2023 年 12 月 25 日

光大证券 EVERBRIGHT SECURITIES

行业研究

梳理全球 AIGC 数据版权规范，哪些领域具备商业化潜力？

——AI 产业前瞻系列报告（二）

要点

用于 AI 模型训练的数据集有多种方式避免侵权，或直接补偿版权持有人。

专有数据：1）版权合作协议：Shutterstock、Axel Springer 等多家版权提供商与 AI 公司建立合作；**2）API 付费访问：**部分专业性强的数据提供商会对 API 访问进行收费，23 年 Reddit、Twitter 的 API 访问由免费转向付费。

开源数据：1）开放许可协议：包括 CC、ODC、CDLA 等；**2）特定的数据抓取策略：**如遵守网站的 Robot.txt 文件；**3）社区监督：**提升数据集透明度。

直接补偿创作者：1）事前补偿：技术难度低，但难以界定合理的补偿额度，如 Shutterstock 建立的贡献者基金；**2）事后补偿：**对 AI 生成内容进行溯源，定价合理但技术尚不成熟，如卡耐基梅隆大学发表的归因模型算法。

专用数据集：直接出售适用于模型训练的数据集，或打包成 MaaS 服务。

海外版权合作协议盈利模式稳定、商业化前景初步展现。AI 生成内容或对版权提供商的传统业务造成一定威胁，版权提供商与 AI 公司的合作是互利共赢。**1）多媒体素材库 Shutterstock：**通过出售模型训练素材创收，推出 AI 生成图片专区，提供由 OpenAI 支持的 AI 工具；**2）出版商 Axel Springer：**向 OpenAI 出售其出版物作为训练素材，共同运用 AI 技术提升用户体验。

从 Shutterstock 看版权库与 AI 公司的合作：AIGC 的利好整体强于利空。
1）利好：Shutterstock 的数据授权收入已较明显体现在业绩端，驱动估值修复和股价回升，23Q3 出售模型训练素材的收入占公司总收入的 19.5%；**2）利空：**23 年以来 Shutterstock 传统业务低迷更多受同业竞争影响，同类公司 Getty Image 业绩稳健，AIGC 对版权库行业的威胁和替代尚不明显。

国内外模型训练数据版权规定尚待完善，版权商股价有望得到密集催化。22M12 一篇论文显示 Stable Diffusion 以像素点级别复制名画的细节。**对 AIGC 的版权问题的争议和相关法规主要可以分为两类：1）AI 生成内容的版权界定：**美国不承认 AI 生成内容拥有著作权，而中国倾向于保护 AI 生成内容的著作权；**2）模型训练数据的版权规定：**美国、欧盟均明确要求使用受版权保护的材料来训练模型，而日本则认定训练数据不受版权保护。

投资建议：整体来看，国内外对于模型训练数据的版权保护技术尚待成熟、政策尚待完善，未来版权提供商股价有望得到密集催化。展望数据归因技术的成熟使版权收入和 AI 生成内容量紧密挂钩，随着 AIGC 下游应用的商业潜力释放，有望持续带动版权提供商的授权收入增量。

AIGC 发展对于版权商的利好多于利空，展望预期差驱动股价回升。**建议关注海外版权商：Adobe、Shutterstock、Getty Image、Elsevier、Thomson Reuters，以及注重数据版权保护的 AI 公司：微软、谷歌、Meta。**

看好国内模型训练数据的版权保护继续完善，带动新闻媒体、影视等各类信息媒介版权提供商的业绩增长。**建议关注：1）AI+出版：**中国出版、中国科传；**2）图片版权库：**视觉中国；**3）影视版权库：**捷成股份、华策影视。

风险提示：AI 技术研发和产品迭代不及预期；AI 行业竞争加剧风险；商业化进展不及预期风险；国内外政策风险。

互联网传媒

买入（维持）

作者

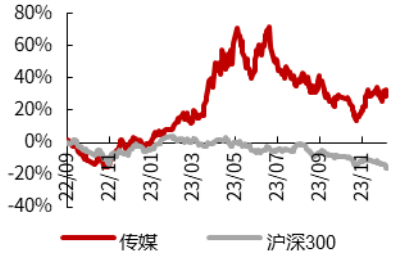
分析师：付天姿

执业证书编号：S0930517040002

021-52523692

futz@ebscn.com

行业与沪深 300 指数对比图



资料来源：Wind

相关研报

谷歌正式发布 Gemini，应用端和硬件端积极布局——AIGC 行业跟踪报告（三十八）（2023-12-07）

美图发布 AI 视觉大模型 4.0 版本，关注应用端落地情况——AIGC 行业跟踪报告（三十六）（2023-12-06）

探讨 AIGC 视频的核心痛点与未来趋势，Pika 1.0 能否带来新变化？——AIGC 行业跟踪报告（三十五）（2023-12-05）

探讨 GPTs 背后的产业逻辑：拉开 AIGC 应用生态的帷幕——AI 产业前瞻系列报告（一）（2023-11-20）

GPT-4 降价辐射 AIGC 应用产业链，定制化和 Agent 赋能使用体验——AIGC 行业跟踪报告（三十三）（2023-11-08）

美股 AIGC 应用端全产业链布局，商业化箭在弦上——AIGC 系列跟踪报告（二十八）（2023-10-14）

微软“AI+操作系统”初见雏形，生态壁垒是 AIGC 核心竞争力——AIGC 系列跟踪报告（二十七）（2023-09-27）

23Q3 美股互联网巨头财报：AIGC 应用各自争先，业绩潜力尚待释放——美国互联网科技公司跟踪专题报告（三）（2023-08-05）

目 录

图目录

表目录

1、 模型训练数据集如何保证版权合规性?

