证券研究报告

2023 年 12 月 31 日

计算机周报（12.25—12.29）

AIGC 产业链近况更新

投资建议： 强于大市（维持）

上次建议： 强于大市

相对大盘走势

作者

分析师：姜青山

执业证书编号：S0590523050001

邮箱：jiangqs@glsc.com.cn

分析师：黄楷

执业证书编号：S0590522090001

邮箱：huangk@glsc.com.cn

➢ **AIGC 正反馈不断加强**

AIGC 是人工智能、大数据、云计算等多个技术领域的整合，是一种跨领域的合作发展模式。AIGC 的四个核心要素：数据、算力、模型和应用，正在逐步实现正反馈。大模型商业化运营、多模态开发等需要更多的算力资源，同时好的产品可以直接带来更多的需求，属于典型的“供给创造需求”。未来随着技术的进步和应用场景的不断拓展，这 4 个要素会不断实现正反馈，推动整个行业的创新和发展。

➢ **算力产业链加速发展**

算力是大模型的底座，呈现加速发展趋势。11 月 13 日英伟达发布了目前世界上最强的 AI 芯片 H200。12 月 15 日，联想两款 AI Ready 的 AI PC 产品正式上市。12 月 19 日，摩尔线程首个全国产千卡千亿模型训练平台——摩尔线程 KUAE 智算中心揭幕。算力网络方面，中国网络设备公司如华为、新华三和锐捷网络，互联网巨头如阿里巴巴、百度、字节跳动、腾讯等在 2023 年下半年陆续加入 UEC 联盟。

➢ **大模型能力不断提升**

国内外大模型厂商持续进行研发投入，大模型能力持续提升。11 月 7 日，OpenAI 推出了 GPT-4 Turbo，对六个方面进行了升级。12 月 6 日，谷歌发布原生多模态大模型 Gemini。12 月 21 日，医渡科技发布自主研发的医疗垂域大模型。12 月 22 日，文心一言等四款国产大模型率先达到国家相关标准。工业和信息化部赛迪研究院预计 2023 年，我国语言大模型市场规模将达到 132.3 亿元，增长率将达到 110%。

➢ **文生视频引领 AIGC 应用**

AIGC 快速发展，已在文字、图像、音频、视频等领域得到应用。文生视频领域进展备受关注，Meta、字节跳动、Stable AI、Pika labs 等均发布相关产品。文生图领先模型 Midjourney 发布 V6 版本，在语义理解、图像质量等方面取得提升。12 月 25 日，昆仑万维 AI Agents 开发平台“天工 SkyAgents”Beta 版正式开放测试。

➢ **投资建议**

建议关注**算力、模型、应用三条主线**，包括：**（1）国产算力基础设施：**海光信息、寒武纪-U、中科曙光、浪潮信息、紫光股份、锐捷网络等；**（2）AI 模型商业化：**百度、科大讯飞、商汤、拓尔思等；**（3）AI 应用：①音视频：**万兴科技、海康威视、大华股份等；**②图像：**美图公司、虹软科技等；**③办公：**金山办公、福昕软件、泛微网络、用友网络等；**④垂直领域：**同花顺、恒生电子、宇信科技、中科软、卫宁健康、医渡科技、中科创达等。

风险提示：AI 技术发展演进不及预期；商业化进程不及预期；法律政策监管风险；行业竞争加剧。

相关报告

1、《计算机：AI 创新推动计算机行业加速发展》2023.12.27

2、《计算机：机器视觉产业链梳理》2023.12.23

正文目录

图表目录

1. AIGC 正反馈持续加强

2023 年我们见证了 AIGC 的快速崛起和破圈发展，虽然过程中有所曲折，但对于 AIGC 领域而言，可能仅仅是一个开始。在过去的深度学习黄金十年，人工智能的感知、理解能力不断增强，为 AIGC 的快速发展奠定基础。如今，随着生成算法、大模型、多模态技术等 AI 技术的持续创新和发展成熟，AI 领域正在经历从感知、理解到生成、创造的跃迁。以 AIGC 为标志 AI 领域正在迎来下一个时代。融合大模型和多模态技术的 AIGC 模型，有望成为新的技术平台深度赋能各行各业。未来，“AIGC+”将在经济社会的各个领域持续大放异彩。

