计算机行业

浅析 AI 大模型训练数据来源与版权挑战

核心观点：

- **AI 大模型训练数据来源广泛。**在算力可获得性提升以及算法同质化趋势下，训练数据成为影响大模型性能的重要因素。区别于传统 AI 模型，大语言模型通常使用公共文本数据集的混合体作为预训练语料库，而多模态大模型则需要大规模的图片和音视频等多模态数据。这些训练数据的来源广泛，包含公开渠道、企业自研、直接购买与合作交换等。

- **内容持有者对AI厂商态度各异。**部分内容持有者针对 AI 平台提出了各种维权诉求，已有数十起版权诉讼正在进行中。同时，另一部分内容持有者则选择了授权合作道路。版权纠纷实质上是商业利益之争，内容持有者具体选择诉讼还是合作取决于其商业模式、内容独特性和行业结构等因素。作家和艺术家们普遍倾向于抵制 AI 公司并控诉其侵权行为，而新闻媒体在版权斗争中则难以形成统一阵线。

- **确保训练数据的合法来源对于 AIGC 发展非常关键。**我们在去年的《从 Adobe 看 AIGC 如何重塑创意工具行业》报告中提到，训练数据的版权问题是 AIGC 商业化落地的重要阻碍。因此，只有解决了这一问题，才能在确保合法的前提下，推动生成式 AI 的商业落地。从 2023 年下半年开始，AI 数据版权诉讼开始进入白热化阶段，而内容合作则于 2024 年上半年加速，表明过去一年中版权问题已经成为 AI 领域的焦点，并且相关法律问题正在被逐步揭示与尝试解决。

- **2024 年有望成为 AI 训练数据版权之争的关键年。**关于 AI 训练数据版权诉讼，国内外尚未达成判例，重点案例的判决将对未来行业发展产生重要意义，需持续关注。同时，越来越多的公司正在明确其立场，显示出行业整体对于训练数据版权问题重视程度的提升。2024 年有望成为 AI 数据版权之争的关键年，将会有更多诉讼、谈判和合作展开，但未来授权合作或快于法律变革与监管介入。

- **当内容合作商对于训练数据版权的立场明确后，大模型研发的不确定性将被消除，应用发展也将进一步加速。**训练数据作为成本项，与下游应用的商业化推广密切相关，二者相辅相成。若数据合作显著加速，这将标志着 AIGC 应用即将迎来商业化落地的飞跃。

- **投资建议：**在众多种类应用中，创意工具软件与办公软件更为受益，标的方面，建议关注万兴科技（300624.SZ）、美图公司（01357.HK，广发传媒覆盖）、金山办公（688111.SH）等。

- **风险提示：**内容价值难以准确量化；行业竞争加剧；数据侵权阻碍下 游应用发展。

行业评级	买入
前次评级	买入
报告日期	2024-07-19

相对市场表现

分析师： 刘雪峰
SAC 执证号：S0260514030002
SFC CE No. BNX004
021-38003675
gfliuxuefeng@gf.com.cn

相关研究：

计算机行业 2024 年中期策略:下半年仍以结构性机会为主，基本面驱动是基础 2024-06-28

计算机行业:GPT-4o 发布，距离 AI 应用普及又近一步 2024-05-14

计算机行业:从 Adobe 看 AIGC 如何重塑创意工具行业 2023-12-27

联系人： 戴亚敏
daiyamin@gf.com.cn

重点公司估值和财务分析表

股票简称	股票代码	货币	最新收盘价	最近报告日期	评级	合理价值（元/股）	EPS(元)		PE(x)		EV/EBITDA(x)		ROE(%)	
							2024E	2025E	2024E	2025E	2024E	2025E	2024E	2025E
万兴科技	300624.SZ	CNY	48.53	2024/04/29	增持	112.83	0.77	1.03	63.03	47.12	50.72	39.42	7.40	9.00
金山办公	688111.SH	CNY	199.20	2024/04/24	增持	388.66	3.52	4.56	56.59	43.68	50.85	39.82	14.00	15.30

数据来源：Wind、广发证券发展研究中心

备注：表中估值指标按照最新收盘价计算

目录索引

图表索引

投资要点

1. 训练数据是构建和优化 AI 模型的基石，大模型常使用文本图片视频等公共数据集混合体作为预训练语料库。

(1) 在算力可获得性提升以及算法同质化趋势下，训练数据成为影响大模型性能的重要因素。具体而言，训练数据可以从数据规模、数据质量和数据即时性等方面对模型的训练效果产生影响。伴随着 AI 大模型的发展，训练数据集在规模和质量上也逐渐提升。 目前，AI 领域正经历从以模型为中心到以数据为中心的转变。

(2) 区别于传统的 AI 模型训练，大语言模型常使用维基百科、书籍期刊、论坛等多样化的公共文本数据集的混合体作为预训练语料库，而多模态大模型则需要大规模的图片、视频和语音等多模态训练数据。这些训练数据的获取方式多种多样，主要包含公开渠道、企业自研、直接购买和合作交换等方式。

2. 内容持有者针对 AI 平台提出的数十起版权诉讼正在进行中，另一部分则走上了授权合作道路。

(1) 目前，众多内容持有者正在针对 AI 平台提出各种维权诉求，有数十起 AI 训练数据版权诉讼正在进行中，指控 AI 厂商因使用受版权保护的内容进行训练，其中原告来自各行各业，包括作家、音乐出版商和新闻媒体等，以集体诉讼为主。