在 AI 模型的训练过程中，数据收集、清洗和标注是重要的前置环节。随着基于大模型的 AIGC 应用逐渐推广和商业化，模型训练数据是否侵权需要纳入考虑，用于模型训练的数据可以分为专有数据、开源数据、专用数据集等类型。

针对不同的数据类型有不同的方式来保证数据的版权，或通过直接补偿创作者的方式，在很大程度上降低了训练数据侵权的风险。随着 AI 模型的不断迭代和性能提升，以及下游应用产业链的繁荣和相关规章制度的成熟，科技公司需要付出越来越多的成本来保证训练数据的版权与合规性。

表 1：模型训练数据集保证版权合规性的具体方式和后续影响梳理

保护版权类型	保护版权方式	具体介绍	后续影响
专有数据	版权合作协议	AI 公司从版权提供商处获取训练数据，通过直接购买或双向合作的方式获得版权	Shutterstock、Getty Image 等版权库和 Axel Springer 等出版商受益
	API 付费访问	专业性强的数据提供商会针对 API 访问收费，而 23 年以来部分免费访问 API 的网站也开始收费	彭博、Elsevier 等专业领域数据和 Reddit、Twitter 等社交平台受益
开源数据	开放许可协议	常见开放许可协议包括知识共享（CC）、开放数据共享（ODC）、社区数据许可协议（CDLA）等	较为成熟的标准化授权协议
	特定的数据抓取策略	AI 公司在抓取网页数据时可以避开有版权保护的信息，网页维护者也可以加强对于数据爬取的审核	仍存在侵权隐患，部分公开网页内容本身存在侵权行为
直接补偿创作者	事前补偿	版权人的作品在被采纳为训练数据时获得补偿，如 Shutterstock 建立的贡献者基金	难以界定合理的补偿额度，仅作为过渡策略
	事后补偿	通过技术手段对训练数据溯源并进行对应的版权补偿，如归因模型可以计算数据源对生成内容的影响	针对性地做出补偿，补偿额度更合理，但技术尚不成熟
训练专用数据集	直接出售	出售能直接运用于 AI 和机器学习模型训练的数据集	专用数据集的数据量较少
	绑定 MaaS 服务	MaaS 服务商提供特定的模型训练集供客户使用，客户训练自己的模型时，用来丰富训练数据	作为 MaaS 服务的一部分进行变现，商业化路径更顺畅

资料来源：各公司官网，光大证券研究所整理

1.1、 专有数据：主要通过版权合作协议、API 付费访问等方式保障版权，商业空间广阔

AI 公司将专有数据用于模型训练，可以直接与版权方交涉，以保证训练数据集的版权合规性。包含特定领域的高质量数据以及未公开授权的私有数据，通常需要付费，但对于进一步提升大模型性能、增强模型的细分垂类能力很十分重要。**AI 公司获取专有数据的两个主要方式是版权合作协议和 API 付费访问。**

1.1.1、版权合作协议：海外 Shutterstock、Axel Springer 等多家版权提供商与 AI 公司建立合作

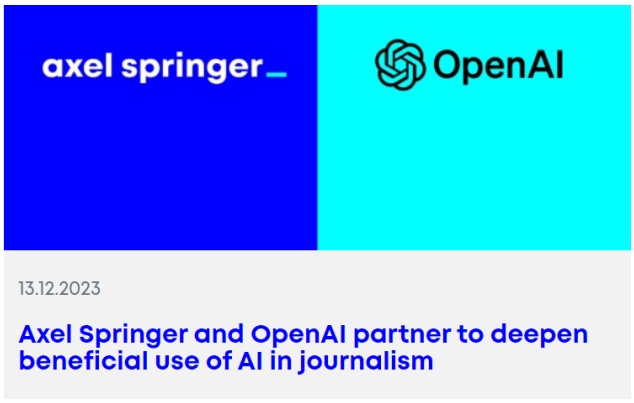
版权提供商的高质量语料对于模型性能提升十分重要，并且能降低数据清洗和标注的工作量。新闻版权商拥有丰富全面且更新及时的信息，文学作品、艺术创作、影视作品中包含大量高质量的训练素材；另外，部分素材库本身就具备针对图片、视频、音乐等素材的标注，能大幅降低数据清洗和标注的工作量。

多媒体版权库 Shutterstock 与 OpenAI、Meta、LG 等公司建立合作，将其图片、视频、音频等素材提供给合作伙伴用于模型训练，并从中获得收入；新闻出版社 Axel Springer 与 OpenAI 合作，其新闻素材将用于丰富 OpenAI 的模型训练数据集；以色列文生图模型公司 Bria AI 与 Getty Images 建立长期合作，采用 Getty Images、Alamy、Envato 等图像版权库的许可内容进行训练。

图 1：23M7 Shutterstock 与 OpenAI 加深合作的声明

资料来源：Shutterstock 官网博客

图 2：23M12 Axel Springer 与 OpenAI 建立合作的声明



资料来源：Axel Springer 官网博客

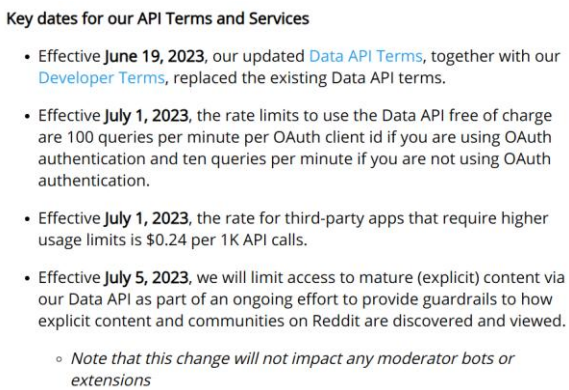
1.1.2、API 付费访问：23 年以来 Reddit、Twitter 等网站的 API 访问由免费转向了付费

通过 API 爬取网络数据也是模型训练数据的重要来源。随着大语言模型在不同细分行业的应用越来越深入，对于专业数据的需求也会水涨船高。

部分含金量高、专业性强的数据提供商会针对 API 访问进行收费。例如，金融领域的彭博 API、新闻媒体领域的纽约时报 API、虚数领域的 Elsevier API、电商领域的亚马逊 API、谷歌地图 API 等均需要付费使用。

社交平台、开源代码平台等非专业数据网站也逐步开始针对 API 访问收费。23M4 社交平台 Reddit 和 Twitter 的 API 访问从免费转向了付费，背后原因可能是大模型训练需求拉动下 API 调用量显著提升，为两家社交平台带来了较高的成本。开源代码平台 Stack Overflow 宣布会向 AI 公司收取训练数据费用。

图 3：23M6 Reddit 开始针对 API 访问进行收费



资料来源：Reddit 官网博客

图 4：23M3 Twitter 推出 Data API 的定价方案

资料来源：Twitter 官网博客

1.2、 开源数据：依靠开放许可协议、特定的数据抓取策略来保障版权，但仍存在侵权的隐患

AI 公司将开源数据用于模型训练，可以通过开放许可协议、特定数据抓取策略、手动筛查、社区监督等方式来保证版权合规性。开放许可协议是一种标准化授权方式，方便著作权持有人将数据授权给他人使用；抓取训练数据时也可以采取精细化的策略，充分尊重网站的 API 政策；此外，模型厂商也可以提升训练数据集的透明度，通过手动筛查和社区监督等方式来保证版权合规性。

