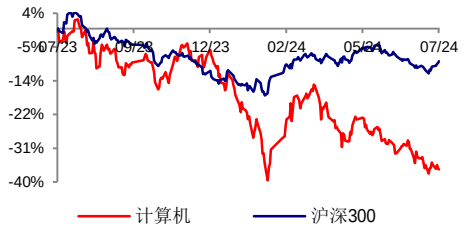
计算机行业

浅析 AI 大模型训练数据来源与版权挑战

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

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。

人工智能产业链联盟

星主： AI产业链盟主

知识星球

微信扫描预览星球详情

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。

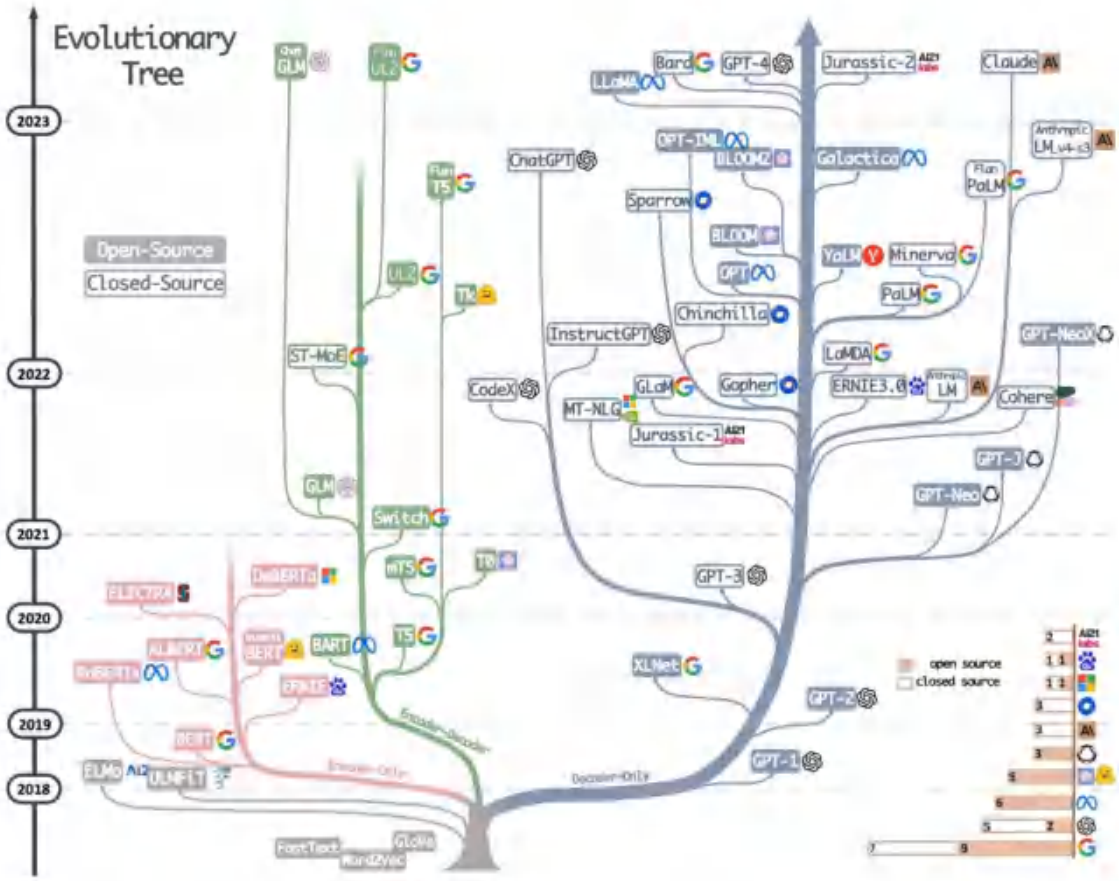
区别于传统的 AI 模型，大语言模型常使用维基百科、书籍期刊、论坛等多样的公共文本数据集的混合体作为预训练语料库，而多模态大模型则需要大规模的图片、视频和语音等多模态训练数据。这些训练数据的获取方式多种多样，主要包含公开渠道、企业自研、直接购买与合作交换等方式。

（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

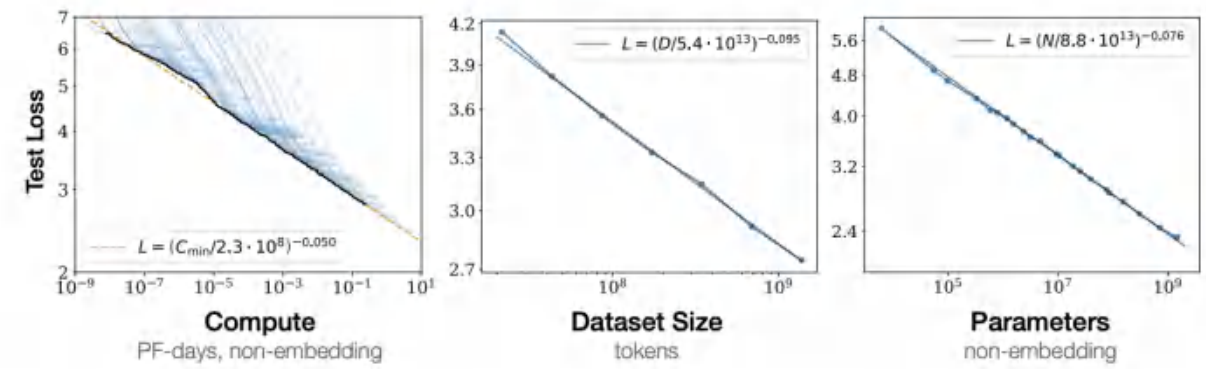
注：粉红色、绿色和蓝色分别为 Encoder-Only、Encoder-Decoder、Decoder-Only 三大技术路线

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。

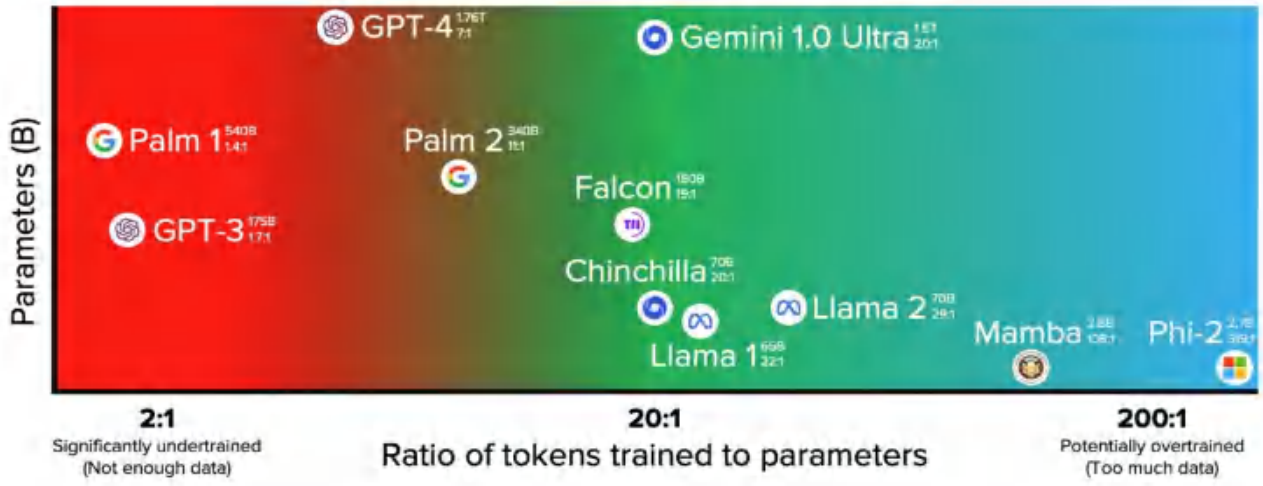
图 2：Scaling Law 提出大模型的性能主要与计算量、训练数据量和模型参数量三者的大小相关



