究

DRCD 数据集：于 2018 年由台达研究中心和台达电子联合发布。一个开放领域的传统中文机器阅读理解数据集，包含来自 2108 篇维基百科文章的 10014 个段落和由注释者生成的 33,941 个问答对。

图表51： DRCD 的问题类型

问题类型	占比（%）	示例关键词
How	5.3	如何
What	28.42	什么
When	13.59	何时
Where	4.98	哪里
which	30.96	何种
Who	10.46	谁
Why	0.27	为何
Other	5.97	X

资料来源：Chih Chieh Shao et al. “DRCD: a Chinese Machine Reading Comprehension Dataset” 2018，华泰研究

FCGEC 数据集：于 2022 年由浙江大学和华为联合发布。FCGEC 用于检测、识别和纠正语法错误，是一个人工标注的多参考语料库，由 41,340 个句子组成，主要来自公立学校语文考试中的选择题。

图表52：不同汉语语法纠错语料库的对比

语料库	来源	范式	句子	#Error	#Refs	#Length
NLPCC(2018)	CFL	代码错误	2000	1983(99.15%)	1.1	29.7
CGED	CFL	代码错误	30145	25837(85.71%)	1.0	46.6
CTC-Qua(2021)	Native	代码错误	972	482(49.59%)	1.0	48.9
MuCGEC(2022)	CFL	重写	7063	6544(92.65%)	2.3	38.5
FCGEC	**Native**	**操作**	**41340**	**22517(54.47%)**	**1.7**	**53.1**

注：#Error 中的数字表示语料库中不正确句子的百分比。#Refs 表示平均每个句子中包含的引用数，#Length 表示每个句子中平均包含的字符数。

资料来源：Lvxiaowei Xu et al. “FCGEC: Fine-Grained Corpus for Chinese Grammatical Error Correction” 2022，华泰研究

E-KAR 数据集：于 2022 年由复旦大学、字节跳动人工智能实验室和 BrainTechnologies,Inc. 联合发布。数据集包含来自公务员考试的 1,655 个（中文）和 1,251 个（英文）问题，这些问题需要深入的背景知识才能解决。

图表53：E-KAR 与以往类比基准的比较

数据集	语言	数据大小 (训练/有效/测试)	候选人中的术语	已经解释
SAT	英文	0 / 37 / 337	2	✗
Google	英文	0 / 50 / 500	2	✗
BATS	英文	0 / 199 / 1,799	2	✗
E-KAR	中文	1,155 /165 / 335	2(64.5%),3(35.5%)	✓
	英文	870 / 119 / 262	2(60.5%),3(39.5%)	✓

资料来源：Jiangjie Chen et al. “E-KAR : A Benchmark for Rationalizing Natural Language Analogical Reasoning” 2022，华泰研究

Douban Conversation Corpus 数据集：于 2017 年由北京航空航天大学、南开大学和微软研究院联合发布。豆瓣会话语料库包括一个训练数据集、一个开发集和一个基于检索的聊天机器人的测试集，测试数据包含 1000 个对话上下文。

图表54：豆瓣会话语料库统计

数据集	训练	有效	测试
#上下文响应对	1M	50K	10K
#每个上下文的候选人	2	2	10
#每个上下文的积极候选人	1	1	1.18
最小值#每个上下文的转数	3	3	3
最大值#每个上下文的转数	98	91	45
平均值#每个上下文的转数	6.69	6.75	6.45
平均值#每句话的字数	18.56	18.5	20.74

资料来源：Yu Wu et al. “Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots” 2017，华泰研究

ODSQA 数据集：于 2018 年由台湾大学发布。ODSQA 数据集是用于中文问答的口语数据集。它包含来自 20 位不同演讲者的三千多个问题。

图表55：ODSQA、DRCD-TTS、DRCD-backtrans 的数据统计

子集	问答对	时长	M-spkrs	F-Spkrs	WER(%)	WER-Q(%)	文档的平均长度	问题的平均长度
(1) ODSQA-test	1,465	25.28	7	13	19.11	18.57	428.32	22.08
(2) DRCD-TTS	16,746	-	-	-	33.63	-	332.80	20.53
(3) DRCD-backtrans	15,238	-	-	-	45.64	-	439.55	20.75

资料来源：Chia-Hsuan Lee et al. “ODSQA: OPEN-DOMAIN SPOKEN QUESTION ANSWERING DATASET” 2018，华泰研究

MATINF 数据集：于 2020 年由武汉大学和密歇根大学联合发布。MATINF 是一个联合标注的大规模数据集，用于中文母婴护理领域的分类、问答和总结。数据集中的一个条目包括四个字段：问题、描述、类别和答案。从中国大型母婴护理 QA 网站收集了近 200 万对问答对，其中包含细粒度的人工标记类，数据清洗后，用剩余的 107 万个条目构建。

图表56： MATINF 中问题、描述和答案的平均字符数和单词数

单词	问题	描述	答案	平均长度
#字符	14.72	64.17	66.91	256
#单词	9.03	41.70	42.32	-

资料来源：Canwen Xu et al. “MATINF: A Jointly Labeled Large-Scale Dataset for Classification,Question Answering and Summarization” 2020，华泰研究

中国开源数据集#2：多模态模型数据集

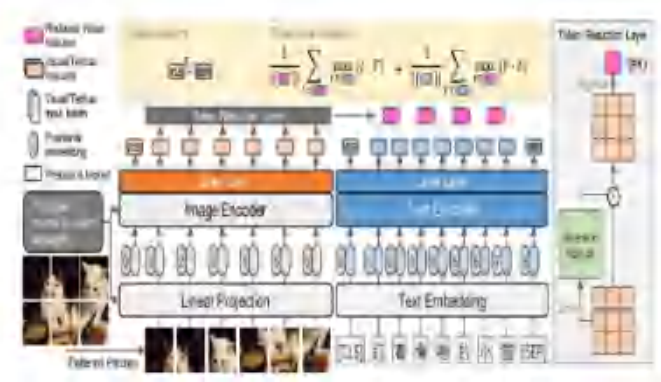
WuDaoMM 数据集：于 2022 年由清华大学和北京智源人工智能研究院联合发布。WuDaoMM 是北京智源人工智能研究院 WuDaoCorpora 开源数据集的一部分。WuDaoMM 是图像和文本的多模态预训练数据，完整的数据集包含 6.5 亿对图像和文本，包含几千万对的强相关数据和 6 亿对弱相关数据，包含 19 大类，分别是：能源、表情、产业、医疗、景观、动物、新闻、花卉、教育、艺术、人物、科学、海洋、树木、汽车、社会、科技、体育等。