1）开放许可协议：开源数据集的常见开放许可协议包括知识共享（CC）、开放数据共享（ODC）、社区数据许可协议（CDLA）等。知识共享协议提供六种选项：① CC BY：需注明作者、允许改编、允许用于商业用途；② CC BY-SA：需注明作者，改编作品必须在相同条款下共享；③ CC BY-NC：需注明作者、允许改编、不允许用于商业用途；④ CC BY-NC-SA：需注明作者，仅允许非商业用途，改编作品必须在相同条款下共享；⑤ CC BY-ND：需注明作者，不允许改编；⑥ CC BY-NC-ND：需注明作者，仅允许非商业用途，不允许改编。

2）特定的数据抓取策略：AI 公司在抓取网页数据时可以采用特定策略避开有版权保护的信息，网页维护者也可以加强对于数据爬取的审核。例如，网页的 Robot.txt 文件规定了搜索引擎抓取工具可以访问哪些网址，noindex 则可以禁止将某个网页编入索引，阻止抓取工具的访问。

3）社区监督：AI 公司可以提升训练数据集的透明度，鼓励社区监督，若训练数据的创作人主张侵权可以进行申诉。这种方法更适用于开源模型，而对于商业化的闭源模型，训练数据集往往会作为开发商技术壁垒的一部分进行保密。

整体来说，开源数据的获取已发展出了一套完善的版权保护制度，但仍存在一定的侵权隐患。例如，部分公开网页不具备完善的开放许可协议和针对 API 抓取的规定，甚至公开网页中的内容可能本身就存在侵权行为。

图 5：开源数据集 Wiki-links 的部分授权声明

You are free:
- to Share -- to copy, distribute and transmit the work;
- to Remix -- to adapt the work;
and to make commercial use of the work.

Under the following conditions:

* Attribution -- You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

With the understanding that:

* Waiver -- Any of the above conditions can be waived if you get permission from the copyright holder.

资料来源：Google Code Archive

图 6：Robot.txt 文件示例

下面是一个包含两条规则的简单 robots.txt 文件：

```
User-agent: Googlebot
Disallow: /nogooglebot/

User-agent: *
Allow: /

Sitemap: https://www.example.com/sitemap.xml
```

以下是该 robots.txt 文件的含义：

1. 名为 Googlebot 的用户代理不能抓取任何以 `https://example.com/nogooglebot/` 开头的网址。
2. 其他所有用户代理均可抓取整个网站。不指定这条规则也无妨，结果是一样的；默认行为是用户代理可以抓取整个网站。
3. 该网站的站点地图文件路径为 `https://www.example.com/sitemap.xml`。

资料来源：谷歌搜索中心

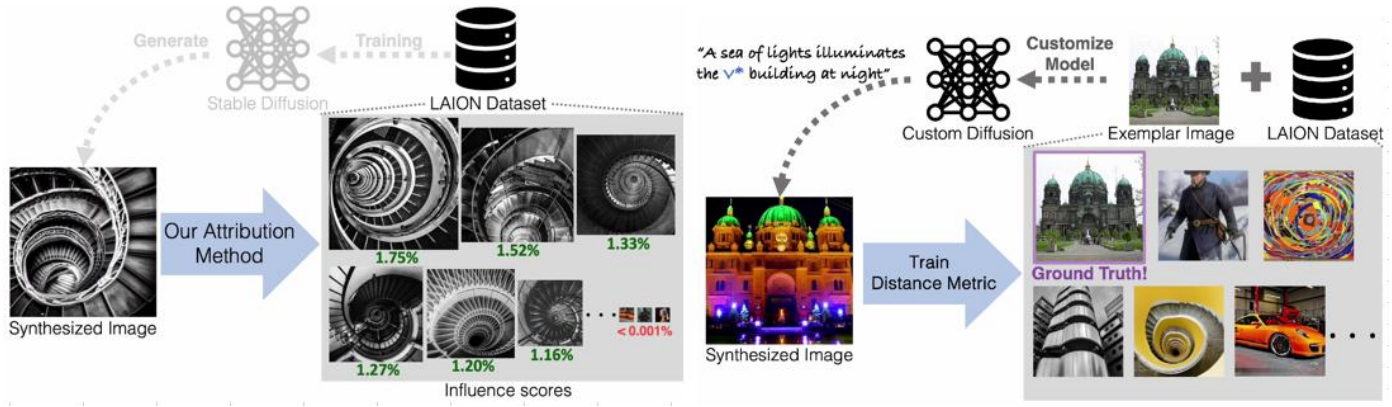
1.3、 直接补偿创作者：海外先进技术识别 AI 生成内容的版权来源，建立基金会为创作者提供补贴

通过直接补偿创作者来保障版权的做法主要分为两种：1）事前补偿：版权人的作品在被采纳为训练数据时获得补偿；**2）事后补偿：**通过特定技术追溯 AI 生成内容的训练数据源，并针对性地给予补偿。

事前补偿的技术难度较低，但难以界定合理的补偿额度。海外知名图片版权库 Shutterstock 建立了贡献者基金，当投稿人创作的内容被用于 AI 模型训练时将获得补偿，并在后续使用模型生成内容时持续获得补偿。此类方法可以保证创作者获得一定的报酬，但不同风格、不同质量的内容对模型训练的贡献各不相同，很难具体量化，会给补偿定价带来一定的难度。

事后补偿指通过技术手段对训练数据溯源并进行对应的版权补偿，定价更合理但技术难度尚不成熟。23M9 卡耐基梅隆大学、Adobe Research 和加州大学伯克利分校合作开发了两种算法，第一种算法可以阻止模型调用受版权保护的作品，第二种算法可以在模型用受版权保护的作品生成内容时为创作者提供补偿，该算法也能提供一种选择，让艺术家随时退出 AI 模型。另外，以色列文生图模型公司 Bria AI 于 23M9 开发了一种归因模型，能够计算数据源对 AI 生成内容的影响，从而对训练数据的版权人提供定价更加合理的报酬。

图 7：卡耐基梅隆大学开发的“评估文本到图像模型的数据归因”算法，可以追溯 AI 生成图像的训练数据来源



资料来源：《Evaluating Data Attribution for Text-to-Image Models》，作者：Sheng-Yu Wang，Alexie A. Efros 等

1.4、 专用数据集：直接出售适用于 AI 和 ML 的数据集，或作为 MaaS 服务的一部分提升用户体验

专用数据集指经过筛选和清洗、直接适用于模型训练的数据集，需要数据集提供方履行数据确权义务。专用数据集为开发者进行机器学习和模型训练相关研究提供强有力的支持，大多数为开源数据集，也有部分数据集被用来出售。对于云服务提供商，往往会将专用数据集打包成 MaaS 服务的一部分提供给用户，帮助用户更好地训练自己的定制化模型。