1.1 AIGC 产业链发展迅速

依托于强大的算力基础设施和海量的通用数据，经过训练和不断调优打造出了 AIGC 大模型，到中间层的垂直化、定制化、个性化的模型工具，再到下游层出不穷的、各种各样的 AIGC 应用和服务，AIGC 的产业生态正在加速形成和发展。AIGC 将创造出很大的经济社会价值，其应用不限于互联网领域，也将给文化、娱乐、教育、金融、医疗、公共服务、交通、制造等诸多领域带来积极影响。经历了 2023 年的喧嚣和炒作，2024 年 AIGC 产业有望迎来更大的发展。

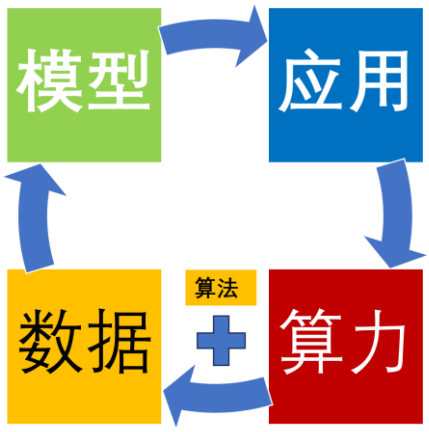
图表1：AIGC 产业链图谱

资料来源：国联证券研究所整理

1.2 AIGC 正反馈不断加强

AIGC 是人工智能、大数据、云计算等多个技术领域的整合，是一种跨领域的合作发展模式。AIGC 的四个核心要素：数据、算力、模型和应用，正在逐步实现正反馈。例如 OpenAI、微软 Office 和文心一言已经率先实现了商业化，开始了商业覆盖，为了满足用户日益增长的访问需求，算力基础设施将会增加。多模态等更强大的模型也需要更多的算力资源，同时好的产品可以直接带来更多的需求，属于典型的“供给创造需求”。未来随着技术的进步和应用场景的不断拓展，这 4 个要素会不断实现正反馈，推动整个行业的创新和发展。

图表2：AIGC 正反馈不断加强



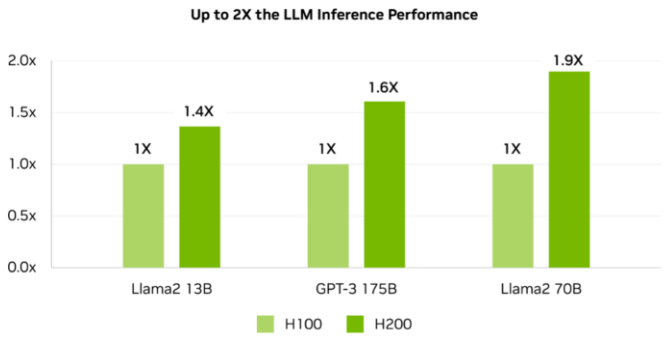
资料来源：国联证券研究所整理

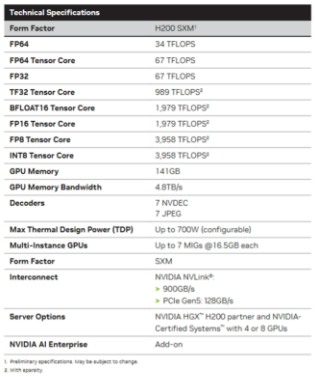
2. 算力产业链加速发展

2.1 英伟达推出 H200 高性能 GPU

英伟达在 2023 年 11 月 13 日，全球超算大会（SC2023）上发布了目前世界上最强的 AI 芯片 H200。这款新的 GPU 是基于 H100 进行了升级，内存带宽提高了 1.4 倍，内存容量提高了 1.8 倍，提高了处理生成式 AI 任务的能力。