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。

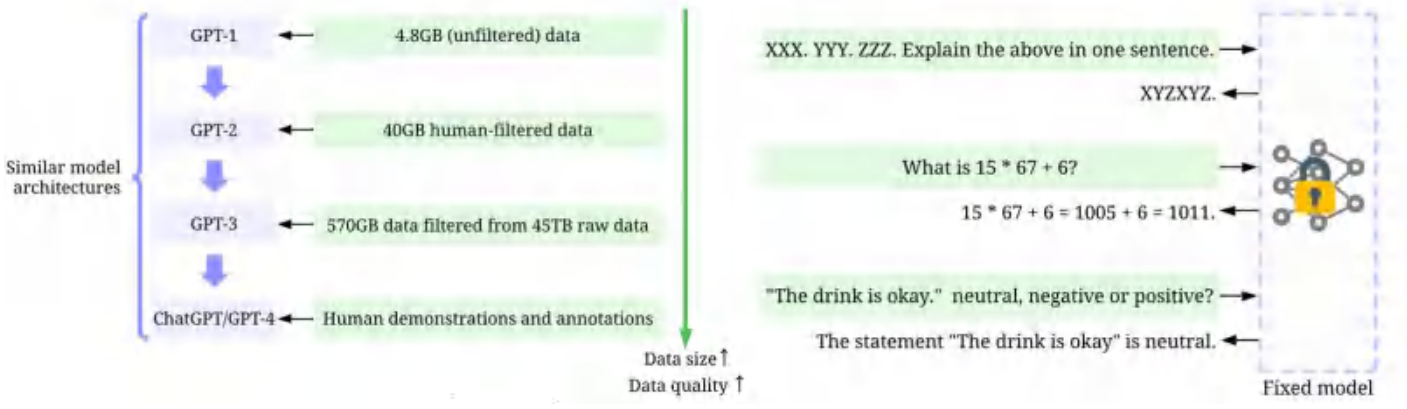
表 1：GPT系列大模型的训练数据集截止时间及模型推出时间梳理

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

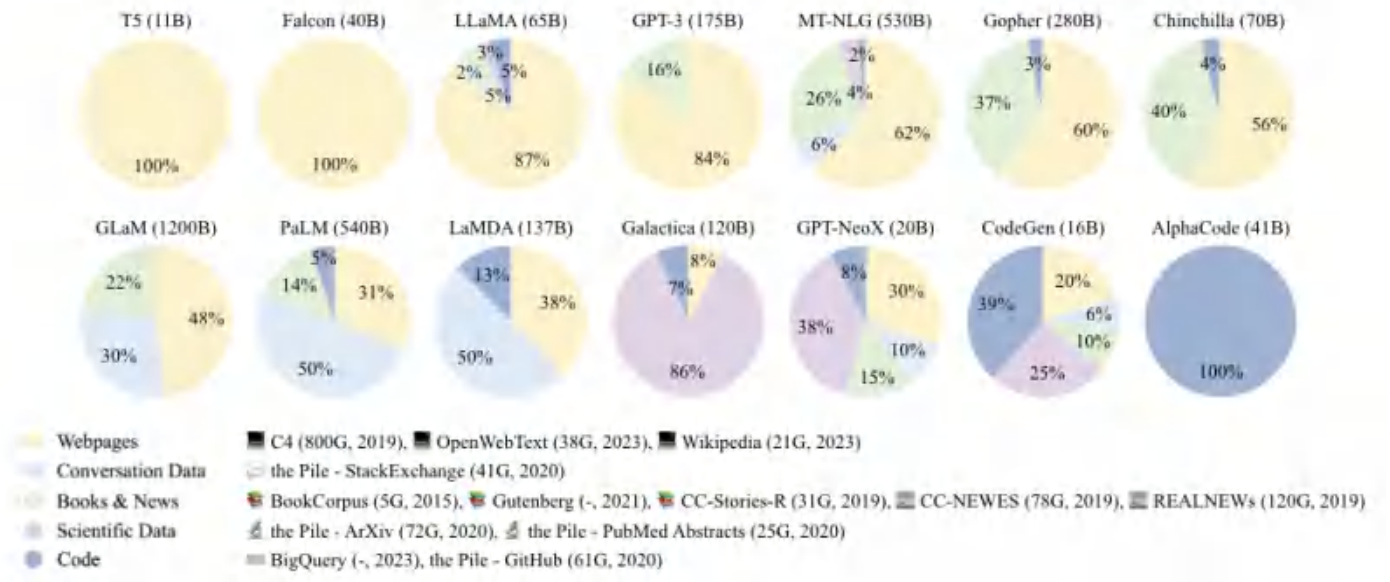
代码数据是大语言模型进行代码生成、代码补全等任务所必备的数据。代码数据不仅包括程序代码本身，还包含丰富的注释信息，通过在大量代码上进行预训练，可以显著提升模型的代码生成效果。与普通的自然语言文本相比，代码通常是一种格式化语言，对应着长程依赖和精确的执行逻辑，其表达中的特定语法结构、关键字以及编程范式对代码的含义与功能起着重要影响。

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



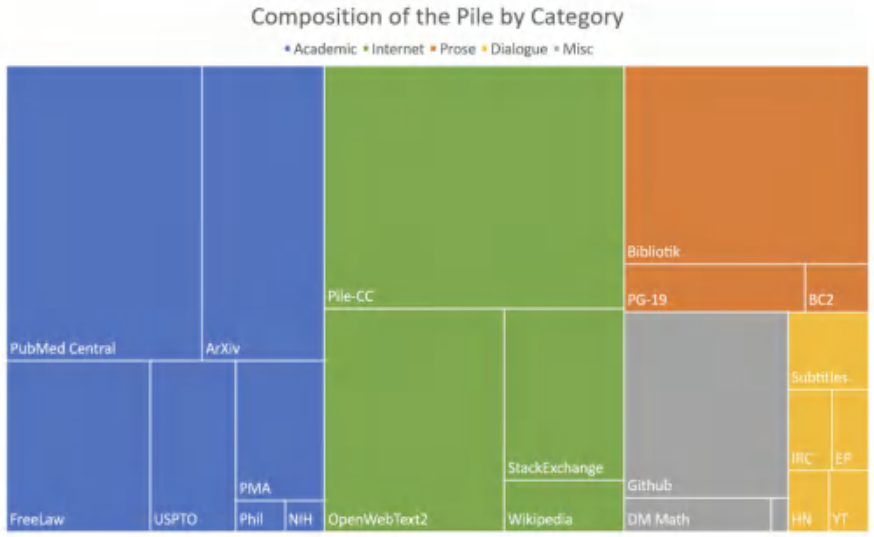
数据来源：《A Survey of Large Language Models》Zhao 等，广发证券发展研究中心

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

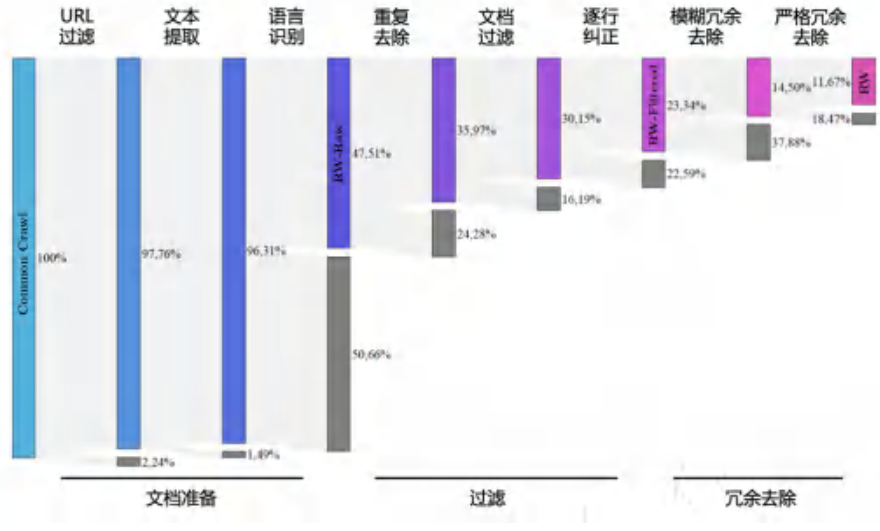


数据来源：《The Pile: An 800GB Dataset of Diverse Text for Language Modeling》Gao等，广发证券发展研究中心

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

数据来源：《中国人工智能系列白皮书——大模型技术（2023 版）》中国人工智能学会，广发证券发展研究中心

（三）AI 大模型训练数据获取途径

数据成为影响 AI 大模型效果的重要差异化环节，其规模、质量与多样性直接影响模型的性能和应用效果。那么以上提到的各种类别的训练数据从何处获取？其获取途径多种多样，主要包含公开渠道、企业自研、直接购买和合作交换等方式。