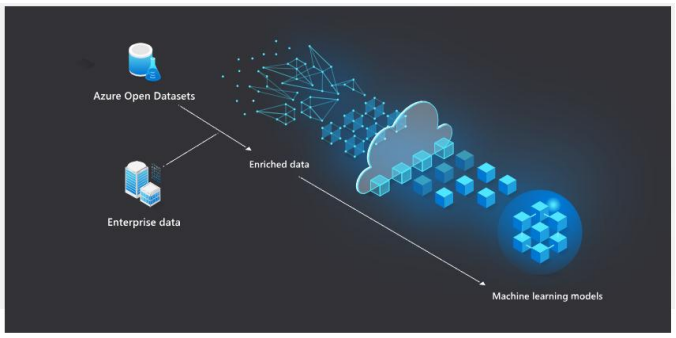
1）直接出售数据集：此类数据集经过了前期的筛选、整理和注释，由标记的示例或输入输出对组成，能直接运用于 AI 和机器学习模型训练。付费方式包括一次性购买和订阅制，具体价格受到数据量、准确度、覆盖时间和地区等因素的影响。例如，数据集商店 DataStock 售卖高质量、结构化的网页爬取数据集，涵盖零售、医疗、旅行等多个领域；数据交易平台 Datarade 划分出了 AI & ML 训练数据专区，供提供者和开发者进行数据集交易。

2）作为 MaaS 的一部分提供给用户：微软、亚马逊、谷歌等云服务提供商均提供帮助客户进行 AI 模型训练和应用程序开发的 MaaS 服务，包括自研和第三方 AI 模型调用接口、围绕模型训练技术细节的配套服务和指导等。对于定制化模型，数据集一般是客户的个性化数据，但部分 MaaS 平台也会提供特定的模型训练数据集供客户使用。例如，微软 Azure 云平台为客户提供特选数据集，使用可公开获取的数据制成，可在模型训练过程中随时访问。

图 8：模型训练数据集商店 DataStock 包含的行业数据

资料来源：DataStock 官网

图 9：Azure 提供开源数据集，与企业数据共同丰富训练数据



资料来源：微软 Azure 官网

2、 版权合作协议：盈利模式稳定、海外商业化成效初步展现

2.1、 版权提供商与 AI 公司的合作是互利共赢

AI 生成内容的快速增加，对于图片素材库、新闻出版社等版权提供商来说构成一定的威胁。1）AI 生成内容可能被上传至版权库混淆视听。随着大模型性能的不断突破，AI 生成内容的质量逐渐提升，甚至难以与人类作者和艺术家创作的内容区分。若版权素材库中被上传了大量的 AI 生成内容，可能会影响用户的付费意愿。**2）AI 生成内容可能成为版权素材库的替代品。**随着 AIGC 产品的推广和普及、未来大模型成本的不断降低，以及相关政策的不断完善，AI 生成内容将被越来越多地运用于商业化产品中，从而挤压传统版权素材库的生存空间。**因此，版权提供商也需要积极拥抱 AIGC 潮流，探索传统业务与 AI 技术结合的新形势。**

对于 AI 公司来说，模型训练需要海量的高质量数据，且 AIGC 产品也需要与更多信息源产生联动。公开渠道的数据存在侵权的风险，且需要耗费更多精力进行数据清洗和数据标注，为了模型后续的商业化和公司的长期健康发展，从版权提供商获取高质量训练数据是更好的方式。另外，版权提供商也可以丰富 AIGC 产品的信息来源和产品功能，赋能用户使用体验。

2.1.1、海外多媒体版权库 Shutterstock：出售模型训练素材创收，通过基金会为创作者提供补偿

海外知名多媒体版权库 Shutterstock 紧随 AIGC 浪潮，推出了 AI 生成图片专区，并提供由 OpenAI 支持的 AI 文生图工具。Shutterstock 拥有 100 多万投稿者贡献的超 4.5 亿张图片，提供的多媒体素材主要包括：**1）图片：**矢量图、

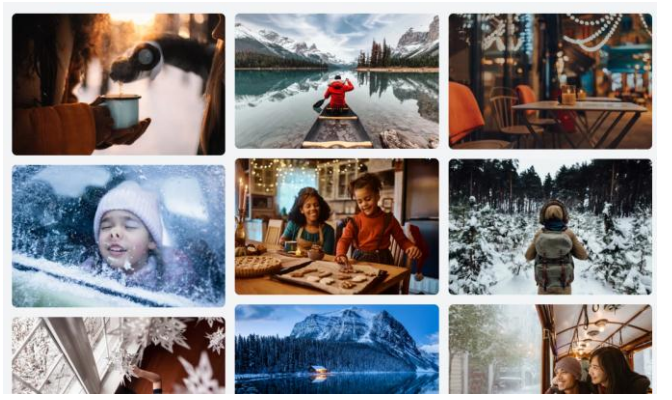
照片、AI 生成的图片等；**2）Pond5 视频平台：**镜头、AE 素材、音效、3D 模型等；**3） 设计：**商业营销模板、社交媒体模板等。此外，Shutterstock 还提供设计工具，包括图片编辑器、抠图工具、AI 生成图片工具等。

Shutterstock 与 OpenAI 的双向合作始于 2021 年。2021 年 Shutterstock 与 OpenAI、LG 开始合作；23M7 OpenAI 与 Shutterstock 进一步加深合作关系，签订了为期六年的合作协议。

图 10：Shutterstock 图片素材主页，包含 AI 图片生成工具

资料来源：Shutterstock 官网

图 11：Shutterstock 拥有丰富的图片版权资源



资料来源：Shutterstock 官网

Shutterstock 与 AI 公司的合作可以概括为三个方面：

1）Shutterstock 向 OpenAI 提供图片素材版权用于模型训练。签订协议后，OpenAI 有权访问 Shutterstock 的图像、视频、音乐等素材用于 AI 模型的训练数据。Shutterstock 拥有丰富且高质量的内容素材版权，在多样性和数据标注上处于行业领先地位，使其在训练 AI 模型上具备较大的优势。

2）Shutterstock 设立了贡献者基金，当投稿人创作的图片被用于 AI 图像模型训练时将获得补偿。Shutterstock 是首个推出贡献者基金的公司，截至 23M7，该基金已为数十万创作者提供补偿，并通过与新生成资产许可活动相关的版税为创作者们提供持续补偿。

3）AIGC 文生图和图片编辑工具集成进 Shutterstock 平台，并得到 OpenAI 的文生图模型 DALL·E 的支持。创作的图片被用于模型训练的投稿人将获得 AI 文生图工具的长期使用权。除 OpenAI 外，Shutterstock 还与英伟达、Meta、LG 等公司建立合作，共同开发文本、图像、3D 等领域的 AIGC 创作工具。