图表3：英伟达 H200 计算和推理能力显著提升





Technical Specifications	
Form Factor	H200 SXM[1]
FP64	34 TFLOPS
FP64 Tensor Core	67 TFLOPS
FP32	67 TFLOPS
TF32 Tensor Core	989 TFLOPS[2]
BFLOAT16 Tensor Core	1,979 TFLOPS[2]
FP16 Tensor Core	1,979 TFLOPS[2]
FP8 Tensor Core	3,958 TFLOPS[2]
INT8 Tensor Core	3,958 TFLOPS[2]
GPU Memory	141GB
GPU Memory Bandwidth	4.8TB/s
Decoders	7 NVDEC 7 JPEG
Max Thermal Design Power (TDP)	Up to 700W (configurable)
Multi-Instance GPUs	Up to 7 MIGs @16.5GB each
Form Factor	SXM
Interconnect	NVIDIA NVLink® › 900GB/s › PCIe Gen5: 128GB/s
Server Options	NVIDIA HGX™ H200 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs
NVIDIA AI Enterprise	Add-on

1. Preliminary specifications. May be subject to change.
2. With sparsity.

资料来源：英伟达《NVIDIA H200 Tensor Core GPU Datasheet》，国联证券研究所

此外，12 月初英伟达 CEO 黄仁勋在新加坡表示，中国市场占英伟达销售额的 20%左右，英伟达将为中国市场提供一套符合美国政府最新规定的新产品。

2.2 国产算力持续进步

12 月 19 日，摩尔线程首个全国产千卡千亿模型训练平台——摩尔线程 KUAE 智算中心揭幕仪式在北京成功举办，宣告国内首个以国产全功能 GPU 为底座的大规模算力集群正式落地。

图表4：摩尔线程 MTT S4000 智算加速卡发布

资料来源：摩尔线程官网，国联证券研究所

摩尔线程大模型智算加速卡 MTT S4000，采用第三代 MUSA 内核，单卡支持 48GB 显存和 768GB/s 的显存带宽。基于摩尔线程自研 MTLink1.0 技术，MTT S4000 可以支持多卡互联，助力千亿大模型的分布式计算加速。同时，MTT S4000 提供先进的图形渲染能力、视频编解码能力和超高清 8K HDR 显示能力，助力 AI 计算、图形渲染、多媒体等综合应用场景的落地。尤为重要的是，借助摩尔线程自研 MUSIFY 开发工具，MTT S4000 计算卡可以充分利用现有 CUDA 软件生态，实现 CUDA 代码零成本迁移到 MUSA 平台。

- **KUAE 智算中心软硬一体解决方案发布**

摩尔线程 KUAE 智算中心解决方案以全功能 GPU 为底座，是软硬一体化的全栈解决方案，包括以 KUAE 计算集群为核心的基础设施、KUAE Platform 集群管理平台以及 KUAE ModelStudio 模型服务，旨在以一体化交付的方式解决大规模 GPU 算力的建设和运营管理问题。该方案可实现开箱即用，大大降低传统算力建设、应用开发和运维运营平台搭建的时间成本，实现快速投放市场开展商业化运营。