公开渠道是获取训练数据的重要途径之一。公开数据集通常由研究机构、大学、政府组织或开源社区提供，涵盖领域广泛。例如，Wikipedia 提供了大量经过验证的百科全书内容，Common Crawl 数据集包含了从互联网中抓取的大量网页数据，而 Reddit 则提供了丰富的社交媒体讨论和用户生成内容，研究者们可以使用这些数据集进行大模型训练，有效推动 AI 技术发展。

企业自研数据是指企业通过自身渠道生成和收集的数据。这些数据通常具有更高的质量与针对性，能够更好地满足特定应用场景的需求。例如，谷歌通过扩展服务条款，利用公开的谷歌文档、谷歌地图上的餐厅评论和其他在线资料，为其 AI 产品提供服务。众多企业通过自身业务流程和用户互动，积累了大量结构化和非结构化数据，为行业特定的 AI 应用提供了宝贵的训练素材。但需要注意的是，企业自研数据在使用过程中要保证合法性。

直接购买也是获取训练数据的常见方式。市场上有许多提供有偿数据服务的商业团队和公司，其根据数据类型、数据规模或是否需要标注等规则向 AI 开发者提供高质量的数据集。例如，Scale AI 等数据标注公司提供大规模、高质量的标注服务，而数据市场平台如 Kaggle 和 AWS Data Exchange 则允许开发者购买和使用各种类型的数据集，涵盖从金融数据到医疗记录的广泛领域。通过与这些数据商合作，AI 公司可以使用高质量的数据集来训练其模型。

最后，数据交换和合作也是获取高质量训练数据的重要手段。许多公司和研究机构通过合作来共享各自的数据资源，实现互利共赢。在某些行业，企业之间通过数据联盟和共享平台，交换非竞争性的数据，例如医疗行业中的研究机构和医院共享匿名化的患者数据。此外，政府和公共机构也与私营企业合作，共享公共数据资源，以推动技术创新和公共服务提升。

我们总结，AI 大模型的训练离不开高质量的数据来源，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、直接购买和交换合作等方式。然而，随着 AI 技术的快速发展和广泛应用，AI 厂商在获取和使用数据时，常面临法律和道德上的挑战，围绕数据版权的争议也在日益增多。

二、AI 大模型训练面临的数据版权挑战

生成式 AI 领域发展迅速，然而伴随的却是日益增多的数据版权纠纷。版权纠纷主要聚焦于模型训练阶段未经授权的版权利用行为，此外，也包含 AI 模型输出本身对于版权的侵犯。目前，内容持有者正在针对 AI 平台提出各种维权诉求，有数十起版权诉讼正在进行中，另一部分内容持有者则走上了授权合作的道路。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，并面临着多重机会与挑战。

（一）训练数据需求下，数据版权诉讼激增

生成式 AI 领域发展迅速，数据版权纠纷日益增多。因为 AI 大模型需要大量数据进行训练，为了获得这些数据，众多 AI 公司冒着被起诉的风险，“抓取”互联网内容来获取数据，或在其它受知识产权保护的内容上训练模型，因此导致了数据版权诉讼激增。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。

表 5：AI 训练数据版权诉讼统计

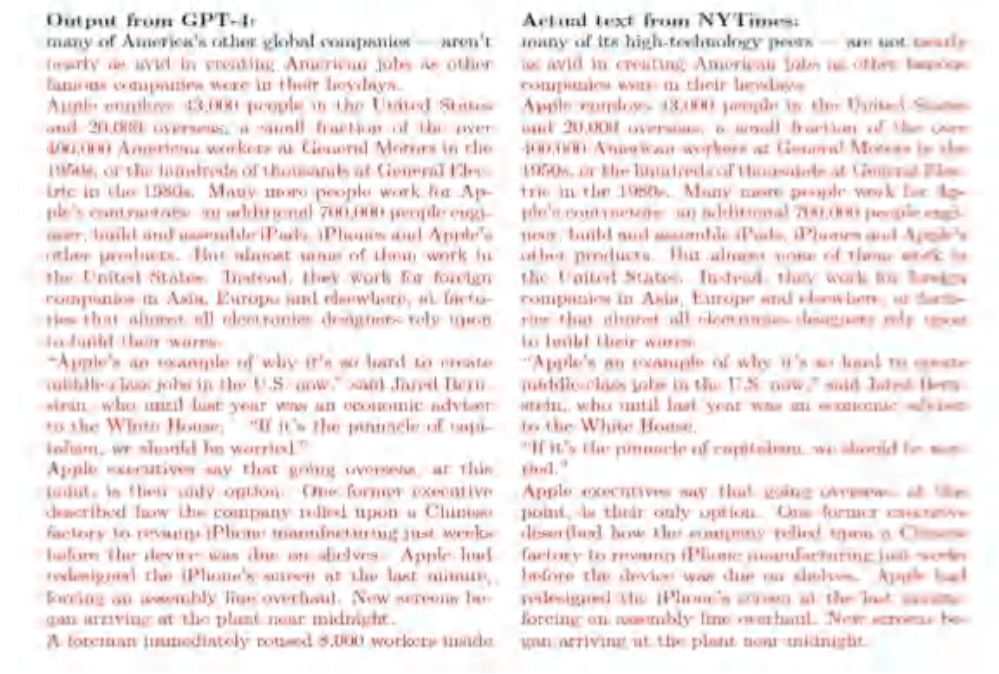
诉讼方	诉讼方分类	被诉方	提交日期
Thomson Reuters	信息服务提供商	ROSS Intel. Inc.	2020/5/6
J. Doe 1 和 J. Doe 2	匿名个人	GitHub / OpenAI / 微软	2022/11/3
Sarah Andersen 等	视觉艺术家集体	Stability AI / Midjourney / DeviantArt / Runway AI	2023/1/13
Getty Images	视觉媒体	Stability AI	2023/2/3
Tremblay 等	作家集体	OpenAI	2023/6/28
Kadrey 等	作家集体	Meta	2023/7/7
Leovy 等	个人（互联网用户和版权持有人集体）	Alphabet Inc.	2023/7/11
Authors Guild	作家集体	微软、OpenAI	2023/9/19
Huckabee 等	作家集体	Bloomberg / Meta / 微软	2023/10/17
Concord Music Group 等	音乐出版商集体	Anthropic PBC	2023/10/18
The New York Times	新闻媒体	微软、OpenAI	2023/12/27
Intercept Media，Raw Story Media，AlterNet Media	三家新闻媒体	微软、OpenAI	2024/2/28
Nazemian 等	作家集体	英伟达	2024/3/8
Daily News, LP; Chicago Tribune Company, LLC 等	八家报纸出版商	微软、OpenAI	2024/4/24

调查报道中心（Center for Investigative Reporting）	非营利新闻机构	微软、OpenAI	2024/6/27

数据来源：BakerHostetler，广发证券发展研究中心

版权纠纷主要聚焦于模型训练阶段未经授权的内容使用行为。2023 年 12 月，美国报业巨头纽约时报公司向 OpenAI 及微软提起诉讼，指控其未经许可使用《纽约时报》的数百万篇文章训练 ChatGPT 模型，侵害了纽约时报的版权，并构成不正当竞争。诉讼称 OpenAI 和微软将纽约时报的文章输入至其大语言模型的内存中，以便 ChatGPT 和 Copilot 可以访问这些信息。在纽约时报提出的一些例子中，ChatGPT 向用户提供的文章近乎逐字摘录《纽约时报》而没有适当引用，但这些文章实际需要付费订阅才能阅读。

图 9：《纽约时报》提供的 ChatGPT 输出文本与该报文章类似的例子



Output from GPT-4:

many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capitalism, we should be worried."

Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone's screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight.

A foreman immediately roused 8,000 workers inside

Actual text from NYTimes:

many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House.

"If it's the pinnacle of capitalism, we should be worried."

Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone's screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight.