2.1.2、海外新闻出版商 Axel Springer：为 OpenAI 提供文本训练数据，通过链接为创作者引流

出版社的优质文章素材是大模型训练的高质量文本语料来源，有助于加快大模型性能迭代，促进 AI 生成内容的版权制度完善。

2023 年 12 月 13 日，德国数字出版商 Axel Springer 和 OpenAI 宣布建立全球伙伴关系，并成为全球第一家与 OpenAI 合作的新闻社。

1）对于 OpenAI：OpenAI 将付费使用 Axel Springer 出版物的内容，完善其 AI 模型训练数据库。ChatGPT 用户将收到 Axel Springer 旗下媒体品牌精选的全球新闻摘要。当 ChatGPT 使用 Axel Springer 出版物中的信息回答用户问题时，将在答案下方提供来源链接，确保内容版权方获得信用、补偿和流量。

2）对于 Axel Springer：可通过向 AI 公司提供优质内容素材开辟新业务线，获

取潜在收入增量，同时利用 OpenAI 的技术支持改进其产品。通过与 OpenAI 合作，利用 AI 来增强内容体验和创造新的发展机会，探索新闻业的未来方向。

OpenAI 曾多次因未经允许使用新闻媒体的文章训练模型引发争议。美国头部新闻机构《华尔街日报》、《纽约时报》都曾因版权问题与 OpenAI 发生过纠纷。23M2，New Corp 道琼斯部门的总法律顾问 Jason Conti 在给彭博新闻社的一份声明中表示，任何使用《华尔街日报》培训 AI 的企业应该向道琼斯公司寻求许可；23M8《纽约时报》更新服务条款，禁止其新闻报导和图片用于开发应用软件和训练 AI 模型，并警告如果持续引发争议将起诉 OpenAI。

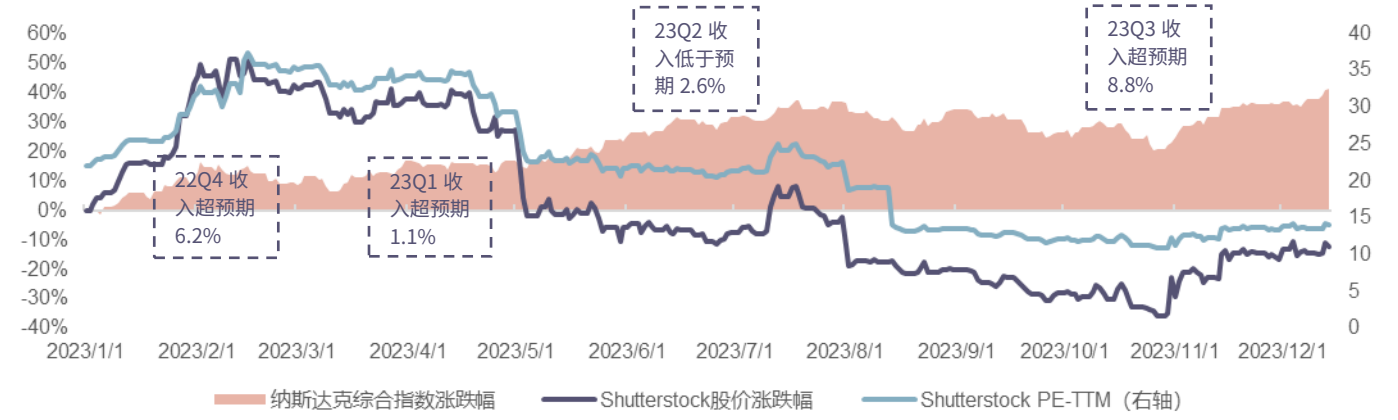
与 Axel Springer 的商业合作是 OpenAI 与世界各地出版商互利共赢的起点。OpenAI 首席运营官 Brad Lightcap 宣布 OpenAI 将致力于与世界各地的出版商和创作者合作，确保他们从先进的 AI 技术和新的收入模式中受益。

2.2、 从 Shutterstock 看多媒体版权库与 AI 公司的合作：AIGC 的利好整体强于利空

2.2.1、Shutterstock 的数据授权收入已较明显体现在业绩端，驱动估值修复和股价回升

随着向合作伙伴出售数据的业务逐渐释放业绩潜力，Shutterstock 股价触底反弹。23M1-23M5，Shutterstock 股价出现了快速上升和回调，后续股价呈现下跌趋势，直至 23Q3 业绩发布后股价开始反弹。

图 12：2023 年 1 月 1 日-2023 年 12 月 14 日纳斯达克综合指数、Shutterstock 股价涨跌幅与 Shutterstock PE-TTM 变化趋势



资料来源：Wind，Refinitiv，Shutterstock 公告，光大证券研究所整理

1）23M1-23M4：受 AIGC 行业投资逻辑催化，股价大幅拉升。AIGC 投资热点下，市场开始挖掘潜在受益的产业，Shutterstock 作为 2021 年起就与 OpenAI 建立合作的公司而受到关注，且大模型训练拉动训练数据版权需求的逻辑非常顺畅，2023 年以来最高涨幅达 51.1%。**2）23M5-23M10：**市场开始担忧 AI 文生图的快速发展挤压 Shutterstock 的传统业务图片版权出售。

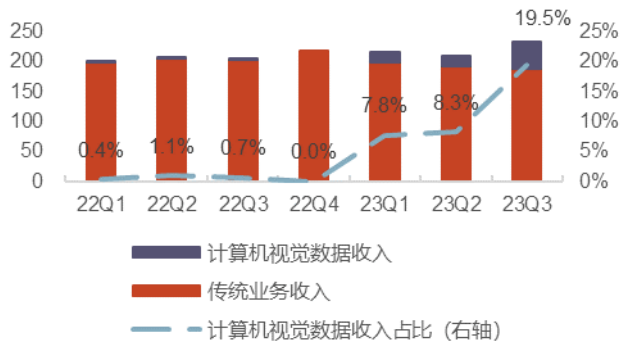
3）随着 Shutterstock 出售 AI 模型训练数据授权的收入快速增长，股价触底反弹。Shutterstock 披露的向合作伙伴提供计算机视觉数据项目（Computer Vision Data Partnerships Offering）代表向大型科技公司提供的图片、视频、音乐、3D 模型等素材授权，用于训练生成式 AI 和机器学习模型。23Q3 该项收入达到 4550 万美元，占公司总收入的 19.5%；23 年前九个月该项收入达到 7950 万美元，占公司总收入的 12.1%。