图表5：摩尔线程 KUAE 智算中心软硬一体解决方案

资料来源：摩尔线程官网，国联证券研究所

- **大模型训练效率提升**

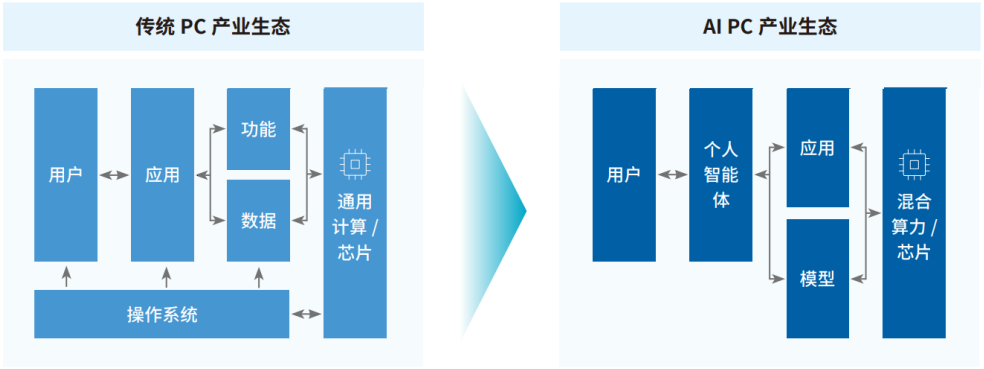
摩尔线程支持包括 LLaMA、GLM、Aquila、Baichuan、GPT、Bloom、玉言等各类主

流大模型的训练和微调。基于摩尔线程 KUAE 千卡集群，70B 到 130B 参数的大模型训练，线性加速比均可达到 91%，算力利用率基本保持不变。以 2000 亿训练数据量为例，智源研究院 700 亿参数 Aquila2 可在 33 天完成训练；1300 亿参数规模的模型可在 56 天完成训练。此外，摩尔线程 KUAE 千卡集群支持长时间连续稳定运行，支持断点续训，异步 Checkpoint 少于 2 分钟。

2.3 AIPC 提升端侧 AI 能力

联想集团联合 IDC 共同发布的《AIPC 产业（中国）白皮书》（以下简称“《白皮书》”）指出：在 AI PC 的推动下，PC 产业生态将从应用为本转向以人为本，从应用驱动转变为意图驱动。传统 PC 产业生态以操作系统为基础，用户在系统界面中直接进行操作，并管理和应用各式各样的应用程序。AI PC 产业生态中，个人智能体将成为第一入口，在大模型与应用生态的支持下，理解用户指令，给出反馈，跨应用进行调度，进而完成复杂任务。

图表6：传统 PC 与 AIPC 产业生态差异



资料来源：联想集团联合 IDC 共同发布的《AIPC 产业（中国）白皮书》，国联证券研究所

模型、应用、算力厂商都需要围绕 AI PC（终端）形态下新的以人为本的需求做出改变，在研发工作中对 AI 的高效运行予以充分的考量，以适应 AI PC 新时代。

➢ **联想持续推动 AI PC 的加速落地**

12 月 15 日，在 2023 英特尔新品发布会暨 AI 技术创新派对上，联想集团副总裁、中国首席战略官阿不力克木·阿不力米提正式公布：联想 ThinkPad X1 Carbon AI 、联想小新 Pro 16 AI 酷睿版两款 AI Ready 的 AI PC 产品，于当天 15:00 正式上市，预约预售同步开启。