(2) 版权纠纷实质上是商业利益之争，各大巨头争夺的重点在于背后的经济利益。尽管生成式 AI 发展浪潮不可阻挡，传统内容持有者仍希望在这一过程中获得更有利的筹码，以避免被时代淘汰。

(3) 另一部分内容持有者则走上了授权合作道路，OpenAI、苹果、谷歌等公司与内容持有者签署了数十个内容许可协议，并有许多协议正在洽谈中。授权合作不仅可以为内容持有者带来与诉讼和解相当甚至更多的现金收益，而且速度更快，同时有助于将 AI 应用于其业务优化。而 AI 公司通过合作可以获取高质量的训练数据以改进模型效果，并避免侵犯版权。因此，这种合作对双方皆有利。

(4) 从行业属性来看，文本类数据集目前以新闻媒体为主， 已经拓展至 Reddit 论坛，但是书籍期刊的授权进展较为缓慢；从格式分类来看，数据授权合作也呈现从文本类拓展至图像、视频和语音等多模态数据的趋势。

(5) 关于授权的定价方式， 目前以直接订阅收费为主，此外还有采取分享收益间接付费，以及以标注出处作者等提供附加价值的方式。未来定价模式可能更多基于内容对 AI 模型的贡献，通过采用利润分享、按 API 访问次数收费等多种定价策略，内容持有者可以获取经常性收入，从而获得更合理的收益。这种定价方式不仅能够

反映内容的实际价值，还能够促进版权方和 AI 公司之间的合作，共同推动技术进步和商业模式创新。

(6) 内容持有者具体选择诉讼还是合作取决于其商业模式、内容独特性和行业结构等。作家和艺术家们普遍倾向于抵制 AI 公司并控诉其侵权行为，而新闻媒体在版权斗争中则难以形成统一阵线。

(7) 内容持有者面临着多重机会与挑战。① 机会端，首先，同一数据集可被用于训练多个模型，因此授权一般不具排他性；同时，内容持有者可以通过增加内容稀缺性以提升议价能力。② 挑战端，若不能与 AI 厂商达成协议，便有可能出局，因此内容持有者将会面临两难局面，起诉的高成本也可能带来压力，迫使其考虑和解；同时，由于缺乏统一标准和透明的评估机制， 内容持有者在谈判时可能处于不利地位，难以确保自身内容的合理定价 ；此外，内容持有者还将面临由于 AI 模型输出内容侵权而带来的法律问题。

3. 确保训练数据的合法来源对于 AIGC 的发展非常关键，2024 年有望成为 AI 数据版权之争的关键年，未来授权合作或快于法律变革与监管介入。

(1) 确保训练数据的合法来源对于 AIGC 的发展非常关键，只有解决了这一问题，才能在确保法律合规的前提下，推动生成式 AI 的广泛应用与商业落地。从 2023 年下半年开始，AI 数据版权诉讼开始进入白热化阶段，而内容合作则于 2024 年上半年加速，表明过去一年中版权问题已经成为 AI 领域的焦点，并且相关法律问题正在被逐步揭示与尝试解决。

(2) 关于 AI 训练数据版权诉讼， 国内外尚未达成判例。由于版权法的复杂性和模糊性，使得很难明确区分哪些行为构成侵权或不构成侵权，提升了判决难度。这种不确定性导致双方在法庭争议中浪费大量资源，可能需要数年时间才能确定这些诉讼中的具体指控与结果。重点案例的判决将对 AI 训练数据的版权界定有较大参考意义，有望在今年内初步了解法院对于此类版权诉讼请求的态度。

(3) 越来越多的公司正在明确其立场，显示出行业整体对于训练数据版权问题重视程度的提升。2024 年有望成为 AI 数据版权之争的关键年，将会有更多诉讼、谈判和合作展开，更多的公司和机构将明确其立场和策略，进一步推动版权争议的解决。

(4) 授权合作或快于法律变革与监管介入。具体节奏方面，在 2024 年下半年，部分案件可能会有初步判决结果，为后续案件提供参考，在诉讼过程中也可能出现和解的情况，推动法律和合作并行发展。而 2024 年第一批合作协议的签署与公开将为行业提供范例，在 2025-2026 年，部分 AI 数据合作将进入落地实施阶段，合作的可行性将得到初步验证，定价模式也将逐渐明确。随着更多案件进入判决阶段，预计将逐步形成较为明确的法律框架，为未来的版权保护和 AI 数据使用提供指导。

4. **投资建议：**(1) 数据将成为决定未来 AI 大模型效果的关键因素之一，进而成为 AI 公司的核心竞争力。随着训练数据成本的上升，只有大型科技公司才能负担得起这一资源，头部公司将因此受益。(2) 当内容合作商对于训练数据版权的立场进

一步明确后，大模型研发的不确定性将被消除，应用发展也将进一步加速。训练数据作为成本项，与 AIGC 应用的商业化推广密切相关，二者相辅相成。若数据合作显著加速，这将标志着 AIGC 应用即将迎来商业化落地的飞跃。在众多种类的应用中，创意工具软件与办公软件更为受益，前景广阔。标的方面，建议关注万兴科技（300624.SZ）、美图公司（01357.HK，广发传媒覆盖）、金山办公（688111.SH）等。