2.2.2、Shutterstock 传统业务下滑原因众多，AIGC 对于版权提供商的威胁和替代尚不明显

我们认为，Shutterstock 传统业务下滑并非受 AI 文生图的影响，更可能源于竞争压力等多种因素影响。我们将 Shutterstock 的传统业务（不包含计算机视觉数据收入）与其竞争对手 Getty Image 的收入进行对比。Shutterstock 的传统业务代表排除了出售用于大模型训练数据之外的其他业务，包括电商业务（客户可以按月订阅，或按需付费下载图片）以及企业服务，为客户提供图片库、视频等素材，与 Getty Image 的收入更具可比性。图片版权提供商 Getty Images 凭借其丰富的高质量图片资源在图片库市场展现出强有力的竞争力。

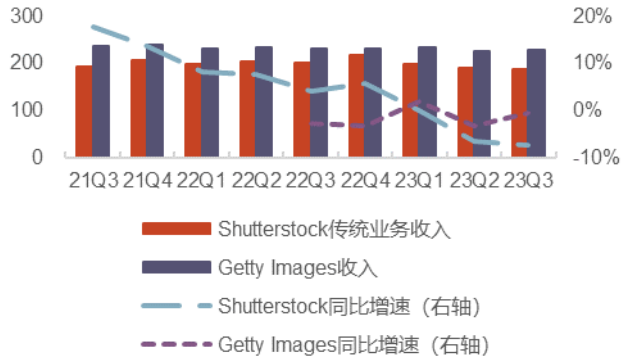
23 年以来，Getty Image 的收入维持稳定，并未明显受到 AI 文生图的影响。作为 Shutterstock 的竞争对手，Getty Images 并未出售 AI 模型训练数据，其近两年总收入相对稳定，23Q3 总收入为 2.3 亿美元，同比下降 0.5%。相比 Getty Image， Shutterstock 传统业务收入自 22Q4 以来连续下滑，23Q3 跌至 1.9 亿美元，同比下降 7.3%，同时 23Q3 Shutterstock 订阅用户数量和付费下载量也呈下降趋势。我们认为，Shutterstock 的传统业务收入下滑更多受到同业竞争压力的影响，但计算机视觉数据出售也成为了业绩的新增长点。

图 13： 22Q1-23Q3 Shutterstock 传统业务收入，计算机视觉数据收入和占总收入的比例（单位：百万美元）



资料来源：Shutterstock 公告，光大证券研究所整理

图 14：21Q3-23Q3 Shutterstock 传统业务（不包含计算机视觉数据收入）、Getty Image 收入（单位：百万美元）



资料来源：各公司公告，光大证券研究所整理

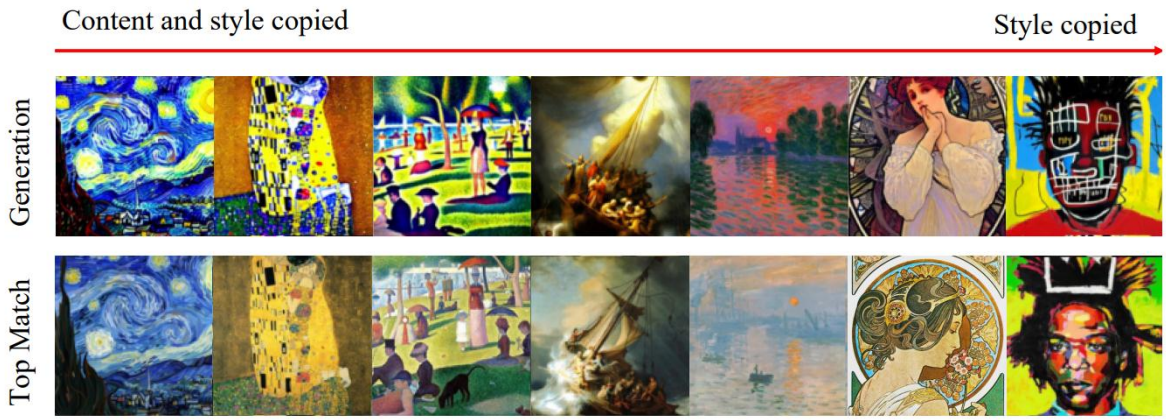
3、国内外模型训练数据版权规定尚待完善，版权商股价有望得到密集催化

截至 23 年底，公众对于 AI 文生图和其他多模态生成的反感情绪仍较为强烈。2023 年 12 月 6 日，春晚吉祥物“龙辰辰”被质疑是 AI 作图，受到了国内民众的广泛批评。自 Stable Diffusion、Midjourney 等文生图软件走入公众视野，便引发了关于 AI 生成图片是否侵权的持续讨论。22M12 马里兰大学帕克分校和纽约大学合作发布的一篇论文显示，一些参数量较小的文生图模型会直接复制用于训练的图片素材的某个部分，而当时较为成熟的文生图产品 Stable Diffusion 也出现了以像素点级别复制名画的细节、结构和绘画风格的情况。

公众对于 AI 多模态生成的质疑主要来自于：1）模型训练采用的图片素材是否获得授权；**2）**通过机器学习生成图片是否可以被定义为学习和创作的过程；**3）**AI 生成图片过程中，运用于训练数据的图片素材是否被简单粗暴地拼接。

逐渐扭转公众对于 AI 多模态生成的消极情绪和片面认知，是 AI 图片、AI 视频等技术推广至生产生活、释放商业化潜力的必要前提。随着 AIGC 的影响力快速扩大，科技公司也需要付出更多成本以确保模型训练数据和生成内容的版权和合规性，以应对未来可能的法律挑战。

图 15：22M12 一篇论文显示，Stable Diffusion 以像素点级别复制了名画的细节、结构和绘画风格



资料来源：《Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models》(2022-12-12)，作者：Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, Tom Goldstein

对 AIGC 的版权问题的争议和相关法规主要可以分为两类：

1）AI 生成内容的版权界定：指由 AI 生成的文字、图片等内容是否受到版权保护，以及版权应当归属于用户、模型提供商、训练数据提供者等哪一方。对于 AI 生成内容版权的明确，是 AIGC 产品大规模商业化的重要前提。

2）模型训练数据的版权规定：指 OpenAI、Stability AI 等模型供应商在训练基础模型时采用的数据集是否受到版权保护，模型供应商应该以怎样的方式获得训练数据集的版权。对于训练数据的版权规定，是 AIGC 产业良性可持续发展、以及令公众消除对 AI 生成内容的消极情绪的关键。