MUGE 数据集：于 2021 年由清华大学和阿里巴巴联合发布，包括图像描述、图像文本检索和基于文本的图像生成 3 种多模态理解和生成任务。

图表57： MUGE 数据集——多模态数据示例

资料来源：Junyang Lin et al. “M6: A Chinese Multimodal Pretrainer” 2021，华泰研究

图表58： WuDaoMM 数据集——强相关性图像-文本对示例



资料来源：Sha Yuan et al. “WuDaoMM: A large-scale Multi-Modal Dataset for Pre-training models” 2022，华泰研究

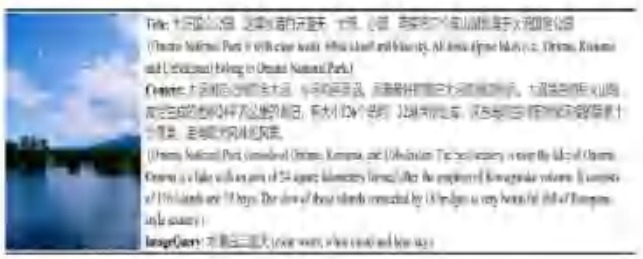
Noah-Wukong 数据集：于 2022 年由华为诺亚方舟实验室和中山大学联合发布。诺亚悟空数据集是一个大规模的多模态中文数据集，包含 100 万对图文对，数据集中的图像根据大小和宽高比进行过滤，数据集中的文本根据其语言，长度和频率进行过滤。隐私和敏感词也被考虑在内。

Zero 数据集：于 2022 年由北京航空航天大学、清华大学、奇虎 360 人工智能研究所联合发布。Zero 是一种大规模的中文跨模态基准测试，其中包含目前最大的公共预训练数据集 ZERO-Corpus 和五个用于下游任务的人工注释微调数据集。ZERO-Corpus 包含 2.5 亿张图片和 7.5 亿篇文字描述，另外五个微调数据集中的两个也是目前中国跨模式下游任务中最大的数据集。

图表59：Noah-Wukong 数据集——模型概述

资料来源：Jiaxi Gu et al. “Wukong: A 100 Million Large-scale Chinese Cross-modal Pre-training Benchmark” 2022，华泰研究

图表60：Zero 数据集——示例



资料来源：Chunyu Xie et al. “ZERO and R2D2: A Large-scale Chinese Cross-modal Benchmark and a Vision-Language Framework” 2022，华泰研究

COCO-CN 数据集：于 2018 年由中国人民大学发布。COCO-CN 是一个双语图像描述数据集，通过手动编写的中文句子和标签丰富了 MS-COCO。新数据集可用于多种任务，包括图像标记、字幕和检索，所有这些都在跨语言环境中完成。COCO-CN 拥有 20,342 张图片，27,218 个中文句子和 70,993 个标签，为跨语言图像标注、字幕和检索提供了一个统一平台。

Flickr8k-CN & Flickr30k-CN 数据集：于 2017 年由浙江大学和中国人民大学联合发布。Flickr8k-cn 是公共数据集，每个测试图像与 5 个中文句子相关联，这些句子是通过手动翻译 Flickr8k 中对应的 5 个英文句子获得的。Flickr30k-cn 是 Flickr30k 的双语版本，通过其训练/有效集的英译汉机器翻译和测试集的人工翻译获得。

图表61：COCO-CN 数据集——示例

资料来源：Xirong Li et al. “COCO-CN for Cross-Lingual Image Tagging, Captioning and Retrieval” 2019，华泰研究

图表62：Flickr30k-CN 数据集——跨语言图像字幕示例

资料来源：Weiyu Lan et al. “Fluency-Guided Cross-Lingual Image Captioning” 2017，华泰研究

Product1M 数据集：于 2021 年由北京交通大学、阿里巴巴和中山大学联合发布。Product1M 是用于实际实例级检索的最大的多模式化妆品数据集之一，包含超过 100 万个图像对并且由两种样品类型组成，即单产品和多产品样品，其中包括各种化妆品品牌。

AI Challenger 图像中文描述数据集：数据来自 2017 AI Challenger，数据集对给定的每一张图片有五句话的中文描述。数据集包含 30 万张图片，150 万句中文描述。数据集包含人类关键点检测(HKD)、大规模属性数据集(LAD)和图像中文字幕(ICC)三个子数据集。

图表63： Product1M 数据集——多模态实例级检索

资料来源：Xunlin Zhan et al. “Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal Pretraining” 2021，华泰研究

图表64： AI Challenger 数据集——示例

资料来源：Jiahong Wu et al. “AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding” 2017，华泰研究

国内数据要素市场建设逐步完善，助力优质数据集生产流通

数字中国建设规划明晰，数据要素为发展框架中关键环节之一。2023 年 2 月 27 日，中共中央、国务院印发《数字中国建设整体布局规划》，文件中明确数字中国建设按照“2522”的整体框架进行布局，即夯实数字基础设施和数据资源体系“两大基础”，推进数字技术与经济、政治、文化、社会、生态文明建设“五位一体”深度融合，强化数字技术创新体系和数字安全屏障“两大能力”，优化数字化发展国内国际“两个环境”。《规划》提出要释放商业数据价值潜能，加快建立数据产权制度，开展数据资产计价研究，建立数据要素按价值贡献参与分配机制。构建国家数据管理体制机制，健全各级数据统筹管理机构，推动公共数据汇聚利用。

图表65： 数据要素是数字中国发展框架中的重要环节之一

资料来源：中国信息通信研究院，华泰研究

我国重视数据要素发展，组建国家数据局，数据要素政策频出。2023 年 3 月 10 日，党的二十届二中全会通过了《党和国家机构改革方案》，方案提出组建国家数据局。国家数据局负责协调推进数据基础制度建设，推进数字基础设施布局建设，统筹数据资源整合共享和开发利用，统筹推进数字中国、数字经济、数字社会规划和建设等，由国家发展和改革委员会管理。这对于充分激活数据要素潜能、发挥数字经济对经济社会的基础性作用而言是场及时雨。