数据来源：美国纽约南区地方法院文件，广发证券发展研究中心

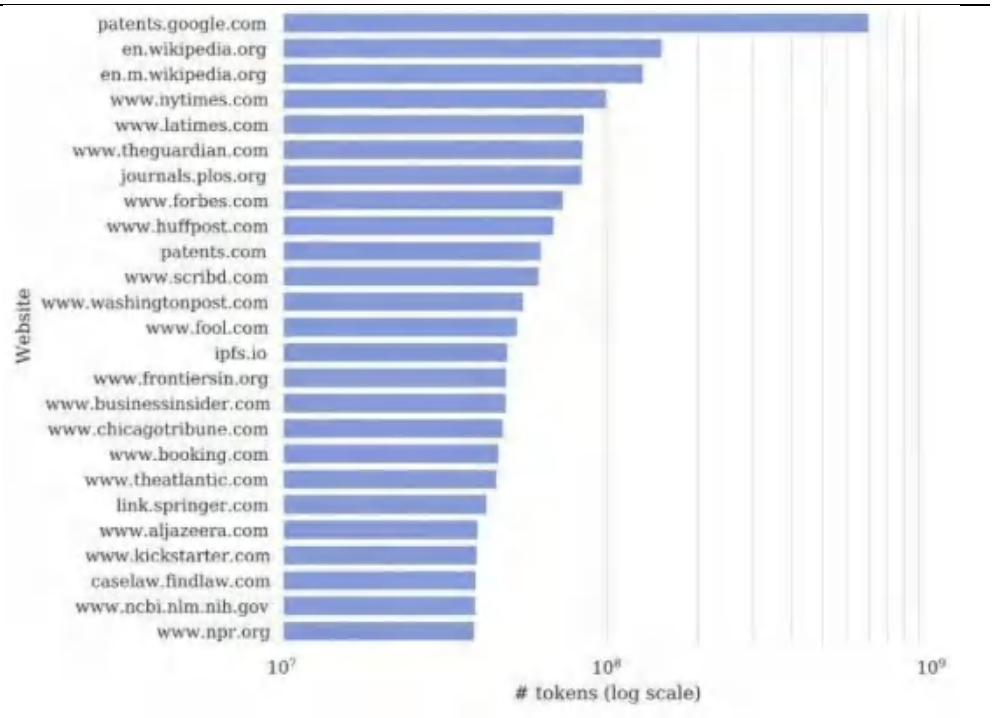
版权纠纷也包含 AI 模型输出本身对于版权的侵犯。2023 年 2 月，美国视觉媒体公司 Getty Images 对 Stable Diffusion 的开发者 Stability AI 提起诉讼，指控 Stability AI 未经许可从其数据库复制了超过 1200 万张图像，公然侵犯了其版权与商标保护权。在下图中，Stable Diffusion 生成的图片便带有 Getty 的商标标识。

图 10：Getty 的原始图片和由 Stable Diffusion 生成的带有 Getty 商标的图片

数据来源：The Verge，广发证券发展研究中心

训练数据的版权纠纷主要源于数据集中的“脏数据”，比如本身便为盗版数据。C4 数据集是谷歌 T5 和 Meta LLaMA 等很多知名大模型的训练材料，美国艾伦人工智能研究院为研究该数据集里具体包含哪些材料来源，对其进行了拆解。结果显示，其实际包含的约 1000 万个网站数据中，有很大一部分是盗版电子书网站等非正当的数据源。而一些诸如创意产品众筹网站、个人博客也包含其中且排名靠前，表明这些数据虽然使用权重较高，但数据版权方可能并未获得任何授权或报酬。基于这一思路，《纽约时报》起诉 OpenAI 或有迹可循，因为 ChatGPT 的训练使用了 Common Crawl 数据集（C4 数据集是其过滤版本），而《纽约时报》正是 C4 数据集中除了谷歌专利数据之外最大的数据来源。

图 11：C4 数据集拆分



数据来源：《Documenting Large Webtext Corpora:A Case Study on the Colossal Clean Crawled Corpus》Dodge等，广发证券发展研究中心

传统内容从业者面对 AI 浪潮，不仅担心作品被未经授权使用，还忧虑新技术威胁其职业稳定性。对于作家等创作者来说，生成式 AI 未经授权使用其作品会造成版税损失，而 AI 模仿、总结或解释其作品而产生的衍生作品，对作品市场的侵占同样会威胁其生计。而对于新闻媒体来说，常见的流量转化逻辑是，通过优化使其内容在搜索引擎中排名靠前以获取更多流量，即 SEO（搜索引擎优化）。然而，聊天类 AI 绕过了“检索”步骤，让用户直接获取新闻内容，而无需在乎报道者。尽管 AI 可以引用报道链接，甚至在正文中标明“根据某媒体报道”，但用户通常不会点击这些链接，这对媒体导流作用明显弊大于利。因此，在聊天机器人中使用新闻媒体的内容可能会转移读者，从而减少订阅、广告、许可和附属收入。

我们总结，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。

（二）授权合作，内容持有者的新道路

面对生成式 AI 的发展浪潮，部分内容持有者选择抵制 AI 公司并控诉其侵权行为，但同时另一部分则走上了授权合作道路。

对于内容持有者来说，授权合作可以带来与诉讼和解相当甚至更多的现金收益，而且速度更快，同时这些交易还有助于其将 AI 应用于业务优化；对于 AI 公司，通过与内容持有者合作，不仅能获取高质量的新闻数据以改进模型效果，还能确保数据的合法来源，避免侵犯版权。因此，这种合作对双方皆有益。目前，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。

OpenAI 是与内容持有者合作最频繁的 AI 厂商之一，动作主要集中在新闻行业。因为新闻数据具有及时性、真实性和广泛覆盖面，可以帮助解决 AI 内容生成过程中的幻觉问题，但同时又存在版权方面的挑战。因此，OpenAI 希望通过与新闻出版商的合作来获取高质量的新闻数据改进模型，并确保数据的合法来源。目前，OpenAI 已签署大约十几个出版商协议，并且有许多协议正在进行中，部分合作案例包括：

1. 2023 年 7 月，OpenAI 获得美联社授权使用其新闻故事存档来训练 AI 模型，但是这项合作未授权在 ChatGPT 的用户输出中使用美联社的内容。
2. 2023 年 12 月，OpenAI 与全球新闻出版商 Axel Springer 达成内容使用协议。OpenAI 在未来三年内将支付数千万欧元以使用其出版物内容，ChatGPT 将以摘要形式生成答案，并在答案下方包含指向信息原始来源的链接。
3. 2024 年 4 月，OpenAI 与《金融时报》达成战略合作，并签署内容许可协议。通过此次合作，ChatGPT 用户将能够看到《金融时报》精选摘要、引述以及新闻报道的链接，并回应相关查询。
4. 2024 年 5 月，OpenAI 与全球型媒体“新闻集团”达成合作，将使用其旗下的媒体数据来训练和增强 ChatGPT 模型。
5. 2024 年 5 月，OpenAI 宣布与社交媒体网站 Reddit 建立合作关系，利用 Reddit 上的内容训练 ChatGPT 模型。
6. 2024 年 6 月，OpenAI 与《时代》杂志签署了一份多年期内容协议，该协议允许 OpenAI 访问《时代》杂志的新闻内容档案。

此外，苹果、谷歌等科技巨头也纷纷与众多新闻媒体、视觉内容持有者、社交平台等开展合作。

1. 2023 年 12 月，苹果和 NBC 新闻、康泰纳仕和 IAC 等新闻机构达成了至少 5000 万美元的多年合约。
2. 2024 年 4 月，苹果与 Shutterstock 达成一笔价值 5000 万美元的协议，苹果将从 Shutterstock 手中获得数百万张图片用于训练其 AI 模型。
3. 2024 年 2 月，谷歌与 Reddit 达成每年 6000 万美元的协议，允许谷歌使用 Reddit 的数据来训练其 AI 系统。

表 6：AI 公司与内容持有方的授权合作案例

AI 公司	内容持有方	内容方性质	协议形式与金额	授权内容	授权时间
OpenAI	Shutterstock	视觉内容持有者	金额未公布	训练	2021 年开始，2023 年宣布合作再延长六年
OpenAI	美联社	新闻媒体	金额未公布	部分训练	2023/7
OpenAI	Axel Springer	新闻媒体	未来三年内支付数千万欧元	训练、展示	2023/12
OpenAI	Le Monde	新闻媒体	长期协议，金额未公布	训练、展示	2024/3
OpenAI	Prisa	新闻媒体	金额未公布	训练、展示	2024/3
OpenAI	金融时报	新闻媒体	500-1000 万美元	部分训练、展示	2024/4
OpenAI	Dotdash Meredith	出版商	金额未公布	训练、展示	2024/5
OpenAI	Reddit	社交平台	6000 万美元	训练、展示	2024/5
OpenAI	新闻集团	新闻媒体	5 年协议，超过 2.5 亿美元	训练、展示	2024/5
OpenAI	Vox Media	新闻媒体	金额未公布	训练	2024/5
OpenAI	The Atlantic	新闻媒体	金额未公布	训练、展示	2024/5
OpenAI	Times	新闻媒体	多年期，未披露金额	训练、展示	2024/6
苹果	康泰纳仕，NBC 新闻，IAC	出版商、新闻媒体	5000 万美元的多年期合同	训练	2023/12
苹果	Shutterstock	视觉内容持有者	5000 万美元	训练	2024/4
谷歌	Reddit	社交平台	6000 万美元/年	训练、展示	2024/2