5. **风险提示：**内容价值难以准确量化；行业竞争加剧；数据侵权阻碍下游应用发展。

一、大模型常使用文本图片视频等公共数据集混合体作为预训练语料库

随着算力的可获得性提升，以及算法同质化趋势逐渐显现，数据成为影响 AI 大模型效果的重要差异化环节，其影响可以体现在数据规模、数据质量和数据即时性等方面。因此，AI 大模型的训练数据在规模与质量上逐渐提升，AI 领域也正经历从“以模型为中心”到“以数据为中心”的转变。

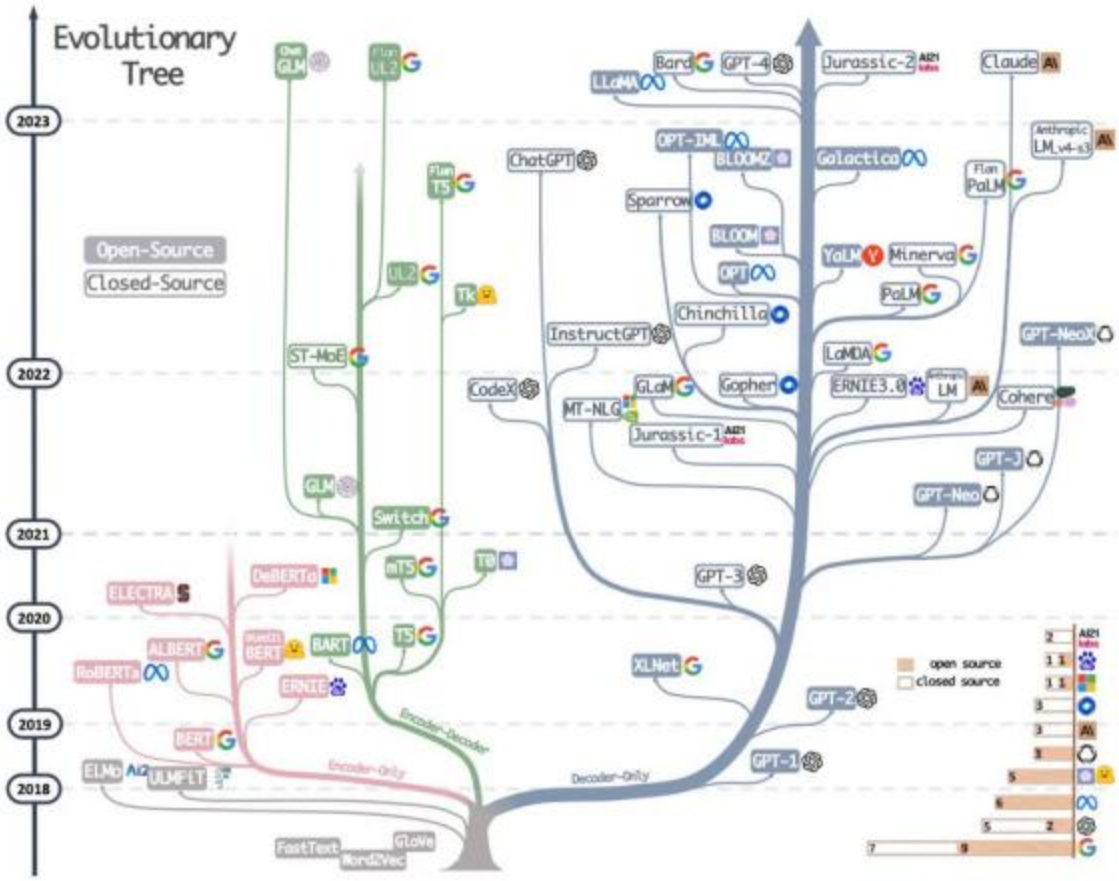
区别于传统的 AI 模型，大语言模型常使用维基百科、书籍期刊、论坛等多样的公共文本数据集的混合体作为预训练语料库，而多模态大模型则需要大规模的图片、视频和语音等多模态训练数据。这些训练数据的获取方式多种多样，主要包含公开渠道、企业自研、直接购买与合作交换等方式。

（一）数据成为影响 AI 大模型效果的重要差异化环节

训练数据是构建和优化 AI 模型的基石，AI 系统从输入的训练数据中进行学习。大模型训练数据包含文本、图像、语音、视频等结构化与非结构化的多种形式，大规模、高质量、多样化的训练数据集使得模型能够更深刻地理解上下文，并生成准确性与相关性更高的回复，相反，规模较小、低质量、缺乏多样性的数据集可能会导致模型结果产生偏差或生成无效回复。因此，训练数据在提升 AI 大模型的性能和应用效果中扮演着重要角色。

算力可获得性提升及算法同质化趋势显现，数据成为真正影响与区分 AI 大模型效果的重要环节。2017年，Transformer 架构的出现奠定了大模型算法架构的基石。Transformer 架构包含编码器（Encoder）和 解码器（Decoder），基于此诞生了三大技术路线——Decoder-Only、Encoder-Only 和 Encoder-Decoder。一方面，目前大模型的技术路径多集中在这三大技术路线，呈现同质化趋势；另一方面，算力可获得性在持续提升，瓶颈效应逐渐减弱。此外，有研究发现，在不同的 AI 大模型中使用相同的数据集，最终会表现出较为相似的行为。因此，在算力可获得性提升以及算法同质化趋势下，模型效果的独特性受到输入的训练数据集影响，训练数据成为区分且影响大模型性能的重要因素之一。