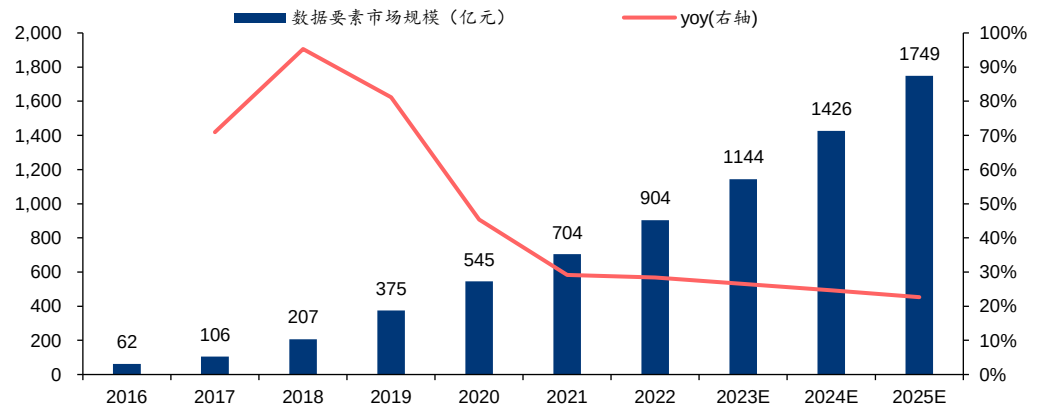
图表66：我国数据要素相关政策

发布日期	发布主体	政策名称	主要内容
2022 年 12 月	国务院	《关于构建数据基础制度更好发挥数据要素作用的意见》	确立了数据要素市场的四大原则体系："数据产权、流通交易、收益分配、安全治理"，标志着我国数据要素基础制度顶层设计开始启动。
2022 年 6 月	国务院	《关于加强数字政府建设的指导意见》	在构建开放共享的数据资源体系方面，创新数据管理机制，深化数据高效共享，促进数据有序开发利用，充分释放数据要素价值。
2022 年 4 月	国务院	《中共中央国务院关于加快建设全国统一大市场的意见》	加快培育数据要素市场，建立健全数据安全、权利保护、跨境传输管理、交易流通、开放共享、安全认证等基础制度和标准规范，深入开展数据资源调查，推动数据资源开发利用。
2021 年 12 月	网信办	《"十四五"国家信息化规划》	提出建立数据要素资源体系，以数据治理为突破提升数据质量，以数据开发利用为抓手激活数据要素，以 立法规范为重点保障数据安全，加快完善与我国发展实际相吻合的数据要素资源体系，释放数据要素价值。
2021 年 12 月	国务院	《"十四五"数字经济发展规划》	要求规范数据交易管理，培育规范的数据交易平台和市场主体，建立健全数据资产评估、登记结算、交易撮合、争议仲裁等市场运营体系，提升数据交易效率。
2021 年 11 月	工信部	《"十四五"大数据产业发展规划》	建立数据价值体系，制定数据要素价值评估指南，开展评估试点；健全要素市场规则，发展数据资产评估、 交易撮合等市场运营体系；提升要素配置作用，加快数据要素化
2021 年 1 月	国务院	《建设高标准市场体系行动方案》	建立数据资源产权、交易流通、跨境传输和安全等基础制度和标准规范
2020 年 4 月	国务院	《关于构建更加完善的要素市场化配置体制机制的意见》	首次提出土地、劳动力、资本、技术、数据五个要素领域的改革方向，将数据列为生产要素。

资料来源：各政府官网，华泰研究

随着数据要素快速融入生产、分配、流通、消费和社会服务管理等各个环节，我们预计未来几年我国的数据要素市场将会蓬勃发展，并实现快速增长。根据国家工信安全发展研究中心数据，2021 年我国数据要素行业市场规模为 815 亿元，预计到 2025 年将达到 1749 亿元左右，2020-2025 年 CAGR 为 26.26%。

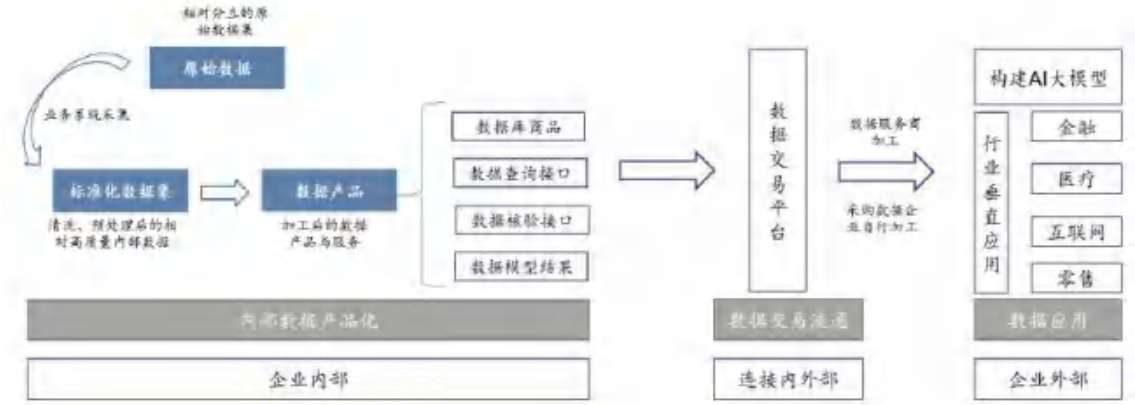
图表67：我国数据要素市场规模及预测



资料来源：国家工业信息安全发展研究中心，中国信息通信研究院，华泰研究

数据要素产业包括数据的内部产生，流通交易，数据加工，行业应用等流程。从企业内原始数据到企业外可以应用的数据产品，需要经历内部数据产品化，数据交易流通，外部数据加工等过程。企业通过在内部将数据清洗，预处理，加工等将数据变为数据产品，并将数据产品放在数据交易平台上交易。在应用端，采购数据企业可以采购交易平台中数据，后自行加工使用于垂直行业应用领域。