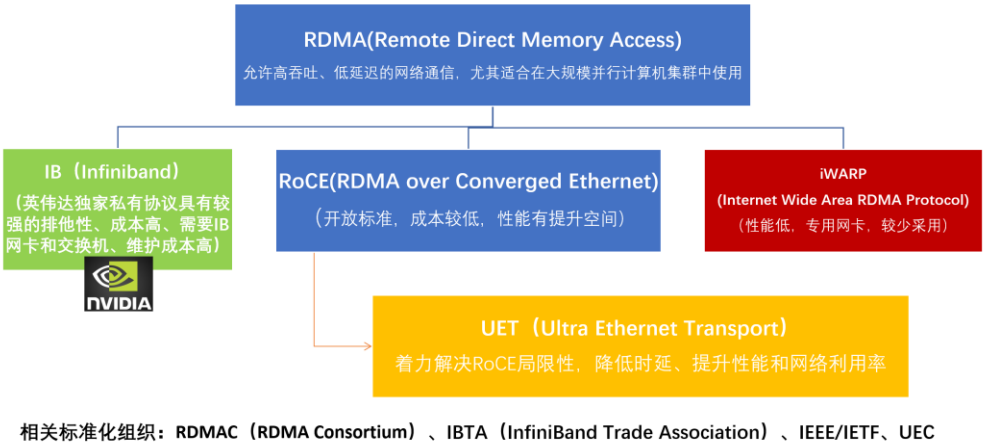
图表7：联想上市两款 AI PC 产品

资料来源：大联想伙伴网，国联证券研究所

2.4 算力网络：中国公司陆续加入超以太网联盟

目前 AI/HPC 训练的网络主要是基于 RDMA(Remote Direct Memory Access)，技术原理主要是允许 CPU、GPU、TPU 等加速器将数据直接从发送方内存传输到接收方内存。RDMA 的网络层协议有三种选择:分别是 InfiniBand(简称 IB)、iWarp(internet Wide Area RDMA Protocol)、RoCE（RDMA over Converged Ethernet)。

图表8：当前 AI 训练算力网络技术路线汇总



资料来源：星融元官网《网络融合大趋势下 RDMA 的发展演进过程》、UEC，国联证券研究所整理

InfiniBand 是一种专为 RDMA 设计的网络协议，由 IBTA（InfiniBand Trade Association)提出，从硬件级别保证了网络无损，具有极高的吞吐量和极低的延迟。但是 InfiniBand 交换机是英伟达旗下 Mellanox 提供的专用产品，采用私有协议，而绝大多数现网都采用 IP 以太网协议，采用 InfiniBand 无法满足互通性需求。同时封闭架构也存在厂商锁定的问题，由于未来需要大规模弹性扩展的业务系统，更多用户选择开放标准 RoCE。iWarp 由 IEEE/IETF 制定，由于性能较低，已经不是主流解决方案。

➢ **RoCE 用于未来 AI/HPC 网络的局限性**

RoCE（RDMA over Converged Ethernet）通过 Verbs API 表达的方式可追溯到上世纪末，由 InfiniBand 贸易协会（InfiniBand Trade Association，IBTA）进行标准化。随着人工智能模型规模扩大、通信模式及计算方法的多样化，传统基于 RoCE 的 RDMA 方案存在诸多问题：

1、RoCE 和 DCQCN 是拥塞控制算法，用于避免链路超限并提高速率。但 DCQCN 对其下方网络和负载性质敏感，需手动调整性能。未来的 AI 网络需要一种适用于任何数据中心的传输协议。

2、网络运营商在“无损”网络（InfiniBand 和 RoCE）上运行 RDMA 以避免此行为，但效率低。优先级流量控制（PFC）生成逐跳背压以太网是无损的，但背压（Back Pressure）传播导致拥塞树、队头阻塞、环路死锁等，使网络性能下降。PFC/ECN、DCQCN 需根据网络情况调整、操作和监控，成本高。未来的 AI 网络急需不依赖于无损结构的传输协议。

3、无论是在带宽还是对等点数量方面。Verbs API 设计规模已经捉襟见肘。RC（可靠连接）传输模式如果不减少快速路径状态，就不适合高速率下的高效硬件卸载。此外，固有的流程到流程（N*P*P)的可扩展性问题也是一大限制。这些问题还没有完

美的解决方案，而未来的 AI 网络需要能支撑 Verbs API 设计规模的传输协议。

4、AI 应用程序传输大量数据，受 NIC QP Scale 和 AI 模型数据交互方式限制，传统 RoCE 需仔细进行大象流负载均衡以防止链路过载。AI 工作负载决定了整个计算周期受限于所有流成功交付，而未来更高性能的 AI 网络需要改进的负载均衡技术。