表 2：国内外关于 AI 生成内容和模型训练数据的版权规定与相关纠纷判决

AIGC 版权分类	国家	法规或相关判决	时间	具体介绍
AI 生成内容的版权界定	美国	美国版权局《联邦法规》规定 AI 生成内容不受版权法保护	2023 年 3 月 16 日	区别于有人工参与创作的 Photoshop 作品，通过 Midjourney、Stability AI、ChatGPT 等平台自动生成的作品完全由 AI 完成，并且训练的数据是基于人类创作的作品，因此不受版权法保护。
		美国版权局拒绝 Midjourney 生成图片的版权申请	2023 年 3 月 6 日	美国版权局在批准《Zarya of the Dawn》时，拒绝为小说中 Midjourney 生成的插图提供版权保护。
	中国	深圳法院判定 AI 生成内容受著作权法保护	2020 年 1 月 8 日	深圳市南山区人民法院在腾讯 AI 协作工具 Dreamwriter 引发的著作权纠纷案中做出判决，首次认定 AI 生成内容具有独创性，应当获得著作权法保护。
		北京互联网法院认定 AI 生成图片属于著作权法上的美术作品	2023 年 12 月 1 日	北京互联网法院针对一起人工智能生成图片（AI 绘画图片）著作权侵权纠纷做出一审判决，肯定了 AI 绘画大模型生成的涉案图片属于著作权法上的美术作品，原告对其拥有著作权。为国内 AI 生成图片相关领域著作权的第一案。
模型训练数据的版权规定	美国	美国新闻媒体联盟发布《生成式 AI 监管原则》	2023 年 4 月 26 日	美国新闻媒体联盟代表近 2000 家印刷和数字媒体出版商发布了《生成式 AI 监管原则》，强调生成式 AI 的开发者和部署者必须尊重创作者对其内容的权利。
		Getty Image 起诉 Stability AI 未经允许自动爬取其拥有著作权的图片素材用于模型训练	2023 年 1 月 18 日	Shutterstock 的同类公司、图片素材版权库 Getty Image 起诉 Stability AI 的侵权行为，未经允许从其网站上窃取了数百万张照片。
	欧盟	《人工智能法案》	2023 年 6 月 14 日	要求 OpenAI、谷歌、微软等基础模型的供应商声明是否使用受版权保护的材料来训练 AI，并添加了透明度和风险评估要求。
	日本	重申日本法律认定 AI 训练数据不受版权保护	2023 年 6 月 2 日	日本文部科学大臣永冈桂子表示，日本法律不会保护人工智能使用的原始材料版权，训练数据“无论用于非盈利还是商业目的，无论是否从非法网站或其他方面获取”政策都允许。

资料来源：中国、美国、欧盟等政府官方网站，光大证券研究所整理

对于 AI 生成内容的版权界定，美国不承认 AI 生成内容拥有著作权，而中国倾向于保护 AI 生成内容的著作权。

美国版权局 2023 年 3 月 16 日收录于《联邦法规》中的 AI 版权认定和登记指引政策表明，仅由 AI 生成的作品不受版权保护，包含 AI 生成内容的作品根据情况给予登记；2023 年 3 月 6 日，美国版权局拒绝为一篇小说中包含的 AI 生成插图提供版权登记。中国的《生成式人工智能服务管理暂行办法》未给出明确规定，在 2020 年 1 月和 2023 年 12 月的两起 AI 生成内容引发的著作权纠纷案中，判决都肯定了 AI 生成内容具有独创性，应当获得著作权法保护。

对于模型训练数据的版权规定，美国、欧盟均明确要求使用受版权保护的材料来训练模型，而日本则认定训练数据不受版权保护。

美国新闻媒体联盟于 23M4 发布的《生成式人工智能监管原则》中强调生成式 AI 的开发者和部署者必须尊重创作者对其内容的权利。欧盟于 23M6 投票通过《人工智能法案》，要求 OpenAI、谷歌和微软等基础模型的供应商声明是否使用受版权保护的材料来训练 AI，并添加了透明度和风险评估要求。日本则在 23M6 重申日本法律不会保护人工智能使用的原始材料版权，无论是否从非法网站或其他方式获取训练数据，政策上都是允许的。

表 3：国内外关于人工智能良性健康发展的方向性指导文件

	名称	发布时间	主要内容
中国	互联网信息服务算法推荐管理规定	2021 年 11 月 16 日	对算法推荐技术提供互联网信息服务的企业活动进行了规定
	互联网信息服务深度合成管理规定	2022 年 11 月 3 日	制定了深度合成服务的治理和相关监督管理规定，完善了监管机制
	人工智能战略院发布《中国新一代人工智能科技产业发展 2023》	2023 年 5 月 19 日	人工智能带来的创新生产方式的变革，不仅带来产业的快速发展，而且带来科技创新范式和教育范式的新变革。
	七部门联合发布《生成式人工智能服务管理暂行办法》	2023 年 7 月 15 日	23 年 8 月 15 日开始施行，采取有效措施鼓励生成式人工智能创新发展；划定底线，推动生成式人工智能向上向善；采取更精细化监管举措
美国	《关于通过联邦政府进一步促进种族平等和支持服务不足社区的行政命令》	2023 年 2 月 16 日	该命令对部署人工智能系统的联邦机构规定了新的公平义务，要求各机构“防止和补救歧视，包括通过保护公众免受算法歧视”。
	美国版权局发布于《联邦法规》的生成式 AI 版权规定条例	2023 年 3 月 16 日	区别于有人工参与创作的 Photoshop 作品，通过 Midjourney、Stability AI、ChatGPT 等平台自动生成的作品完全由 AI 完成，并且训练的数据是基于人类创作的作品，因此不受版权法保护。
	美国 NMA《生成式人工智能监管原则》	2023 年 4 月 26 日	生成式 AI 的开发者和部署者必须尊重创作者对其内容的权利
	美国国家科学技术委员会更新《美国人工智能战略》	2023 年 5 月更新	自 2018 年不断更新，强调人工智能的重要性，并呼吁在创新、国家安全和劳动市场等方面采取相应措施。
	美国政府《关于安全、可靠和可信地开发和使用人工智能的行政命令》	2023 年 10 月 30 日	为人工智能安全和安保确立了新的标准，保护了美国人的隐私，促进了公平和公民权利，促进了创新和竞争。
欧盟	《人工智能协调计划》	2018 年 4 月 1 日	欧盟成员国和挪威在 2018 年签署的“合作宣言”为基础，强调愿意在人工智能方面进行更密切的合作，包括计算能力、微电子、TEF、数字创新中心等。
	《人工智能、机器人和相关技术的伦理方面框架》	2020 年 1 月 16 日	包括关于通过机器学习或深度学习开发、实施和技术演变的条款。目标是将监督扩展到高度复杂技术的所有领域。
	《人工智能白皮书》	2020 年 2 月 19 日	在欧盟委员会采取的广泛数字举措的框架内，关于人工智能的白皮书促进了欧盟对人工智能对人类和道德影响的协调方法。
	《人工智能法案》	2023 年 6 月 14 日	要求 OpenAI、谷歌、微软等基础模型的供应商声明是否使用受版权保护的材料来训练 AI，并添加了透明度和风险评估要求。
OECD	《生成式人工智能的初步政策考虑》	2023 年 5 月 19 日	AIGC 的政策挑战包括劳动力市场的潜在变化、版权的不确定性、传播错误和虚假信息、歧视、扭曲公共话语、煽动暴力等。
英国	《支持 AI 创新的监管方法》	2023 年 3 月 29 日	英国希望通过此政策文件消除用户的忧虑并降低 AI 的输出风险，激励企业、民众共同参与到这场 AI 技术变革中
日本	《知识产权战略计划》	2023 年 6 月 9 日	重点关注生成式 AI 的知识产权问题，提出“要适当地应对外界担忧和潜在风险，以促进生成式 AI 的开发、提供和使用”。