图表68：数据要素流通产业链



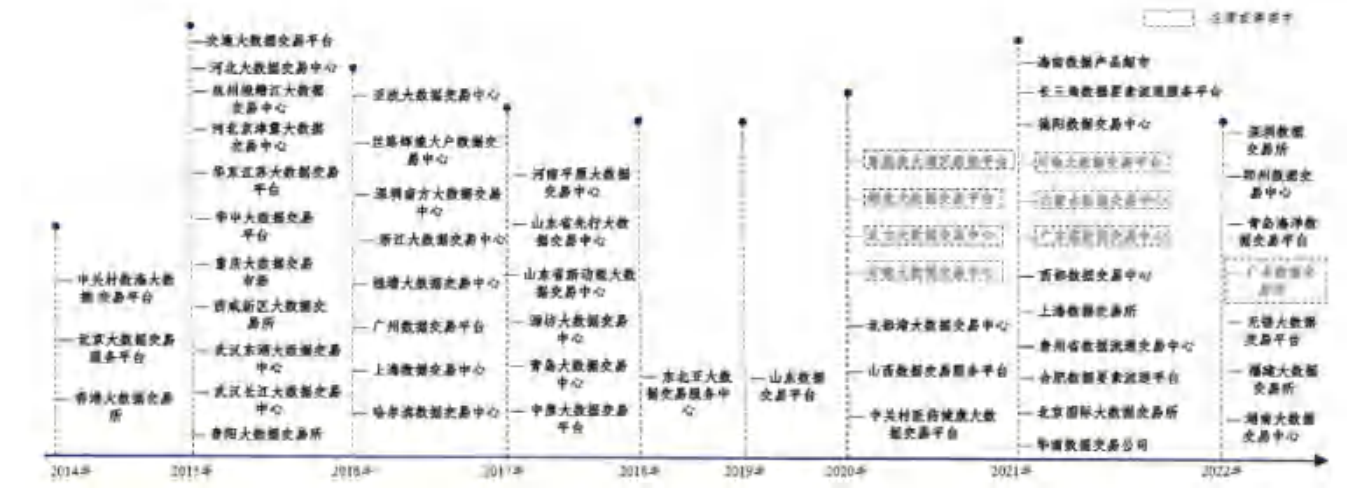
资料来源：中国信息通信研究院，华泰研究

数据交易环节：数据交易所发展进入新阶段，缓解中文数据集数量不足问题

我国数据交易所发展进入新阶段，构筑数据集发展坚实基座。我国自 2014 年开始探索建立类似证券交易所形式的数据交易机构，截至 2022 年 11 月，各地先后成立 48 家，仍有 8 家正在筹备建设中。**第一阶段：**2015 年，党的十八届五中全会提出“实施国家大数据战略，推进数据资源开放共享”，以贵阳大数据交易所为代表的数据交易机构涌现。早期建设的数据交易机构大都没有找到成功的商业模式，多数机构已停止运营或转变经营方向。

第二阶段：2021 年以来北京、上海等大数据交易所成立。近两年来，随着党中央国务院多项重要政策出台，各地新建一批数据交易机构，试图消除供需双方的信息差，推动形成合理的市场化价格机制和可复制的交易制度和规则。我们预计新一阶段，国家将大力发展数据交易所，并将带动上下游产业发展，这将为国内大模型数据集发展注入新动力。

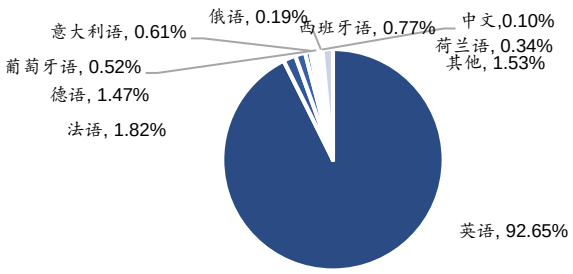
图表69：国内大数据交易所建设历程



资料来源：中国信息通信研究院，华泰研究

由于开源较少，数据无法在市场上自由流通，目前中文优质数据集相对稀缺。对于大模型训练，能否未来得到优质的大模型，与投入的数据质量非常关键，包括通用参数、文本语言、图像、视频音频等等，投入数据的质量高低，会直接影响模型最终生成的内容。以 GPT-3 为例，其训练集中中文数据仅占 0.1%，这一方面因为 GPT 为美国开发，以英文语料为主；另一方面也反映了目前可获得数据集中，中文内容仍然相对稀缺。国内大模型训练基于中文原生内容仍是第一选择，翻译数据训练效果劣于中文原生内容。

图表70： GPT3 训练中各国语言占比



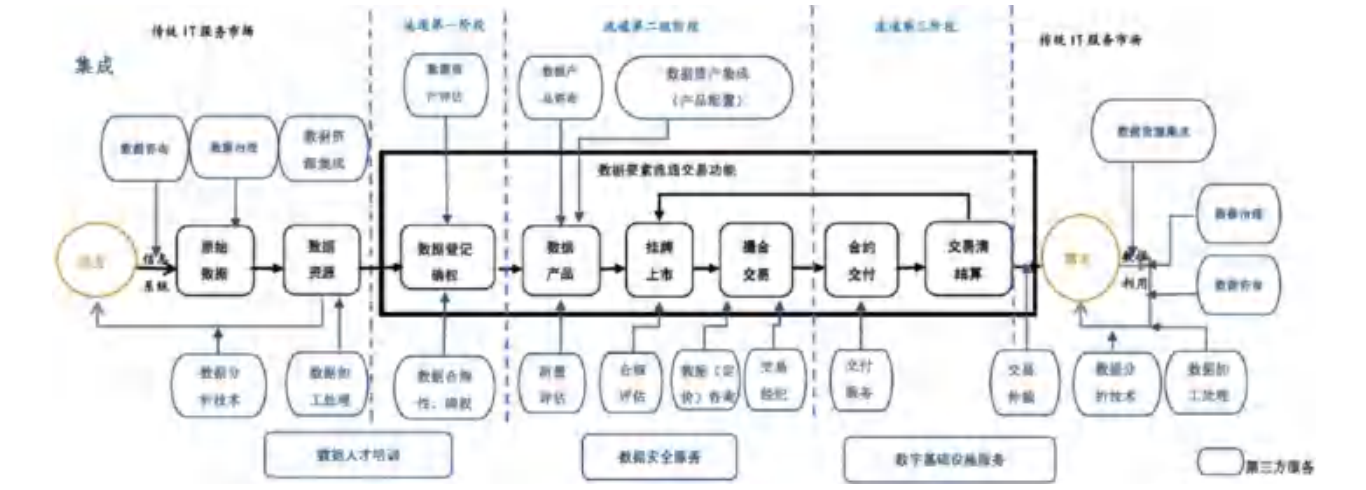
资料来源：中国网信网，华泰研究

未来随着各地积极推动数据交易所建设，数据有望在各行业、各企业之间实现自由流通，缓解国内优质数据集不足问题。据上海数据交易所总经理汤奇峰，上海数据交易所 2023 年场内交易额有望突破 10 亿元。据深圳数据交易所预计，未来 2-3 年，其数据交易规模超过 100 亿元，设立合规工作站 100 家以上，培育、引入数据服务企业 50 家以上。我们认为数据交易所发展将使得国内数据流通更顺畅，未来中小型模型训练企业可以直接从交易所购买各行业的数据产品，这将极大地提升大模型训练数据的可获得性，缓解国内优质数据集不足的问题。