图 1：大模型的技术路径多集中在 Transformer 架构衍生出的三大技术路线



数据来源：《Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond》YANG 等，广发证券发展研究中心

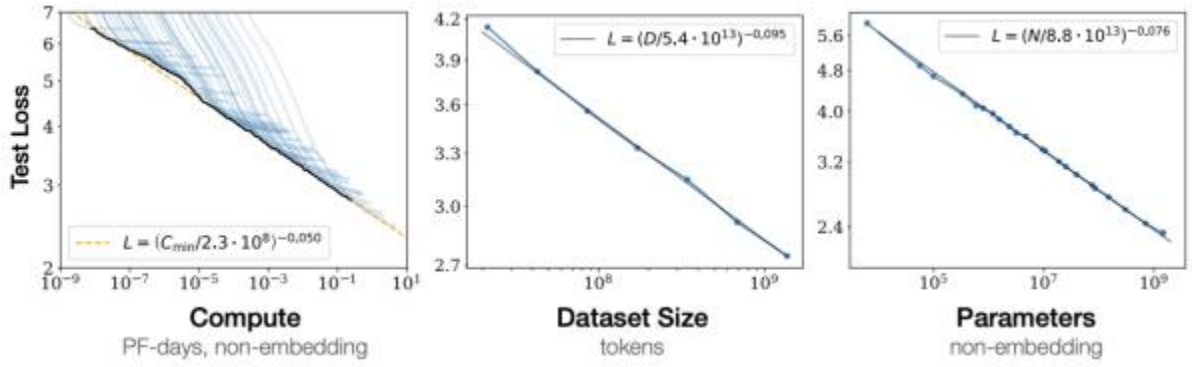
注：粉红色、绿色和蓝色分别为 Encoder-Only、Encoder-Decoder、Decoder-Only 三大技术路线

训练数据的影响可以体现在数据规模、数据质量和数据即时性等方面。在大语言模型的预训练阶段，由于需要消耗较多计算资源，通常不能进行无限次迭代，因此准备大规模高质量的语料库尤为重要。具体而言，训练数据可以从数据规模、数据质量和数据即时性等方面对模型的训练效果产生影响。

1. 从数据规模看，需要收集足够规模的数据才能满足大模型的训练需求。

根据大模型的尺度定律（Scaling Law），提升训练数据量是提升模型效果的重要一环。OpenAI 在 2020 年的一篇论文中最早提出 Scaling Law，Scaling Law 是一个经验性公式，其含义为，大模型性能主要与计算量、模型参数量和训练数据量三者的大小相关，而与模型的层次、深度、宽度等具体结构基本无关。

图 2：Scaling Law 提出大模型的性能主要与计算量、训练数据量和模型参数量三者的大小相关



数据来源：《Scaling Laws for Neural Language Models》Kaplan 等，广发证券发展研究中心

此外，根据 Scaling Law，当模型的参数或计算量按比例扩大时，模型性能也随之成比例提升。但只有当参数规模突破了某个阈值，大模型才会“涌现”出上下文学习、复杂推理等能力。而随着参数规模的增加，需要更多数据来训练模型，即模型参数与训练数据量之间也存在类似的比例关系。因此，为了与大模型的参数量匹配，也需要收集足够规模的训练数据。

图 3：部分经典模型的参数量与训练数据量之间的关系

数据来源：Thompson, A. D. (2024). LifeArchitect.ai.，广发证券发展研究中心

2. 从数据质量来看，使用低质量语料训练会损害大模型的性能。重复的数据会使模型的初始性能恶化，影响训练过程的稳定性，同时也会影响大模型的学习泛化能力。噪声和错误数据则会导致模型学习到不正确的信息，进而产生错误输出。因此，

为了保证模型的高性能，需要尽量使用高质量的语料库，去除其中的重复、噪声和错误数据等低质量语料。

3. 从数据即时性看，在过时的数据上进行训练同样不利于模型达到最优性能。大模型的训练数据通常源于已有的网页、书籍和其它公开数据等，这些数据通常于特

定时间点前被收集。而由于大多数大模型没有内置的实时数据访问或动态更新机制，一旦训练完成并进行部署，其知识也将会停止在最后一次更新训练时，除非进行再次训练和更新，此后发生的任何事件或新信息都不会被模型所学习。

例如在 ChatGPT 刚推出时，训练数据截至 2021 年 9 月，可能导致不准确或过时的回复。因此，相较于大模型的固定训练数据集而言，若能获取最新的新闻数据，则更具有即时性。目前，ChatGPT 的训练数据已更新至 2023 年 12 月。