资料来源：中国、美国、欧盟等政府官方网站，光大证券研究所整理

4、 投资建议

整体来看，国内外对于模型训练数据的版权保护技术尚待成熟、政策尚待完善，未来版权提供商股价有望得到密集催化。随着模型训练数据的版权规定进一步完善，有利于扭转公众对于 AI 多模态生成的消极情绪，促进 AI 生成图片、视频等技术的产品化，释放商业潜力。展望数据归因技术的成熟使版权收入和 AI 生成内容量紧密挂钩，随着 AIGC 下游应用的商业潜力释放，有望持续带动版权提供商的授权收入增量。

复盘 Shutterstock 的业绩和股价表现，AIGC 产业的发展对于版权提供商的利好多于利空，预期差驱动股价回升。Shutterstock 授权给科技公司训练 AI 模型的数据出售收入快速增长，而传统业务的下滑更多受到同业竞争压力的影响，AI 生成图片替代版权提供商传统业务的担忧逐渐消退，而模型训练带动的版权收入快速增长形成了预期差，驱动 Shutterstock 股价回升。**建议关注海外版权提供商：Adobe、Shutterstock、Getty Image、Elsevier、Thomson Reuters，以及注重数据版权保护的 AI 公司：微软、谷歌、Meta。**

看好国内模型训练数据的版权保护继续完善，带动新闻媒体、图片、影视等各类信息媒介版权提供商的业绩增长。**建议关注：1）AI+出版：**中国出版、中国科传；**2）图片版权库：**视觉中国；**3）影视版权库：**捷成股份、华策影视。

5、 风险提示

中美地缘政治摩擦、宏观经济不及预期、AIGC 技术发展和应用落地进度不及预期、AI 行业竞争加剧风险。

行业及公司评级体系

	评级	说明
行业及公司评级	买入	未来 6-12 个月的投资收益率领先市场基准指数 15%以上
	增持	未来 6-12 个月的投资收益率领先市场基准指数 5%至 15%；
	中性	未来 6-12 个月的投资收益率与市场基准指数的变动幅度相差-5%至 5%；
	减持	未来 6-12 个月的投资收益率落后市场基准指数 5%至 15%；
	卖出	未来 6-12 个月的投资收益率落后市场基准指数 15%以上；
	无评级	因无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使无法给出明确的投资评级。
基准指数说明：		A 股市场基准为沪深 300 指数；香港市场基准为恒生指数；美国市场基准为纳斯达克综合指数或标普 500 指数。

分析、估值方法的局限性说明

本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。

分析师声明

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，并对本报告的内容和观点负责。负责准备以及撰写本报告的所有研究人员在此保证，本研究报告中任何关于发行商或证券所发表的观点均如实反映研究人员的个人观点。研究人员获取报酬的评判因素包括研究的质量和准确性、客户反馈、竞争性因素以及光大证券股份有限公司的整体收益。所有研究人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

法律主体声明

本报告由光大证券股份有限公司制作，光大证券股份有限公司具有中国证监会许可的证券投资咨询业务资格，负责本报告在中华人民共和国境内（仅为本报告目的，不包括港澳台）的分销。本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格编号已披露在报告首页。

中国光大证券国际有限公司和 Everbright Securities(UK) Company Limited 是光大证券股份有限公司的关联机构。

特别声明

光大证券股份有限公司（以下简称“本公司”）成立于 1996 年，是中国证监会批准的首批三家创新试点证券公司之一，也是世界 500 强企业——中国光大集团股份公司的核心金融服务平台之一。根据中国证监会核发的经营证券期货业务许可，本公司的经营范围包括证券投资咨询业务。

本公司经营范围：证券经纪；证券投资咨询；与证券交易、证券投资活动有关的财务顾问；证券承销与保荐；证券自营；为期货公司提供中间介绍业务；证券投资基金代销；融资融券业务；中国证监会批准的其他业务。此外，本公司还通过全资或控股子公司开展资产管理、直接投资、期货、基金管理以及香港证券业务。

本报告由光大证券股份有限公司研究所（以下简称“光大证券研究所”）编写，以合法获得的我们相信为可靠、准确、完整的信息为基础，但不保证我们所获得的原始信息以及报告所载信息之准确性和完整性。光大证券研究所可能将不时补充、修订或更新有关信息，但不保证及时发布该等更新。

本报告中的资料、意见、预测均反映报告初次发布时光大证券研究所的判断，可能需随时进行调整且不予通知。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。客户应自主作出投资决策并自行承担投资风险。本报告中的信息或所表述的意见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。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，本公司及作者均不承担任何法律责任。

不同时期，本公司可能会撰写并发布与本报告所载信息、建议及预测不一致的报告。本公司的销售人员、交易人员和其他专业人员可能会向客户提供与本报告中观点不同的口头或书面评论或交易策略。本公司的资产管理子公司、自营部门以及其他投资业务板块可能会独立做出与本报告的意见或建议不相一致的投资决策。本公司提醒投资者注意并理解投资证券及投资产品存在的风险，在做出投资决策前，建议投资者务必向专业人士咨询并谨慎抉择。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问或金融产品等相关服务。投资者应当充分考虑本公司及本公司附属机构就报告内容可能存在的利益冲突，勿将本报告作为投资决策的唯一信赖依据。

光大证券研究所

上海
静安区南京西路 1266 号
恒隆广场 1 期办公楼 48 层

北京
西城区武定侯街 2 号
泰康国际大厦 7 层

深圳
福田区深南大道 6011 号
NEO 绿景纪元大厦 A 座 17 楼

光大证券股份有限公司关联机构

香港
中国光大证券国际有限公司
香港铜锣湾希慎道 33 号利园一期 28 楼

英国
Everbright Securities(UK) Company Limited
6th Floor, 9 Appold Street, London, United Kingdom, EC2A 2AP