数据加工环节：数据服务产业加速发展，助力中文数据集质量提升

数据服务商链接数据要素产业链上下游，助力形成优质数据集。上海数据交易所在全国率先提出“数商”概念，以数据交易为中心激活数据要素上下游产业链。并提出了 15 个的“数商”类别。传统大数据服务商：数据咨询服务商、数据治理服务商、数据资源集成商、数据加工服务商、数据分析技术服务商等。数据交易相关服务商：数据合规评估服务商、数据质量评估商、数据资产评估服务商、数据经纪服务商、数据交付服务商、数据交易仲裁服务商。我们预计数据服务商的参与将会进一步激活数据交易流通市场，提供更多样化的数据产品，将促进我国大模型数据集的发展。

图表71： 数据服务商在数据要素市场中的角色



资料来源：上海数据交易所，华泰研究

数据服务商参与有望提升国内大模型训练数据质量。据 Dimensional Research 全球调研报告，72%的受访者认为至少使用超过 10 万条训练数据进行模型训练，才能保证模型有效性和可靠性，96%的受访者在训练模型的过程中遇到训练数据质量不佳、数量不足、数据标注人员不足等难题。我们认为随着国内数据服务产业蓬勃发展，数据服务商未来将在数据加工处理，数据基础设施建设，数据资源集成，提供数据分析服务等方面协助企业构建高质量数据集，这将进一步提升我国大模型训练的数据质量，从而促进各行业大模型的发展效率提升。

图表72：国内各类型数据服务商企业统计样本数及占比

数商类型	企业数量（万）	占比	数商类型	企业数量（万）	占比
数据咨询服务商	66.6	34.7%	数据资产评估服务商	6.6	3.4%
数据资源集成商	41.1	21.4%	数据合规评估服务商	2.2	1.1%
数据分析技术服务商	27.5	14.3%	数据质量评估商	0.74	0.4%
数据基础设施提供商	13.8	7.2%	数据人才培训服务商	0.47	0.2%
数据加工处理服务商	12.2	6.3%	数据交易经纪服务商	0.46	0.2%
数据安全服务商	10.5	5.5%	数据交易仲裁服务商	0.13	0.1%
数据产品供应商	9.8	5.1%	数据交付、数据治理商	不足百家	0.0%

资料来源：上海数据交易所，华泰研究

AI 时代数据的监管与隐私保护问题

人工智能引发数据隐私关注，需平衡技术发展与隐私保护。随着人工智能技术的不断发展和应用，大量的个人数据被采集、存储和处理，由此引发了人们对于 AI 时代数据的隐私保护的关注和讨论。数据隐私问题的严重性不言而喻，如何平衡人工智能技术的应用与数据隐私保护之间的关系、如何实现人机共存的良好发展是现在亟需解决的问题。

图表73： 大模型数据隐私问题实例

时间	国家	实例
2023.3.21	-	Reddit 上有用户发布截图称遇到了一个 BUG，其 ChatGPT 聊天历史栏出现了不是自己的聊天记录标题。
2023.3.30	韩国	3 月 30 日，据韩国媒体《Economist》报道，三星内部发生三起涉及 ChatGPT 误用与滥用案例，包括两起“设备信息泄露”和一起“会议内容泄露”，相关的半导体设备测量资料、产品良率等内容或已被存入 ChatGPT 学习资料库中。进一步报道表明，三星导体暨装置解决方案（Device Solutions）部门的 A 职员在执行半导体设备测量资料库（DB）下载程序的源代码时出现了错误，便复制出有问题的原始代码到 ChatGPT，并询问了解决方法。另外，三星 DS 部门的 B 职员把为了把握产量、不良设备而编写的源代码输入到 ChatGPT，并要求其优化。另外，三星 DS 部门的 C 职员则将手机录制的会议内容转换为文件后输入到 ChatGPT，要求其制作会议纪录。
2023.3.31	意大利	意大利个人数据保护局宣布，基于欧盟的《通用数据保护条例》（GDPR）法律，暂时禁止 ChatGPT 在其境内运行。此外，意大利数据保护局在上月底对 ChatGPT 涉嫌违反隐私规则展开调查。该机构认为，3 月 20 日 ChatGPT 平台出现了用户对话数据和付款服务支付信息丢失情况。此外平台没有就收集处理用户信息予以告知，缺乏大量收集和存储个人信息的法律依据。Open AI 必须在 20 天内与监管机构进行沟通采取有效保护措施，不然将面临最高 2000 万欧元或全球年营业额的 4%罚款。
2023.4.3	德国	德国联邦数据保护专员 Ulrich Kelber 对德国商报表示，原则上，类似 ChatGPT 这样的 AI 软件在德国被禁用是有可能的。因为，OpenAI 是一家国外公司，是在德国数据保护机构管控范围内。
2023.4.4	意大利	意大利隐私监管机构发布声明，称将于当地时间 4 月 5 日与 OpenAI 代表会面，讨论在该国暂时禁止使用 ChatGPT 的问题。OpenAI 表示，该公司愿意同意大利当局合作，以确保遵守隐私法规。
2023.4.4	法国	法新社消息称加拿大隐私专员办公室（OPC）4 日宣布开始调查 ChatGPT 背后的公司 OpenAI，涉及“指控 OpenAI 未经同意收集、使用和披露个人信息”的投诉。

资料来源：法新社，德国商报，经济学人，华泰研究

方法#1：法律法规技术手段——数据隐私需法律约束，全球出台相关法规加强个人数据保护。数据隐私问题需要法律约束，以确保个人数据得到妥善保护，避免数据滥用和泄露。全球各地区纷纷出台相关法律法规，例如中国的《中华人民共和国个人信息保护法》、欧盟的《通用数据保护条例》、美国的《美国隐私法》等，以加强对个人数据的保护。

图表74：各地区数据隐私相关法律

英美欧数据隐私相关法律

美国	英国	欧盟
《美国隐私法》 《美国计算机欺诈和滥用法》 《美国金融消费者保护法》	《英国数据保护法》 《英国通信数据法》 《英国通用数据保护法规指南》 《英国数据保护、隐私和电子通信(修订等)(欧盟退出)条例》 《英国个人数据保护法》	《关于在个人数据处理中对个人的保护以及此类数据自由流动的指令》 《通用数据保护条例》 《电子隐私条例》