➢ **超以太网联盟应运而生**

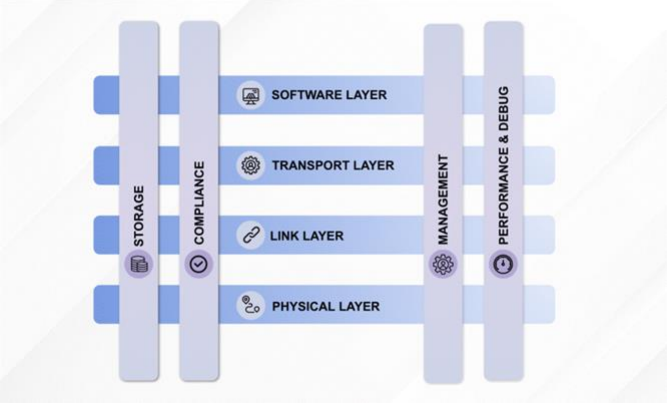
为了突破传统以太网的性能瓶颈，满足 AI 和高性能计算对智能算力日益激增的需求，超以太网联盟（Ultra Ethernet Consortium,UEC）于 2023 年 7 月在 Linux 基金会的牵头下由多家全球头部科技企业联合成立。UEC 早期创始成员包括 AMD、Arista、博通、思科、Eviden、HPE、Intel、Meta 和微软等全球头部科技企业。

图表9：超以太网联盟（UEC）早期创始成员

资料来源：UEC 官网，国联证券研究所

创始会员具备丰富的网络、人工智能、云和高性能计算大规模部署经验，将为 UEC 的四个工作组——物理层、链路层、传输层和软件层做出贡献。

图表10：超以太网联盟工作组设置



资料来源：新华三官网，UEC,，国联证券研究所

➢ **超以太网传输协议（UET）立足解决 RoCE 局限性**

以太网是 IP 网络最成熟和主流的技术，被广泛应用于园区网络、数据中心和云计算环境以及广域网等场景，但面对 AIGC 等新兴技术发展带来的算力需求大增，传统以太网将难以满足数据中心的性能需求。超以太网联盟（UEC）旨在创建一个“基于以太网的完整通信堆栈架构”,用超以太网传输协议取代基于以太网的 RoCE 协议，提高网络吞吐量、降低延迟，增强网络的可靠性和稳定性，为人工智能和高性能计算等领域的发展提供更加坚实的网络基础，同时保留以太网/IP 生态系统的优势。

图表11：超以太网传输协议（UET）主要改进方向

超级以太网主要改进方向
开放协议规范从一开始就设计为在 IP 和以太网上运行
数据包级别负载均衡
流控针对 AI 优化、无需配置
支持乱序包处理
低尾延迟
网内计算（INC）
大规模：可支持 1 百万个端点
无需拥塞算法调整提升性能和网络利用率
高带宽网络：800G/1.6T 和未来更高速端口

资料来源：UEC《 Overview of and Motivation for the Forthcoming Ultra Ethernet Consortium Specification》，国联证券研究所整理

UET 协议将超越传输层，定义标准语义层、改进的低延迟交付机制以及一致的 AI 和 HPC API，并提供标准的多供应商支持，以便通过 UEC 传输协议实现这些 API。为了实现全行业在互操作性方面的合作，UEC 构建了完整的基于以太网的通信堆栈架构，以最好地匹配快速发展的、大规模的 AI/HPC 工作负载，并提供一流的功能、性能、互操作性、TCO 以及开发人员和最终用户友好性。

➢ **中国企业陆续加入超以太网联盟**

中国网络设备公司如华为、新华三和锐捷网络，互联网巨头如阿里巴巴、百度、字节跳动、腾讯等在 2023 年下半年陆续加入 UEC 联盟。

图表12：中国企业陆续加入 UEC

资料来源：UEC 官网，国联证券研究所

3. 大模型能力不断提升

自 2022 年 11 月 30 日 ChatGPT 发布以来，AI 大模型在全球范围内掀起了有史以来规模最大的人工智能浪潮，国内厂商百度、华为、阿里、商汤、科大讯飞等积极参与大模型训练，陆续发布大模型产品。