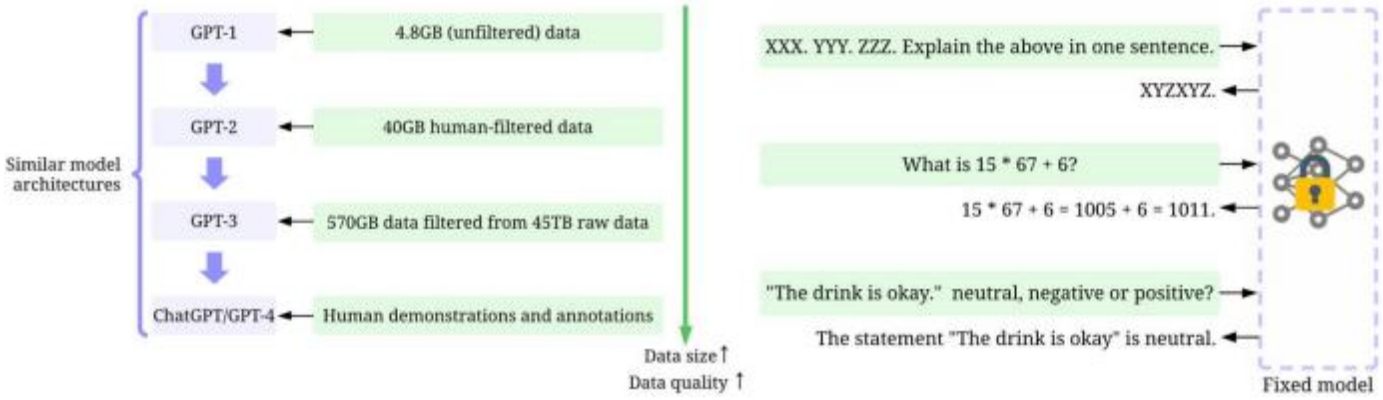
表 1：GPT系列大模型的训练数据集截止时间及模型推出时间梳理

模型名称	训练数据集截止时间	模型推出时间
GPT-1	-	2018 年 6 月
GPT-2	-	2019 年 2 月
GPT-3	-	2020 年 5 月
GPT-3.5	2021 年 9 月	2022 年 3 月
GPT-4	更新至 2023 年 12 月	2023 年 3 月
GPT4-Turbo	更新至 2023 年 12 月	2023 年 11 月
GPT-4o	2023 年 12 月	2024 年 5 月

数据来源：OpenAI，Microsoft Blog，Thompson, A. D. (2024). Models Table. LifeArchitect.ai.，广发证券发展研究中心

AI 大模型的训练数据集在规模和质量上逐渐提升。以 OpenAI 的 GPT 系列模型为例，2018 年的 GPT-1 数据集约为 4.8GB，2019 年的 GPT-2 数据集约为 40GB，而 2020 年的 GPT-3 数据集规模已超过 500GB，质量上也逐渐提升。尽管如此，GPT 系列模型架构并未发生较大变化，都是基于 Transformer 架构。

图 4：AI 大模型的训练数据集在规模和质量上逐渐提升



数据来源：《Data-centric Artificial Intelligence: A Survey》ZHA 等，广发证券发展研究中心

AI 领域正经历从以模型为中心到以数据为中心的转变。吴恩达等学者在 2021 年提出，AI 领域正经历从 Model-Centric AI（以模型为中心）到 Data-Centric AI（以数据为中心）的转变。

1. Model-Centric AI （以模型为中心）主要关注于通过优化模型的架构算法来提高 AI 系统性能。在 Model-Centric AI 中，研究人员的重点是设计更复杂高效的模型架构和算法，以在固定的数据集上获得更好的表现。这种方法已有几十年的历史，积累了丰富经验。然而， 在这种模式下，数据集在训练过程中通常保持不变，若数据质量问题和数据偏差等未被充分处理，模型的精度和性能可能会受到影响。

2. Data-Centric AI（以数据为中心）强调通过系统地工程化和优化数据来提升 AI 系统的性能。在 Data-Centric AI 中，数据的质量和规模是关注焦点，数据分析过程将持续贯穿整个 AI 系统的生命周期。通过对数据进行系统清洗、标注与增强，可以显著提高模型的精度和性能。这种数据优先的策略提升了准确度与一致性，从而实现高质量的 AI 系统。

表 2：Model-Centric AI 与 Data-Centric AI 对比

对比角度	Model-Centric AI	Data-Centric AI
主要关注点	代码	数据
研究人员的关注点	90%	少于 10%
研究跨度	30 年	大约 3 年
数据分析	一次性	持续（N次）
准确性	低	高
质量保证	没有	有
实践	代码优先	数据优先
漂移敏感性	概念和数据都敏感	不敏感
数据检查	仅在训练前	在整个生命周期内
反馈	缓慢且不足	及时
结果的可解释性	复杂	简单
数据准备步骤	有限	全面

数据来源：《A Data-Centric AI Paradigm for Socio-Industrial and Global Challenges》Abdul Majeed 等、广发证券发展研究中心

（二）AI 大模型训练数据来源分类

AI 大模型的训练数据与传统 AI 训练数据有所差异。对于传统 AI 训练，常用的有 MNIST、ImageNet、Open Images 等数据集，这些数据集可用于自然语言处理、计算机视觉和语音识别等传统 AI 应用。研究人员经常使用这些数据集作为创建、评估和对比 AI 模型有效性的标准，用户也可以根据开放许可条款访问、使用、更改和共享这些公开数据集。

大语言模型在训练过程中所需的数据内容由具体阶段所决定。以 ChatGPT 为例，其基础模型训练过程可分为三个主要阶段：预训练、监督微调（SFT）和强化学习

（RLHF），后两个阶段也被称为对齐（Alignment）阶段。有时也需要结合某行业的专业知识进行训练和对齐，即行业模型阶段。通过在各阶段输入不同的训练数据，模型能够提供高效准确的输出并满足特定应用场景需求。

1. **预训练阶段：**在预训练阶段，模型需要输入包括书籍期刊、新闻报道、学术论文、对话文本和代码等在内的多样化数据。该阶段的目标是通过大规模的多样化数据，让模型建立起基本理解与知识架构。因此，这个阶段的训练数据特点是“广”，即涵盖范围广泛。

2. **监督微调阶段（SFT）：**在监督微调阶段，数据由人工标注人员设计，包括具体的问答对示例。通过输入这些标注数据，模型能够在一些未见过的任务中提高判断能力，泛化性得以提升。这一阶段对于训练数据的要求较高，需要精心设计和高质量的人工标注。