中国数据隐私相关法律

相关法律	生效时间	相关内容
《中华人民共和国网络安全法》	2017.06	规定网络运营者应当对其收集的用户信息严格保密，并建立健全用户信息保护制度。网络运营者收集、使用个人信息，应当遵循合法、正当、必要的原则，公开收集、使用规则，明示收集、使用信息的目的、方式和范围，并经被收集者同意。
《中华人民共和国数据安全法》	2021.09	规定了国家机关为履行法定职责的需要收集、使用数据，应当在其履行法定职责的范围内依照法律、行政法规规定的条件和程序进行；对在履行职责中知悉的个人隐私、个人信息、商业秘密、保密商务信息等数据应当依法予以保密，不得泄露或者非法向他人提供。
《中华人民共和国个人信息保护法》	2021.11	规定了自然人的个人信息受法律保护，任何组织、个人不得侵害自然人的个人信息权益。在中华人民共和国境内处理自然人个人信息的活动，适用本法。国家保护个人、组织与数据有关的权益，鼓励数据依法合理有效利用，保障数据依法有序自由流动，促进以数据为关键要素的数字经济发展。

资料来源：各政府官网，华泰研究

方法#2：技术手段——隐私保护计算具体涵盖了安全多方计算、联邦学习、同态加密、差分隐私和机密计算等技术。隐私保护计算是一套包含人工智能、密码学、数据科学等众多领域交叉融合的跨学科技术体系。它能够在不泄露原始数据的前提下，对数据进行加工、分析处理、分析验证，其重点提供了数据计算过程和数据计算结果的隐私安全保护能力。

图表75：隐私保护计算的五大关键技术

安全多方计算	联邦学习	同态加密	差分隐私	机密计算
安全多方计算（Secure Multi-Party Computation，SMPC）旨在解决“一组相互独立且互不信任的参与方各自持有秘密数据，协同计算一个既定函数”的问题。安全多方计算保证了各参与方在获得正确计算结果的同时，无法获得计算结果之外的任何信息。	联邦学习（Federated Learning，FL），可被理解为是由两个或两个以上数据方共同参与，在保证数据方各自原始数据不出其定义的安全控制范围的前提下，协作构建并使用机器学习模型的技术架构。通常情况下，联邦学习需与其它隐私保护计算技术联合使用，才可在计算过程中实现数据保护。	同态加密（Homomorphic Encryption，HE），是一种允许在加密之后的密文上直接进行计算，且计算结果解密后与基于明文的计算结果一致的加密算法，可在不解密以实现数据机密性保护的同时完成计算。根据支持密文运算的程度，同态加密方案可以分为部分同态加密方案和全同态加密方案两类。	差分隐私（Differential Privacy，DP），在保留统计学特征的前提下，去除个体特征以保护用户隐私。差分隐私具有两个重要的优点：一是提出与背景知识无关的隐私保护模型，实现攻击者背景知识最大化的假设；二是为隐私保护水平提供严格的定义和量化评估方法。	机密计算（Confidential Computing，CC），是指通过在基于硬件的可信执行环境中执行计算来保护数据应用中的隐私安全的技术之一。其基本原理是将需要保护的数据和代码存储在可信执行环境中，对这些数据和代码的任何访问都必须经过基于硬件的访问控制，防止他们在使用中未经授权被访问或修改，从而提高机构管理敏感数据的安全水平。

资料来源：中国信通院，华泰研究

数据产业链投资机会

我们认为数据产业链分为数据生产、数据处理、数据使用三大环节。数据使用环节的参与者包括训练、微调大模型的企业，本文不作展开。以下我们对数据生产、数据处理环节进行讨论。

数据生产环节

数据生产环节是数据产业链的上游环节，是数据的源头。环节内的企业或从业务运营中直接产生数据，或作为平台方聚合数据。按照数据的通用程度，我们认为这一环节的公司可以分为通用类型数据及垂直行业数据 2 类。

1） **通用类型数据：**如前文所言，我们认为 AI 大模型需要高质量、大规模、具有多样性的数据。对标海外主要数据集，通用类型数据来自维基百科、书籍期刊、高质量论坛，因此国内的数据或来自文本数据领域的百度百科、中文在线、中国科传、知乎，以及图像视觉领域的视觉中国等公司。

a) 截至 2022 年 6 月，视觉中国拥有超过 2/3 的高水准独家或自有内容，目前提供 4 亿张图片、3,000 万条视频和 35 万首音乐等可销售的各类素材，是全球最大的同类数字版权内容平台之一。

b) 中国科传从事图书出版业务、期刊业务、出版物进出口业务。截至 2022 年底，公司年出版新书超过 3000 种，已累计出版图书超过 5 万种，是国内学科分布最全、出版规模最大的综合性科技出版机构。截至 2022 年底，中国科传出版期刊 554 种，其中中文期刊 254 种，英文期刊 276 种，中英文期刊 5 种，法文期刊 19 种。共有 101 种期刊被 SCI 收录，其中 36 种期刊处于 Q1 区，4 种期刊在国际同学科期刊中排名第一，16 种期刊居国际同学科期刊排名前 10%。

c) 截至 2022 年 6 月，中文在线累积数字内容资源超 510 万种，网络原创驻站作者 440 余万名。

2） **行业数据：**我们认为垂直行业的高价值量数据对于 AI 大模型，尤其是行业大模型的训练和落地至关重要。处于数字化程度领先的行业中的龙头公司在行业数据积累上具有优势，例如：1）计算机视觉领域的海康威视、大华股份；2）城市治理、ToB 行业应用领域的中国电信、中国移动、中国联通等；3）金融领域的同花顺、东方财富等；4）自动驾驶领域的特斯拉、蔚小理、经纬恒润、德赛西威等。

大模型时代数据价值凸显，国内外数据收费为大势所趋，收费方式尚在摸索中。2023 年 4 月 18 日，美国知名论坛 Reddit 宣布计划向通过其 API 使用数据的公司收费。Reddit 尚未公布具体的收费标准，但表示会分为不同的等级，根据使用者的规模和需求来区分。Reddit 是大模型训练的优质语料库，OpenAI 的 GPT-3 训练使用了来自 Reddit 的数据，Meta 旗下的 Facebook AI Research 与华盛顿大学也联合开源了来自 Reddit 数据的 OpenWebText 数据集。对于通用类型数据和行业数据，我们认为其潜在的变现方式可能存在差异：

1） **通用类型数据：**我们认为通用类型数据所有者可能采用开发自有模型/应用、售卖数据 2 种变现方式。例如，知乎联合清华系 AI 公司面壁智能发布中文大模型“知海图 AI”。中文在线则基于自有数据开发了 AI 辅助文字创作工具，并计划售卖数据：根据中文在线 4 月 19 日回复深交所关注函内容，其收费方式为按照采集数据包的大小及数据类别进行基础包加增量包的收费，目前尚未签署具体合作协议。