数据来源：Shutterstock，Bloomberge，The Verge，Axios，Lemond，Prisa，Dotdash Meredith，OpenAI，Mashable，Investopedia，News Cop，Vox Media，The Atlantic，Reuters，NewYorkTimes，BusinessInsider，广发证券发展研究中心

从行业属性来看，文本类数据集授权目前以新闻媒体为主，以及 Reddit 论坛等数据，但是书籍期刊的进展较为缓慢。

我们分析，文本类数据集出现此种趋势的原因如下：首先，新闻内容的生命周期相对较短，难以保持独特性，通常在发布后的短时间内价值达到顶峰，但随着时间推移，价值迅速下降。因此，新闻出版商更愿意在高价值时期内获取更多收益；其次，大部分训练数据并不能捕捉到人类的日常用语，而 Reddit 等论坛则提供了丰富的未过滤口语写作数据，这是 AI 公司能找到的较为有价值的日常表达数据集，能够有效提升模型训练效果；相比之下，期刊书籍等学术出版物通常具有较长的生命周期，在数年甚至数十年内仍然具有较高的研究和参考价值，因此，学术出版商更倾向于长期控制其内容，以维持长久价值，这种考虑导致其授权合作进展较为缓慢。

表 7：不同行业属性文本类数据集比较

	新闻媒体	论坛	书籍期刊
内容生命周期	短，随着时间推移价值会迅速下降	不定，讨论内容可能随时被更新，但某些热门话题会长期保持活跃	长，在数年甚至数十年内仍具有较高研究和参考价值
内容需求	信息传播为主	实时讨论和问题解答，反映人们的真实语言使用和思维过程	用于研究和教育，使用频率相对较低且高度专业化
商业模式	订阅；媒体广告收入	主要依赖广告和会员捐赠	依赖于订阅、会员费和销售
版权归属	通常属于出版商，涉及的法律与版权问题相对较少	版权通常属于发布者和平台，涉及隐私问题	版权可能涉及多个作者和机构

数据来源：广发证券发展研究中心

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。起初，生成式 AI 领域的数据授权合作多集中在文本领域，伴随着多模态大模型的发展，数据授权合作由于训练与应用需求而拓展至图像和音视频等领域。以全球创意平台 Shutterstock 为例，其通过向 AI 公司提供庞大的视觉媒体库，将视觉内容货币化并进行了商业模式的重大转变。Shutterstock 不仅与 Meta、Alphabet、亚马逊和苹果等“主力客户”达成协议，还与 OpenAI 签署了一份为期六年的协议。2023年，Shutterstock 通过与 AI 公司的授权业务创收达 1.04 亿美元。

关于授权的定价方式，目前以直接订阅收费为主。例如图像网站 Photobucket 目前正在与多家科技公司进行商谈，计划将其平台上的 130 亿张照片和视频资源授权给这些公司，用于训练生成式 AI 模型，其中每张照片价值 5 美分到 1 美元，每个视频价值超过 1 美元，具体定价取决于买家和素材种类；AI 数据定制公司 Defined.ai 文本价格为每字 0.001 美元，而一张图片卖 1 到 2 美元，一部短视频卖 2 到 4 美元，一部长片每小时则可以卖到 100 到 300 美元。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来进行授权定价。

未来定价模式可能更多基于内容对 AI 模型的贡献。在当前大模型的发展背景下，内容的价值评估标准正在发生变化。在搜索引擎时代，数字版权的定价模式主要是基于内容的受欢迎程度和流量，广告商愿意为高流量的内容付费。然而，在大模型时代，数据作为一种新的资产类别，AI 公司利用大规模数据进行模型训练和优化，相比于过去单纯依靠广告收入的模式，现在内容的价值更多体现在其对 AI 模型的贡献上。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。因此，我们判断，未来的内容授权合作的定价模式可能更多基于对 AI 模型的贡献，这种以大模型公司的盈利方式来定价较为合理，不仅反映了内容的实际价值，还能促进版权方和 AI 公司之间的合作，共同推动技术进步和商业模式创新。

（三）诉讼或合作？内容持有者面临的选择、机会与挑战

结合对于以上诉讼与授权合作案例的讨论分析，我们有如下发现：

1. 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。我们发现，艺术家们普遍倾向于抵制 AI 公司并控诉其侵权行为，而新闻媒体在版权保护的斗争中却难以形成统一阵线。《金融时报》、美联社和 Axel Springer 等新闻媒体选择与 AI 公司合作，签订付费协议，而《纽约时报》和一些地区性报纸等则选择抵制与诉讼，这种选择差异导致新闻行业在版权保护上的一致行动受到削弱。我们认为，分歧由以下几点原因造成：

(1) 商业模式：艺术家们主要依赖于作品销售和版权使用费，因此，未经授权使用这些内容直接威胁到其经济利益，使得他们更有动机采取法律行动来抵制侵权行为。相比之下，新闻媒体的商业模式发生了较大变化，广告收入减少和在线免费新闻的普及使得新闻媒体在版权保护问题上难以达成一致，无法形成统一的阵线对抗 AI 公司的侵权行为。因为广告依赖型的新闻媒体担心 AI 技术会减少访问量，而依赖授权收入的媒体则可能更愿意与 AI 公司合作。

(2) 内容独特性：音乐绘画等艺术作品通常属于独特的内容创作，具有较高的艺术和商业价值，因此需要较强的版权保护，同时也常拥有强大的版权集体管理组织来保护其利益。而新闻内容独特性较低，同一事件可能由不同媒体报道，削弱了新闻媒体在版权保护上的一致性和力度。此外，新闻内容的生命周期较短，时效性强，且报道往往涉及公共信息，增加了版权保护的复杂性。

(3) 行业结构：音乐等行业的行业格局相对集中，少数大公司主导，这些公司有资源和动机共同抵制 AI 公司的侵权行为。相比之下，新闻行业较为分散，不同媒体公司之间存在激烈竞争，难以形成统一战线。不同新闻媒体的立场和利益可能有所不同，导致在对抗 AI 公司的侵权行为上的一致行动受到削弱。

2. 内容持有者面临的商业机会

合作授权一般不具有排他性，同一数据集可被用于训练多个模型。除非是通过公司的并购交易等方式进行授权或直接买断，内容持有者授权一般不具有排他性。例如，Reddit 同时与 Open AI、谷歌等签订了内容授权协议，而 Shutterstock 也将其图片数据授予给 OpenAI、苹果等多家公司用于训练 AI 模型。

内容的稀缺性可以增强内容持有者在交易中的议价能力。由于新闻媒体难以在版权保护的斗争中形成统一阵线，这种分歧削弱了新闻行业在与 AI 公司谈判时的议价能力。但对于拥有数十年内容的老牌新闻公司，其所拥有的新闻档案对于大模型公司来说可能非常有价值。同时，新兴的新闻公司，若能提供实时数据与见解，也能吸引 AI 厂商的兴趣。此外，视频档案通常比文本数据提供了更多差异化信息，广

播公司和有线网络等拥有大量视频档案的内容所有者同样能够增强其在谈判中的议价能力。

3. 内容持有者面临的挑战

内容持有者可能会面临两难局面。AI 公司训练大模型所需的数据类型与数量有所差异，可能无需从所有内容持有者那里获得许可。对于内容持有者而言，最好的集体结果是抵制授权其内容并将价值保留在其平台内。然而，若不能与 AI 厂商达成协议，便有可能出局，对被拒之门外的恐惧可能会迫使部分内容持有者授权其内容，甚至不断降低授权价格，并开始恶性循环，因此内容持有者将会面临两难局面。此外，起诉的高成本也可能会给内容持有者造成压力，迫使其考虑和解。