3. **强化学习阶段（RLHF）：**在强化学习阶段，模型的目标是通过人类反馈进行调整，使其输出结果更符合认知。这个过程包括对模型回答进行评分与排序，以便模型学习如何更好地回答用户问题。

   强化学习阶段与监督微调阶段的数据需要来自人类的高质量反馈，其特征可以总结为“齐”，即让大模型的输出结果和人类需求对齐。

4. **行业模型：**如将经过微调的模型应用于法律、金融等特定行业，则需要结合该行业的专业知识进行训练与对齐。此时，所需的数据则包括行业数据库、专业文档和特定领域的网站内容等，需要具有较高的专业性和行业深度，其特征可以用“专”来概括，即专业性强。

图 5：大语言模型分阶段训练数据来源

数据来源：《大模型训练数据白皮书》阿里云，广发证券发展研究中心

大语言模型常使用多样的公共文本数据集的混合体作为预训练语料库。具体而言，国内外大语言模型训练数据集的主要来源为维基百科、书籍期刊、论坛、代码、

Common Crawl（CC）网页数据集和其它数据集等，其中部分经典模型所使用的 训练数据分类拆解如下表所示。

表 3：部分模型所使用的训练数据分类

模型	总 tokens (T)	总规模 (GB)	维基百科	书籍期刊	论坛	代码	网页 (CC/C4)
Piper monorepo	37.9	86000				86000	
FineWeb	15	44000					44000
GPT-4	13	40000					
MassiveText ML	5	20000	48	12853	-	2754	4544
PaLM 2	3.6	13000					
Infiniset	2.8	12616	1569	1632	6277	1569	1569
MADLAD-400	3	12000					120000
MassiveText EN	2.35	10550	12.5	2264		3100	5173
Stability New Pile	1.5	5000					
LLaMA	1.2	4749	83	177	78	328	4083
The Pile v1	0.247	825	6	362	166.71	95	227
GPT-3	0.499	753	11.4	122			620
Megatron-11B		161	11.4	4.6			145
GPT-2		40					40
GPT-1		4.6		4.6			

数据来源：Thompson, A. D. (2024). Models Table. LifeArchitect.ai.，广发证券发展研究中心

注：C4 为 Common Crawl 的仅包含英文的过滤版本

我们对于以上五类公共文本数据集进行逐一分析。

1. 维基百科

维基百科是一个多语言协作式在线百科全书，由于其引用、撰写风格较为严谨，以及跨语言与领域的内容，维基百科的文本被视为非常有价值的资源，主要研究实验室通常会使用仅包含英文的过滤版本维基百科作为数据集起点。

2. 书籍期刊

书籍期刊也是大模型训练数据的重要来源。一方面，由虚构和非虚构书籍混合而成的叙述内容对于连贯的故事讲述和回答较为适用，另一方面，因为学术写作的输出涉及众多专业科学领域，且数据格式复杂，因此期刊可以有效提升大语言模型对于科学知识的理解。

目前，有许多书籍数据库收集了涵盖多种语言的公开可用电子书，并将其整理成易于使用的格式，例如 Project Gutenberg 、Smashwords（BookCorpus）、Books3 等数据集。而期刊数据库则包括 ArXiv 和美国国家卫生研究院（NIH）等数据集，ArXiv 主要集中在数学、计算机科学和物理领域，其用 LaTeX 语法编写的论文可以将不同格式数据转换为统一形式，对于公式、符号、表格等内容的表示也较为适合模型学习，使得大模型更好地处理和分析科学文本数据。

3. 论坛

论坛数据指的是来自 StackExchange 等问答网站和 Reddit 等社交媒体平台的对话或视频字幕数据集等。Stack Exchange 是一个围绕用户提供问题和答案的网站，Stack Exchange Data Dump 包含了在 Stack Exchange 网站集合中所有用户贡献的内容的匿名数据集，是截止到 2023 年 9 月为止公开可用的最大的问题-答案对数据集合之一，涵盖了编程、园艺和艺术等广泛主题。而社交媒体平台 Reddit 是一问一答的 QA 内容形式，且基本都是回复的真实情况表达，为了使得回答更符合人类表达模式，AI 厂商非常需要这类数据来进行高质量的预训练和监督微调。