2） **行业数据：**我们认为数据是垂直行业企业的护城河之一，结合具体场景和用户充分挖掘数据能更好地赋能业务。因此垂直行业企业或更偏好基于基础模型，使用自有数据来训练自有模型，并且可能会尽量规避售卖数据。建议关注具有丰富行业数据积累的龙头公司。

数据处理环节

根据 IDC 在 2020 年的数据，百度智能云和海天瑞声是我国 AI 基础数据服务市场中份额最大的两家公司。Appen、Telus international 则是海外数据服务的主要上市公司。其中，百度智能云数据众包是平台型 AI 数据服务提供者，服务涵盖方案设计、数据采集与数据标注全流程，并与政府共建数据标注基地；海天瑞声数据服务涵盖从方案设计到采集、标注直至交付的全流程；慧听科技包括语言语音、多媒体两大类几十余种数据服务；标贝科技提供语音合成整体解决方案及数据服务；Appen 拥有 MatrixGo 数据标注平台；Scale AI 通过帮助机器学习团队生成高质量的地面数据来加速 AI 应用程序的开发;V7 的图像标记平台可应用于医疗保健、生命科学、制造业、自动驾驶、农业科技等领域；Telus international 服务包括数字化战略、创新、咨询和设计、数字化转型和 IT 生命周期服务、数据注释和智能自动化；Lion bridge 是 AI 语言服务提供商。

图表76： 国内外数据处理相关公司

公司名称	公司代码	所属国家	公司简介
百度智能云	9888 HK/BIDU US	中国	百度智能云数据众包作为平台型 AI 数据服务提供者，自 2011 年建立数据采标团队，能力全面，资源丰富。百度智能云数据众包与政府共建数据标注基地，拥有百位数据项目方案专家，2 千名百度山西基地全职标注人力，2 万名签约外场专职标注人员，3 万名百度众包在线标注用户，实现百万级数据标注处理能力
海天瑞声	688787 CH	中国	**2022 年收入 2.6 亿元。**公司是我国领先的 AI 训练数据专业提供商，自 2005 年成立以来，始终致力于为 AI 产业链上的各类机构提供 AI 算法模型开发训练所需的专业数据集。公司所提供的训练数据覆盖智能语音(语音识别、语音合成等)、计算机视觉、自然语言等多个 AI 核心领域，全面服务于人机交互、智能驾驶、智慧城市等多种创新应用场景。
慧听科技	未上市	中国	一家专业的数据服务提供商。团队核心成员拥有十余年的数据制作经验，负责完成过语音识别、语音合成、语音评测、语言文本类、多媒体类等多领域数据制作，并参与过语音合成、语音识别、输入法系统的研发。目前提供语言语音、多媒体两大类几十余种数据服务。
标贝科技	未上市	中国	一家专注于智能语音交互和 AI 数据服务的人工智能公司，基于 AI+SaaS 开放平台，为客户提供 AI 数据服务、技术能力、智能语音交互方案赋能服务，包括通用场景的语音合成和语音识别，以及 TTS 音色定制，声音复刻，情感合成和声音转换在内的语音技术产品；AI 数据业务涵盖语音合成、语音识别、图像视觉、NLP 等采标服务和平台化自研工具能力。
Appen	APX AU	澳大利亚	**2022 年收入 3.9 亿美元。**公司是全球领先的图像、文本、语音、音频、视频等 AI 训练数据服务提供商，拥有业内最先进的人工智能辅助数据标注平台、一体化的 AI 数据及资源管理平台及全球 100 多万名技能娴熟的众包资源。
Telus international	TIXT CN	加拿大	**2022 年总收入 24.7 亿美元，其中 AI 数据解决方案占比 13%（约 3.2 亿美元）**。公司业务包括数字化战略、创新、咨询和设计、数字化转型和 IT 生命周期服务、数据注释和智能自动化等等。
Lion bridge	未上市	美国	一家覆盖 800+语言、每年翻译 36 亿文字的 AI 语言服务提供商。
Scale AI	未上市	美国	公司通过帮助机器学习团队生成高质量的地面数据来加速 AI 应用程序的开发。其先进的 LiDAR，图像，视频和 NLP 注释 API 允许 OpenAI，Lyft，Pinterest 和 Airbnb 等公司的机器学习团队专注于构建差异化模型和标签数据。
V7	未上市	英国	一家机器学习技术研发商，基于人工智能技术，推出了图像标记平台用于为电脑视觉项目创建培训数据，可应用于医疗保健、生命科学、制造业、自动驾驶、农业科技等领域。

注：9888 HK/BIDU US 为百度代码
资料来源：公司官网，公司公众号，公司公开投资者展示材料，IT 桔子，彭博，华泰研究

风险提示

1） AI 技术落地不及预期。虽然 AI 技术加速发展，但由于成本、落地效果等限制，相关技术落地节奏可能不及我们预期.

2） 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。

图表77：全文提及公司列表

公司	代码	公司	代码
百度	9888 HK/BIDU US	理想汽车	2015 HK/LI US
阿里巴巴	9988 HK/BABA US	同花顺	300033 CH
腾讯	700 HK	东方财富	300059 CH
商汤-W	20 HK	海天瑞声	688787 CH
中文在线	300364 CH	谷歌	GOOG US
中国科传	601858 CH	微软	MSFT US
视觉中国	000681 CH	Meta	META US
海康威视	002415 CH	Appen	APX AU
大华股份	002236 CH	Telus international	TIXT CN
中国移动	600941 CH/941 HK	标贝科技	未上市
中国联通	600050 CH/762 HK	慧听科技	未上市
中国电信	601728 CH/728 HK	Lion bridge	未上市
经纬恒润	688325 CH	Scale AI	未上市
德赛西威	002920 CH	V7	未上市
蔚来	9866 HK/NIO US	Stability AI	未上市
小鹏汽车	9868 HK/XPEV US	OpenAI	未上市

资料来源：彭博，华泰研究

免责声明

分析师声明

本人，黄乐平、余熠，兹证明本报告所表达的观点准确地反映了分析师对标的证券或发行人的个人意见；彼以往、现在或未来并无就其研究报告所提供的具体建议或所表达的意见直接或间接收取任何报酬。

一般声明及披露

本报告由华泰证券股份有限公司（已具备中国证监会批准的证券投资咨询业务资格，以下简称“本公司”）制作。本报告所载资料是仅供接收人的严格保密资料。本报告仅供本公司及其客户和其关联机构使用。本公司不因接收人收到本报告而视其为客户。