内容持有者面临的另一挑战是量化其内容的商业价值。由于缺乏统一的标准和透明的评估机制，内容持有者在与 AI 公司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，AI 公司对内容需求的多样性和动态变化使得内容持有者在确定内容价值时面临更多不确定性。这种量化挑战迫使内容持有者需要在创新商业模式和保护自身利益之间找到平衡。

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。当 AI 公司在训练模型时使用了未经授权的受版权保护内容，可能导致生成内容的侵权，并让内容提供者面临法律诉讼的风险。因此，尽管是 AI 公司进行模型训练，但内容持有者可能因提供了这些数据而被卷入法律纠纷，被指控间接侵犯版权。

三、AI 巨头将持续加码数据合作，需关注数据版权纠纷重点案例

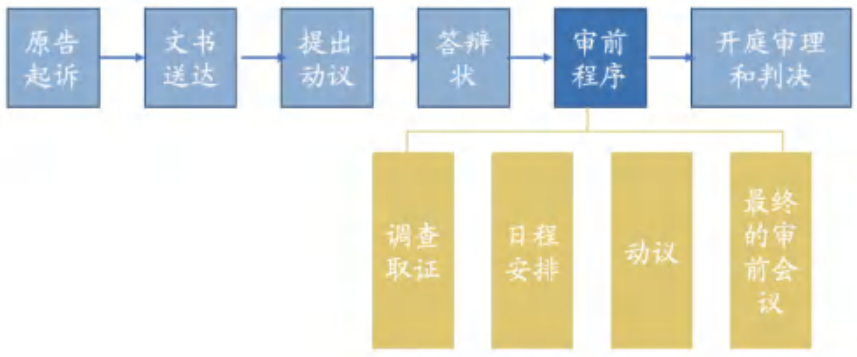
确保训练数据的合法来源与合理使用对于 AIGC 的发展非常关键。在 AIGC 的发展过程中，大模型训练面临着数据使用的合法性问题，例如未经授权使用数据进行训练将可能造成侵权行为。若在初期阶段未能妥善解决，可能会导致一系列法律纠纷与侵权指控，不仅会对模型的合法合规产生影响，还会给研发带来较大的法律风险和不确定性，甚至可能阻碍整个项目的进展。因此，确保训练数据的合法来源和合理使用对于 AIGC 的发展非常关键。只有解决了这一问题，才能在确保法律合规的前提下，推动生成式 AI 的广泛应用与商业落地。

我们对于 AI 数据版权诉讼与内容合作的时间节奏进行梳理，发现从 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。

（一）数据版权纠纷尚无判例，需关注重点案例

关于 AI 训练数据版权诉讼，国外尚未达成判例。最早的 AI 训练数据版权诉讼可以追溯到 2020 年，然而，各方在诉讼程序中反复拉锯，目前尚未宣判。OpenAI、微软和《纽约时报》之间的诉讼尚处于审前程序阶段，2023 年 12 月 27 日，《纽约时报》提交针对微软和 OpenAI 的诉讼，2024 年 7 月 1 日，OpenAI 要求《纽约时报》提供证据证明受版权保护的作品是作者的原创作品。如果在审前程序阶段没有和解或其它解决办法，那么在审前程序阶段完成后，案件将进入庭审阶段并进行最终庭审和判决。根据目前进展，该案件可能会持续数月甚至数年，具体时间取决于案件的复杂性、法庭日程安排以及双方的合作程度与法律策略。

图 12：美国民事诉讼流程



数据来源：法保网，广发证券发展研究中心

表 8：纽约时报与 OpenAI、微软的诉讼时间轴

日期	事件
2023 年 12 月 27 日	《纽约时报》提交针对微软和 OpenAI 的投诉
2023 年 12 月 27 日	原告提交相关案件声明
2024 年 1 月 2 日	《纽约时报》和作者公会的案件相关
2024 年 2 月 23 日	原告提交介入动议作为首起集体诉讼原告
2024 年 2 月 26 日	微软提交反对介入动议的意见，OpenAI 提交对介入动议的回复，OpenAI 提交驳回动议，OpenAI 请求进行口头辩论
2024 年 3 月 4 日	微软提交驳回动议，微软请求进行口头辩论
2024 年 3 月 8 日	原告提交发现计划报告
2024 年 3 月 11 日	原告提交反对 OpenAI 驳回动议的动议，原告请求对 OpenAI 部分驳回动议进行口头辩论
2024 年 3 月 18 日	《纽约时报》提交反对微软部分驳回动议，OpenAI 提交支持其驳回动议的回复
2024 年 3 月 25 日	微软提交支持其部分驳回动议的回复
2024 年 4 月 1 日	南区法院驳回 Tremblay 原告的介入动议
2024 年 4 月 15 日	Tremblay 原告提交上诉通知
2024 年 6 月 3 日	OpenAI 提交有条件反对《纽约时报》修正投诉申请的动议
2024 年 6 月 10 日	《纽约时报》提交支持其修正投诉申请的回复
2024 年 6 月 13 日	OpenAI 提交合并《纽约时报》和《每日新闻》案件的动议
2024 年 6 月 14 日	微软提交加入支持 OpenAI 的合并动议
2024 年 7 月 1 日	OpenAI 要求《纽约时报》提供证据证明受版权保护的作品是作者的原创作品

数据来源：BakerHostetler，美国纽约南区地方法院文件，广发证券发展研究中心

国内针对关于 AI 生成内容的版权纠纷已有相关判例，但聚焦于生成内容被侵权而非训练数据侵权。2023 年 5 月，原告使用 Stable Diffusion 模型生成的图片未经许可被发布后，将被告起诉至北京互联网法院。最终法院认定原告的图片具备“独创性”，符合作品的定义，受到著作权法保护，这也成为了国内首例 AI 生成图片著作权侵权案。然而关于训练数据未经授权的问题，国内相关案件同样未做出生效判决。例如，2023 年 6 月，北京笔神作文公司宣布起诉其合作伙伴学而思，指控其通过“爬虫”技术，非法访问、缓存其服务器数据多达 258 万次，以此开发 MathGPT 的新产品“作文 AI 助手”。笔神作文公司要求学而思其公开道歉、删除数据资源并求偿 1 元，但目前双方已达成和解。

版权法的复杂性与模糊性导致 AI 数据版权纠纷的判决尚需时间。2005 年，美国作家协会通过集体诉讼向法院起诉谷歌图书项目侵犯版权，该诉讼历时十一年。法官最后判定，谷歌图书馆计划中的作品利用属于合理使用，仅复制了作品的一小部

分并改变了原作，是转换性使用，因此不构成侵权。我们认为，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。在《纽约时报》诉 OpenAI 案例中，由于原被告双方分别为美国老牌媒体与生成式 AI 巨头，其诉讼会深度触及 AI 数据训练以及生成内容与训练素材关系的合法性判断，不但会影响 AIGC 文字内容的版权纠纷，也会对 Midiourney 和 Stable·Diffusion 一众图形生成 AI 的版权诉讼产生明显影响。一旦相关判决落地，将成为里程碑式的案件，对以后 AI 训练数据的版权界定有较大参考意义。因此建议关注《纽约时报》诉 OpenAI 等重点案例，案件判决将厘清新兴法律轮廓，引导 AI 技术与版权法之间的关系发展，并成为生成式 AI 技术历史上的标志性事件之一。

有望在今年内初步了解法院对于此类训练数据版权诉讼请求的态度。汤森路透公司曾在 2020 年对 Ross Intelligence 提起诉讼，指控其使用 Westlaw 案例摘要作为其 AI 系统分析法律问题的训练数据，但 Ross Intelligence 认为其自身只是合理使用了这些案例摘要。法官驳回了诉讼双方的简易判决动议，认为在侵权和合理使用的指控上存在有争议的事实问题。该诉讼预计在 2024 年 8 月 26 日进行庭审，我们预计本次庭审将初步揭示法院对于这类训练数据版权诉讼请求的态度。

（二）AI 巨头将持续加码数据合作，确保数据的合法来源

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。根据相关文献，对 Wikipedia、Common Crawl 和 WebText 数据集组合的模型中主要资源或域名的未加权 Token 总数进行排名。我们发现在排名前 50 的域中，有 17 家新闻媒体类（4 家重复）公司，其中不重复的 13 家中，目前有 3 家正在与 AI 公司合作（Reuters、Business Insider 和 The Atlantic），2 家正在进行诉讼（The NY Times 和 Chicago Tribune），2 家正在谈判（BBC 和 CNN），还有 6 家未表明立场。从 2023 年下半年开始，越来越多的公司开始明确其立场。