4. 代码

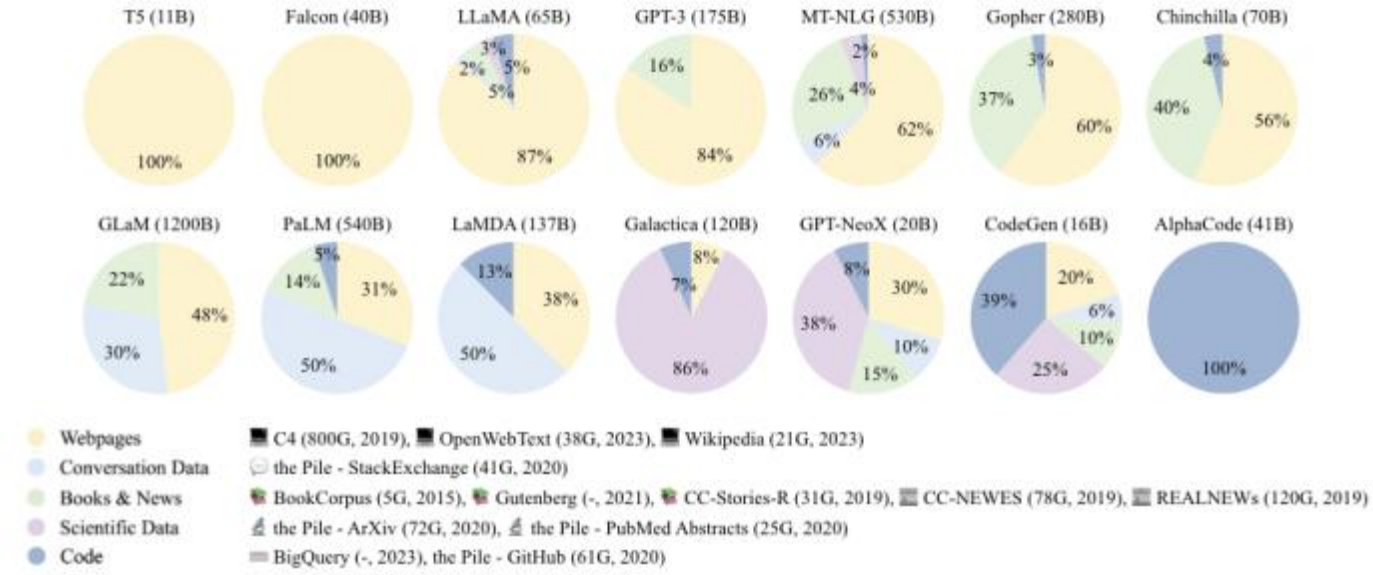
代码数据是大语言模型进行代码生成、代码补全等任务所必备的数据。代码数据不仅包括程序代码本身，还包含丰富的注释信息，通过在大量代码上进行预训练，可以显著提升模型的代码生成效果。与普通的自然语言文本相比，代码通常是一种格式化语言，对应着长程依赖和精确的执行逻辑，其表达中的特定语法结构、关键字以及编程范式对代码的含义与功能起着重要影响。

代码数据主要来源于 GitHub 等代码仓库以及 StackOverflow 等编程问答社区。在代码仓库中，包含了各种编程语言在内的开源代码，应用范围广阔，这些代码库中的代码通常经过严格的代码评审和实际使用测试，因此具有较高质量与可靠性；而在 StackOverflow 等编程问答社区中，数据则包含了开发者提出的问题、其他开发者的回答以及相关的代码示例，提供了丰富的语境和真实的代码使用场景。

5. 网页

网页数据包含 Common Crawl （CC）数据集和 C4 数据集等。Common Crawl 是一个自 2008 年起持续抓取的大规模 Web 爬虫数据集，包括原始网页、元数据和文本摘录，涵盖了不同语言和领域的文本。Common Crawl 每月爬取数十亿个页面，将这些数据存储在可搜索的数据库中，并提供一些列开源工具，帮助用户下载和分析数据。Common Crawl 所有抓取数据均免费开放，无需注册或申请许可，使得任何人都能够访问大量的网络信息并进行研究与开发。CC 数据集规模庞大，包含数十亿个页面和数百 TB 的数据，覆盖全球众多网站，主要研究实验室通常使用其仅包含英文的过滤版本 C4 作为数据集的起点。CC 数据集最新的数据是在 2024 年 5 月抓取的，存档包含 2.70 亿个页面。

图 6：部分经典大语言模型所使用的训练数据组成情况



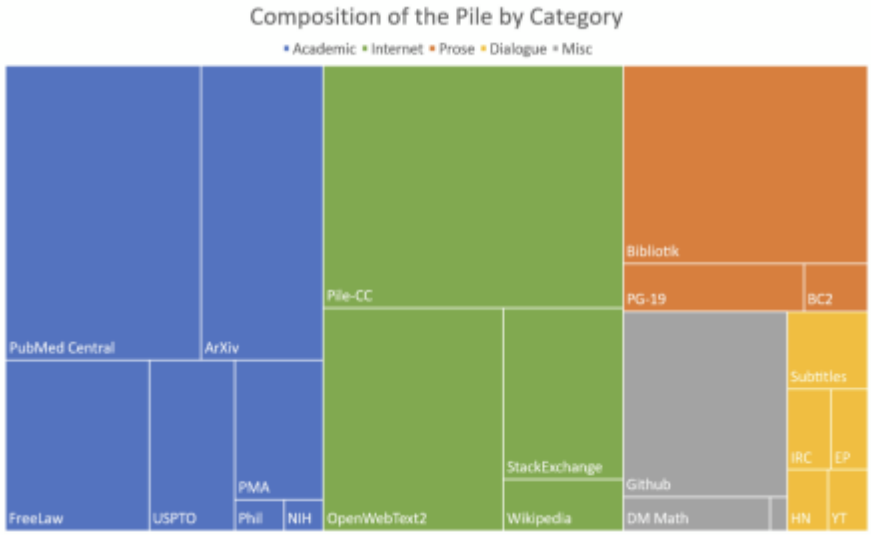
数据来源：《A Survey of Large Language Models》Zhao 等，广发证券发展研究中心

多模态大模型需要大规模的多模态训练数据。在大语言模型迅速发展的同时，大模型开始迁移到图像、视频和语音等其他模态领域，并与大语言模型融合，形成多模态大模型。多模态大模型把各种感知模态结合起来，可以以更全面综合的方式理解和生成信息，最终实现更丰富的应用。多模态大模型的训练需要有大模型的多模态数据，例如图像-文本对、视频-文本对等数据集。图像-文本对包含了图像以及描述该图像内容的文本数据，让模型可以学习组成图像的像素之间、文字与图像的关联。视频-文本对则包含了视频以及描述视频的文本，让模型不仅可以学习单个画面，还可以理解视频中的时间序列和动态变化。