本报告基于本公司认为可靠的、已公开的信息编制，但本公司及其关联机构(以下统称为“华泰”)对该等信息的准确性及完整性不作任何保证。

本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，华泰可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。华泰不保证本报告所含信息保持在最新状态。华泰对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司不是 FINRA 的注册会员，其研究分析师亦没有注册为 FINRA 的研究分析师/不具有 FINRA 分析师的注册资格。

华泰力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽。该等观点、建议并未考虑到个别投资者的具体投资目的、财务状况以及特定需求，在任何时候均不构成对客户私人投资建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。对依据或者使用本报告所造成的一切后果，华泰及作者均不承担任何法律责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。华泰不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。

华泰及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，华泰可能会持有报告中提到的公司所发行的证券头寸并进行交易，为该公司提供投资银行、财务顾问或者金融产品等相关服务或向该公司招揽业务。

华泰的销售人员、交易人员或其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。华泰没有将此意见及建议向报告所有接收者进行更新的义务。华泰的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。投资者应当考虑到华泰及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。有关该方面的具体披露请参照本报告尾部。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布的机构或人员，也并非意图发送、发布给因可得到、使用本报告的行为而使华泰违反或受制于当地法律或监管规则的机构或人员。

中国香港

本报告由华泰证券股份有限公司制作,在香港由华泰金融控股（香港）有限公司向符合《证券及期货条例》及其附属法律规定的机构投资者和专业投资者的客户进行分发。华泰金融控股（香港）有限公司受香港证券及期货事务监察委员会监管，是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。在香港获得本报告的人员若有任何有关本报告的问题,请与华泰金融控股（香港）有限公司联系。

香港-重要监管披露

- 华泰金融控股（香港）有限公司的雇员或其关联人士没有担任本报告中提及的公司或发行人的高级人员。
- 有关重要的披露信息，请参华泰金融控股（香港）有限公司的网页 https://www.htsc.com.hk/stock_disclosure

其他信息请参见下方 **“美国-重要监管披露”**。

美国

在美国本报告由华泰证券（美国）有限公司向符合美国监管规定的机构投资者进行发表与分发。华泰证券（美国）有限公司是美国注册经纪商和美国金融业监管局（FINRA）的注册会员。对于其在美国分发的研究报告，华泰证券（美国）有限公司根据《1934 年证券交易法》（修订版）第 15a-6 条规定以及美国证券交易委员会人员解释，对本研究报告内容负责。华泰证券（美国）有限公司联营公司的分析师不具有美国金融监管（FINRA）分析师的注册资格，可能不属于华泰证券（美国）有限公司的关联人员，因此可能不受 FINRA 关于分析师与标的公司沟通、公开露面和所持交易证券的限制。华泰证券（美国）有限公司是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。任何直接从华泰证券（美国）有限公司收到此报告并希望就本报告所述任何证券进行交易的人士，应通过华泰证券（美国）有限公司进行交易。

美国-重要监管披露

- 分析师黄乐平、余熠本人及相关人士并不担任本报告所提及的标的证券或发行人的高级人员、董事或顾问。分析师及相关人士与本报告所提及的标的证券或发行人并无任何相关财务利益。本披露中所提及的“相关人士”包括 FINRA 定义下分析师的家庭成员。分析师根据华泰证券的整体收入和盈利能力获得薪酬，包括源自公司投资银行业务的收入。
- 中国移动（600941 CH）：华泰证券股份有限公司、其子公司和/或其联营公司在本报告发布日之前 12 个月内曾向标的公司提供投资银行服务并收取报酬。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或不时会以自身或代理形式向客户出售及购买华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或其高级管理层、董事和雇员可能会持有本报告中所提到的任何证券（或任何相关投资）头寸，并可能不时进行增持或减持该证券（或投资）。因此，投资者应该意识到可能存在利益冲突。
- 本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽，亦不试图促进购买或销售该等证券。如任何投资者为美国公民、取得美国永久居留权的外国人、根据美国法律所设立的实体（包括外国实体在美国的分支机构）、任何位于美国的个人，该等投资者应当充分考虑自身特定状况，不以任何形式直接或间接地投资本报告涉及的投资者所在国相关适用的法律法规所限制的企业的公开交易的证券、其衍生证券及用于为该等证券提供投资机会的证券的任何交易。该等投资者对依据或者使用本报告内容所造成的一切后果，华泰证券股份有限公司、华泰金融控股（香港）有限公司、华泰证券（美国）有限公司及作者均不承担任何法律责任。

评级说明

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力（含此期间的股息回报）相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数），具体如下：

行业评级

增持：预计行业股票指数超越基准
中性：预计行业股票指数基本与基准持平
减持：预计行业股票指数明显弱于基准

公司评级

买入：预计股价超越基准 15%以上
增持：预计股价超越基准 5%~15%
持有：预计股价相对基准波动在-15%~5%之间
卖出：预计股价弱于基准 15%以上
暂停评级：已暂停评级、目标价及预测，以遵守适用法规及/或公司政策
无评级：股票不在常规研究覆盖范围内。投资者不应期待华泰提供该等证券及/或公司相关的持续或补充信息

法律实体披露

中国：华泰证券股份有限公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J

香港：华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809

美国：华泰证券（美国）有限公司为美国金融业监管局（FINRA）成员，具有在美国开展经纪交易商业务的资格，经营业务许可编号为：CRD#:298809/SEC#:8-70231

华泰证券股份有限公司

南京
南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019

电话：86 25 83389999/传真：86 25 83387521
电子邮件：ht-rd@htsc.com

北京
北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层/邮政编码：100032
电话：86 10 63211166/传真：86 10 63211275
电子邮件：ht-rd@htsc.com

深圳
深圳市福田区益田路 5999 号基金大厦 10 楼/邮政编码：518017
电话：86 755 82493932/传真：86 755 82492062
电子邮件：ht-rd@htsc.com

上海
上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120
电话：86 21 28972098/传真：86 21 28972068
电子邮件：ht-rd@htsc.com

华泰金融控股（香港）有限公司
香港中环皇后大道中 99 号中环中心 58 楼 5808-12 室
电话：+852-3658-6000/传真：+852-2169-0770
电子邮件：research@htsc.com
http://www.htsc.com.hk

华泰证券（美国）有限公司
美国纽约公园大道 280 号 21 楼东（纽约 10017）
电话：+212-763-8160/传真：+917-725-9702
电子邮件: Huatai@htsc-us.com
http://www.htsc-us.com