表 9：混合的文本数据集前 50 个域排名

排名	来源 / 域名	数据集	未加权的 Tokens 数量 (M)	域名分类	归属	目前状态
1	Biography	Wikipedia	834	Wiki	Wikipedia	
2	Google Patents	Common Crawl	750	专利类	Google	
3	Geography	Wikipedia	531	Wiki	Wikipedia	
4	Google	WebText	514	-	Google	
5	Culture and Arts	Wikipedia	474	Wiki	Wikipedia	
6	History	Wikipedia	297	Wiki	Wikipedia	

7	Biology, Health, and Medicine	Wikipedia	234	Wiki	Wikipedia	
8	Archive	WebText	199	学术类	Archive	
9	Sports	Wikipedia	195	Wiki	Wikipedia	
10	Blogspot	WebText	152	博客类	Google	
11	Business	Wikipedia	144	Wiki	Wikipedia	
12	GitHub	WebText	138	代码类	Microsoft	
13	Other society	Wikipedia	132	Wiki	Wikipedia	
14	The NY Times	WebText	111	新闻类	The New York Times Company	诉讼
15	WordPress	WebText	107	博客类	Automattic	
16	Science & Math	Wikipedia	105	Wiki	Wikipedia	
17	WashingtonPost	WebText	105	新闻类	Nash Holdings	-
18	Wikia	WebText	104	社区类	Fandom, Inc.	
19	BBC	WebText	104	新闻类	BBC	谈判
20	The NY Times	Common Crawl	100	新闻类	The New York Times Company	诉讼
21	Los Angeles Times	Common Crawl	90	新闻类	Nant Capital	-
22	The Guardian	Common Crawl	90	新闻类	Scott Trust	-
23	PLoS	Common Crawl	90	学术类	Public Library of Science	
24	TheGuardian	WebText	82	新闻类	Scott Trust	-
25	Forbes	Common Crawl	80	新闻类	Integrated Whale Media Investments	-
26	HuffingtonPost	Common Crawl	75	新闻类	BuzzFeed	-
27	Patents.com	Common Crawl	71	专利类	Patents.com	
28	Scribd	Common Crawl	70	社区类	Scribd, Inc.	
29	eBay	WebText	70	电商类	eBay	
30	Pastebin	WebText	70	代码类	Pastebin.com	
31	CNN	WebText	66	新闻类	Warner Bros. Discovery	谈判
32	Washington Post	Common Crawl	65	新闻类	Nash Holdings	-
33	Yahoo	WebText	65	信息类	Apollo Global Management	
34	HuffingtonPost	WebText	62	新闻类	BuzzFeed	-
35	Go	WebText	62	搜索引擎类	Google	
36	The Motley Fool	Common Crawl	61	财经类	The Motley Fool	
37	Reuters	WebText	61	新闻类	Thomson Reuters	合作（交易方未公布）
38	IMDb	WebText	61	信息类	Amazon	
39	IPFS	Common Crawl	60	技术类	Protocol Labs	
40	Frontiers Media	Common Crawl	60	学术类	Frontiers	
41	Business Insider	Common Crawl	60	新闻类	Axel Springer	与 OpenAI 合作
42	Chicago Tribune	Common Crawl	59	新闻类	Tribune Publishing	诉讼
43	Booking.com	Common Crawl	58	旅游类	Booking Holdings	
44	The Atlantic	Common Crawl	57	新闻类	Emerson Collective	与 OpenAI 合作

45	Springer Link	Common Crawl	56	学术类	Springer Nature	
46	Al Jazeera	Common Crawl	55	新闻类	Al Jazeera Media Network	-
47	Kickstarter	Common Crawl	54	众筹类	Kickstarter, PBC	
48	Goo	WebText	54	搜索引擎类	NTT Resonant Inc.	
49	FindLaw Caselaw	Common Crawl	53	法律类	Thomson Reuters	-
50	NCBI	Common Crawl	53	学术类	National Center for Biotechnology Information	-

数据来源：《What's in my AI》Alan D. Thompson，Financial Times，Bloomberg，广发证券发展研究中心

注：绿色底色为新闻媒体类公司，显示的排名是基于数据集中未加权的 Token 总数，有些资源存在重复（例如《纽约时报》既出现在 WebText 数据集，也出现在过滤后的 Common Crawl 数据集）。

2024 年有望成为 AI 数据版权之争的关键年。从当前的法律诉讼和合作谈判情况来看，内容持有者已在积极采取行动保护其版权，并通过诉讼或合作的方式来处理与 AI 公司的关系。目前趋势与数据表明，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在这一年内明确其立场和策略，进一步推动版权争议的解决。

我们判断未来授权合作或快于法律变革与监管介入。根据目前趋势，内容持有者和 AI 公司之间的合作与谈判将持续进行，预计在未来一两年内能出现成果，特别是如果双方都能达成满意协议的情况下，速度将更快；相比之下，版权诉讼尚未达成判例，考虑到当前诉讼数量与复杂性，可能需要两三年才能形成明确的法律框架；政府和国际组织等监管机构同样可能介入并制定新的标准，但这通常需要数年时间，如果紧迫性增加，可能会在未来三五年内看到明显进展。综合来看，全面解决 AI 训练数据版权之争可能需要三到五年。然而，在这一过程中可能会有阶段性的进展和部分解决方案逐步出台。

具体节奏方面，我们预计，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，在诉讼过程中也很有可能出现和解的情况，推动诉讼和合作并行发展。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，合作模式的可行性和有效性将得到初步验证，合作的定价模式也将逐渐明确。随着更多案件进入判决阶段，预计将出台多个具有代表性的法律判例，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。

四、投资建议

基于上述分析，我们认为，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，微软、谷歌、脸书等头部公司将因此受益。这些公司具备强大的资金实力和资源获取能力，能够负担起训练数据成本，从而在数据竞争中占据优势，并在大模型研发和应用中保持领先。

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，AIGC 应用的发展也将进一步加速。值得注意的是，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，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，创意工具软件建议关注万兴科技（300624.SZ）、美图公司（01357.HK，广发传媒覆盖）等；办公软件建议关注金山办公（688111.SH）等。

海外市场方面，建议关注以下应用方向：

1. 创意工具软件（如 Adobe ），将 AI 解决方案深度融入应用程序及生态中，赋能生产力变革，重塑创意工作流；
2. AI 解决方案（如 Palantir 、C3.ai ），通过提供定制化 AI 解决方案，帮助企业应对复杂的数据分析和业务决策挑战，受益于企业与公共部门对 AI 解决方案需求的增加而增长；
3. 企业服务软件（如 Salesforce、ServiceNow ），基于 AI 技术提升产品力，从而满足客户多样化需求；
4. 数据管理与服务软件（如 MongoDB、Datadog ），提供基于云的数据库、数据分析和监控解决方案等，AI 有望成为其长期增长驱动力；
5. 以及 AI 音乐创作服务、AI 搜索、AI 翻译、协同 OA 等一些 AI 初创公司所处的其它赛道。

表 10：部分海外 AI 初创公司主营与融资信息

公司名称	主营信息	当前融资轮次	当前估值
Suno AI	**AI 音乐创作服务提供商**，允许通过输入文本提示或歌词来创作原创歌曲，AI 则会根据此生成旋律、和声和完整作曲。	B 轮	未披露
Perplexity AI	**智能对话式搜索引擎提供商，**专注于开发基于 AI 聊天的对话搜索引擎，使用大语言模型（OpenAI API）和搜索引擎，提供问题答案。	B+轮	10 亿美元
DeepL	**AI 翻译软件提供商，**专注于为用户提供以欧洲语系为主的即时翻译服务，主要为企业提供翻译服务，强项是与中小型组织合作。	C 轮	未披露
Notion	**协同 OA 软件提供商，**集便笺、任务、Wiki 和数据库于一体，对于同一份数据支持多种视图的切换和展示，拥有文档编辑与管理、个人及团队知识库、项目协作以及多端支持等功能。	C 轮	100 亿美元

Scale AI	**AI 数据平台，**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，助力 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。	F 轮	138 亿美元

数据来源：烯牛数据、广发证券发展研究中心

此外，值得注意的是，虽然内容持有者可以通过直接收费、分享收益、标注出处等方式进行授权定价，但 AI 厂商可以与多家内容持有者合作。因此，内容持有者在定价与谈判中面临一定挑战，需要在维持收益和吸引 AI 厂商之间寻求平衡点，这使得训练数据授权合作对于单个内容持有者的收入贡献较为有限。