基于上述数据，建立了 Pile 数据集、RefinedWeb 数据集等许多经典的训练数据集，以及一批涵盖多种模态的大模型数据集。

1. Pile 数据集是一个用于大语言模型训练的大规模文本语料库，由 Common Crawl、Wikipedia、OpenWebText、ArXiv、PubMed 等 22 个不同的高质量子集构成。Pile 数据集包含了大量不同领域和主题的文本，从而提高了训练数据集的多样性和丰富性，总计规模大小超过 800G，其数据类型组成如下图所示。

图 7：Pile 数据集组成分类

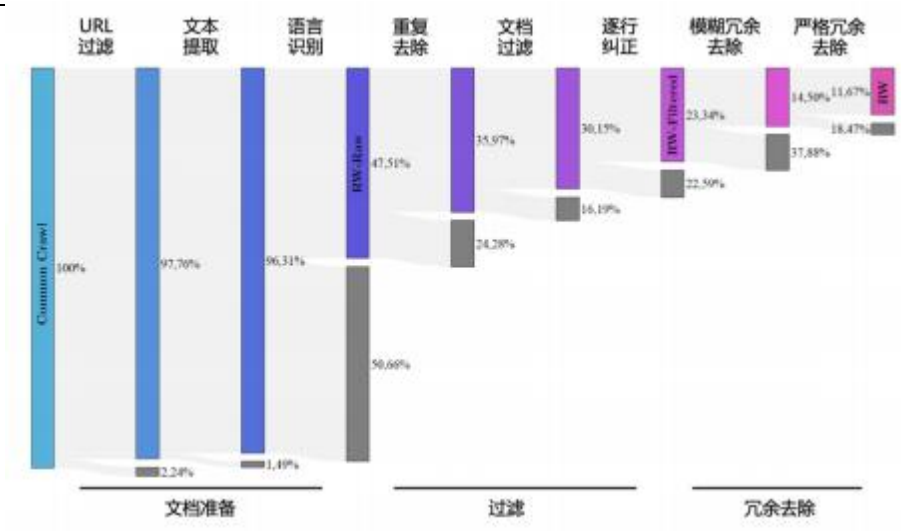


数据来源：《The Pile: An 800GB Dataset of Diverse Text for Language Modeling》Gao等，广发证券发展研究中心

注：所占面积大小表示数据在整个数据集中所占的规模。

2. RefinedWeb 是由位于阿布扎比的技术创新研究院在开发 Falcon 大语言模型时同步开源的大语言模型预训练集合，主要由从 CommonCrawl 数据集过滤的高质量数据组成，下图展示了由 CommonCrawl 数据集得到 RefinedWeb 数据集的 Pipeline 过程。

图 8：由 CommonCrawl 数据集得到 RefinedWeb 数据集的 Pipeline 过程



数据来源：《The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only》Penedo 等，广发证券发展研究中心

3. 此外，常见的还包括 ALIGN、VAST-27M、WebVid-2.5M 等多模态数据集。大模型常用的公开数据集如下表所示。

表 4：大模型常用的公开数据集

数据集类型	数据集名称	数据量和简介
语言大模型预训练数据集	BookCorpus	2.24G，包括超过 11000 本电子书，涵盖广泛的主题和类型（如小说和传记）
	OpenWebText	38G，从 Reddit 上共享的 URL 中提取的 Web 内容，且至少获得了 3 次赞成
	Common Crawl	PB 级规模，一个大型网站抓取数据集，包含原始网页数据，元数据提取和文本提取等内容
	The Pile	825G，一个大规模、多样化、开源的文本数据集，内容包括书籍、网站、代码、和社交媒体等
语言大模型指令微调数据集	Stanford Alpaca	21.7M，开源的 SFT 多样化数据集，包含 52000 条指令数据，涵盖创作、生成、设计等多维度
	static-hh	90M，开源的 SFT 多样化数据集，包含 100000 条人类对话数据
语言大模型强化学习微调数据集	ShareGPT	1.8G，由用户共享的对话 SFT 数据集，包含了超过 1 亿条来自不同领域主题的对话样本，
	HH-RLHF	120M，Anthropic 创建的大型 RLHF 训练数据集，包含 161000 条人工标注的数据
	zhihu_rlhf_3k	16M，知乎开源的 RLHF 数据集，包含 3000 条基于知乎问答的人类偏好数据
	BeaverTails	52M，北京大学开源的 RLHF 数据集，包含 302000 个数据对，覆盖 7774 个问题
图片-文本多模态	SBU	1M，图片-标题对
图文数据集	COCO	330K，图片/1.5M 标题
	Visual Genome	108K，图片-标题对
	Conceptual	12M，图片标注对
	ALIGN	1.8B，图片-标题对
	COYO-700M	747M，图片-标题对
视频-文本多模态数据集	HowTo100M	136M，视频标注对 / 134500 小时
	WebVid-2.5M	2.5M，视频标注对 / 13000 小时
	YT-Temporal-180M	1.8M，视频标注对
	HD-VILA-100M	100M，视频-标题对
图文音多模态数据集	VALOR-1M	1M，视频-音频-文本数据组
	VAST-27M	27M，视频-字幕-文本数据组

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/198056111021006114