五、风险提示

（一）内容价值难以准确量化

内容持有者因缺乏统一标准和透明评估机制，以及 AI 公司需求的多样性和动态变化，难以准确量化其内容的价值，从而在谈判中处于不利地位。

（二）行业竞争加剧

AI 厂商具备合作灵活性，内容持有者在定价与谈判中面临挑战，一旦行业竞争加剧，对被拒之门外的恐惧可能会迫使部分内容持有者授权其内容，甚至不断降低授权价格，并开始恶性循环。

（三）数据侵权阻碍下游应用进展

若数据诉讼或者合作节奏不及预期，可能引起侵权指控，并给研发团队带来较大法律风险和不确定性，下游应用的开发和推广将受到阻碍，导致商业落地困难。

AI人工智能产业链联盟

#每日为你摘取最重要的商业新闻#

更新 · 更快 · 更精彩

Zero

AI音乐创作人
水墨动漫联盟创始人
百脑共创联合创始人
人工智能产业链联盟创始人
中关村人才协会秘书长助理
河北北大企业家分会秘书长
墨玫星辰智能科技有限公司CEO
河北清华发展研究院智能机器人中心线上负责人
中关村人才协会数字体育与电子竞技专委会秘书长助理

主要业务:AI商业化答疑及课程应用场景探索，各类AI产品学习手册，答疑及课程

欢迎扫码交流

提供：学习手册/工具/资源链接/商业化案例/行业报告/行业最新资讯及动态

人工智能产业链联盟创始人

邀请你加入星球，一起学习

人工智能产业链联盟报告库

星主：人工智能产业链联盟创始人

每天仅需0.5元，即可拥有以下福利！
每周更新各类机构的最新研究成果。立志将人工智能产业链联盟打造成市面上最全的AI研究资料库，覆盖券商、产业公司、研究院所等...

知识星球

微信扫码加入星球

广发计算机行业研究小组

刘 雪 峰：首席分析师，东南大学工学士，中国人民大学经济学硕士，1997 年起先后在数家 IT 行业跨国公司从事技术、运营与全球项目管理工作。2010 年就职于招商证券研究发展中心负责计算机组行业研究工作，2014 年加入广发证券发展研究中心。

吴 祖 鹏：资深分析师，中南大学材料工程学士，复旦大学经济学硕士，曾先后任职于华泰证券、华西证券，2021 年加入广发证券发展研究中心。

李 婉 云：资深分析师，西南财经大学金融学硕士，2022 年加入广发证券发展研究中心。

周 源：资深分析师，慕尼黑工业大学硕士，2021 年加入广发证券，曾任职于 TUMCREATE 自动驾驶科技公司，负责大数据相关工作。

许 晟 榕：高级研究员，香港大学金融科技硕士，2023 年加入广发证券发展研究中心。

王 钰 翔：研究员，哥伦比亚大学运筹学硕士，2024 年加入广发证券发展研究中心。

戴 亚 敏：研究员，北京大学金融硕士，2024 年加入广发证券发展研究中心。

广发证券—行业投资评级说明

买入：预期未来 12 个月内，股价表现强于大盘 10%以上。

持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-10%～+10%。

卖出：预期未来 12 个月内，股价表现弱于大盘 10%以上。

广发证券—公司投资评级说明

买入：预期未来 12 个月内，股价表现强于大盘 15%以上。

增持：预期未来 12 个月内，股价表现强于大盘 5%-15%。

持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-5%～+5%。

卖出：预期未来 12 个月内，股价表现弱于大盘 5%以上。

联系我们

	广州市	深圳市	北京市	上海市	香港
地址	广州市天河区马场路 26 号广发证券大厦 47 楼	深圳市福田区益田路 6001 号太平金融大厦 31 层	北京市西城区月坛北街 2 号月坛大厦 18 层	上海市浦东新区南泉北路 429 号泰康保险大厦 37 楼	香港湾仔骆克道 81 号广发大厦 27 楼
邮政编码	510627	518026	100045	200120	-
客服邮箱	gfzqyf@gf.com.cn				

法律主体声明

本报告由广发证券股份有限公司或其关联机构制作，广发证券股份有限公司及其关联机构以下统称为“广发证券”。本报告的分销依据不同国家、地区的法律、法规和监管要求由广发证券于该国家或地区的具有相关合法合规经营资质的子公司/经营机构完成。

广发证券股份有限公司具备中国证监会批复的证券投资咨询业务资格，接受中国证监会监管，负责本报告于中国（港澳台地区除外）的分销。

广发证券（香港）经纪有限公司具备香港证监会批复的就证券提供意见（4 号牌照）的牌照，接受香港证监会监管，负责本报告于中国香港地区的分销。

本报告署名研究人员所持中国证券业协会注册分析师资质信息和香港证监会批复的牌照信息已于署名研究人员姓名处披露。

重要声明

广发证券股份有限公司及其关联机构可能与本报告中提及的公司寻求或正在建立业务关系，因此，投资者应当考虑广发证券股份有限公司及其关联机构因可能存在的潜在利益冲突而对本报告的独立性产生影响。投资者不应仅依据本报告内容作出任何投资决策。投资者应自主作出投资决策并自行承担投资风险，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者口头承诺均为无效。

本报告署名研究人员、联系人（以下均简称“研究人员”）针对本报告中相关公司或证券的研究分析内容，在此声明：（1）本报告的全部分析结论、研究观点均精确反映研究人员于本报告发出当日的关于相关公司或证券的所有个人观点，并不代表广发证券的立场；（2）研究人员的部分或全部的报酬无论在过去、现在还是将来均不会与本报告所述特定分析结论、研究观点具有直接或间接的联系。

研究人员制作本报告的报酬标准依据研究质量、客户评价、工作量等多种因素确定，其影响因素亦包括广发证券的整体经营收入，该等经营收入部分来源于广发证券的投资银行类业务。

本报告仅面向经广发证券授权使用的客户/特定合作机构发送，不对外公开发布，只有接收人才可以使用，且对于接收人而言具有保密义务。广发证券并不因相关人员通过其他途径收到或阅读本报告而视其为广发证券的客户。在特定国家或地区传播或者发布本报告可能违反当地法律，广发证券并未采取任何行动以允许于该等国家或地区传播或者分销本报告。

本报告所提及证券可能不被允许在某些国家或地区内出售。请注意，投资涉及风险，证券价格可能会波动，因此投资回报可能会有所变化，过去的业绩并不保证未来的表现。本报告的内容、观点或建议并未考虑任何个别客户的具体投资目标、财务状况和特殊需求，不应被视为对特定客户关于特定证券或金融工具的投资建议。本报告发送给某客户是基于该客户被认为有能力独立评估投资风险、独立行使投资决策并独立承担相应风险。

本报告所载资料的来源及观点的出处皆被广发证券认为可靠，但广发证券不对其准确性、完整性做出任何保证。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价。广发证券不对因使用本报告的内容而引致的损失承担任何责任，除非法律法规有明确规定。客户不应以本报告取代其独立判断或仅根据本报告做出决策，如有需要，应先咨询专业意见。

广发证券可发出其它与本报告所载信息不一致及有不同结论的报告。本报告反映研究人员的不同观点、见解及分析方法，并不代表广发证券的立场。广发证券的销售人员、交易员或其他专业人士可能以书面或口头形式，向其客户或自营交易部门提供与本报告观点相反的市场评论或交易策略，广发证券的自营交易部门亦可能会有与本报告观点不一致，甚至相反的投资策略。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且无需另行通告。广发证券或其证券研究报告业务的相关董事、高级职员、分析师和员工可能拥有本报告所提及证券的权益。在阅读本报告时，收件人应了解相关的权益披露（若有）。

本研究报告可能包括和/或描述/呈列期货合约价格的事实历史信息（“信息”）。请注意此信息仅供用作组成我们的研究方法/分析中的部分论点/依据/证据，以支持我们对所述相关行业/公司的观点的结论。在任何情况下，它并不（明示或暗示）与香港证监会第 5 类受规管活动（就期货合约提供意见）有关联或构成此活动。

权益披露

(1)广发证券（香港）跟本研究报告所述公司在过去 12 个月内并没有任何投资银行业务的关系。

版权声明