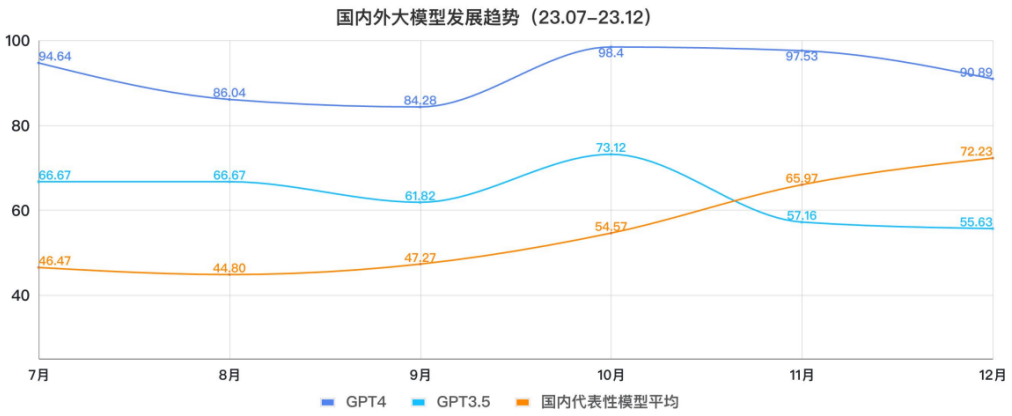
图表13：AI 大模型迅速发展

资料来源：SuperCLUE《中文大模型基准测评 2023 年度报告》，国联证券研究所

AI 大模型能力持续提升。11 月 7 日，OpenAI 举行首届开发者大会，推出了 GPT-4 Turbo，对六个方面进行了升级。12 月 6 日，谷歌正式向公众发布新一代大语言模型（LLM）Gemini，最大亮点之一为原生多模态大模型。

2023 年下半年，国内领军大模型企业持续追赶，每个月都有稳定且明显的提升，根据到 SuperCLUE 中文大模型基准测评结果，以文心一言、通义千问、ChatGLM 为代表的国内模型在 11 月已经完成总分上对 GPT3.5 的超越；12 月国内第一梯队模型与 GPT4 的差距在缩小。但仍有较大的距离需要追赶。

图表14：过去六个月国内外代表性模型的发展趋势



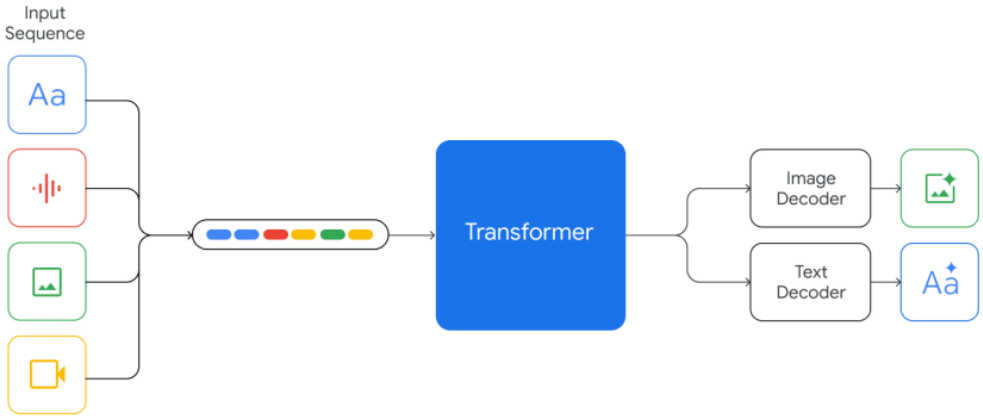
资料来源：SuperCLUE《中文大模型基准测评 2023 年度报告》，国联证券研究所

3.1 谷歌发布原生多模态大模型

12 月 6 日，谷歌正式向公众发布新一代大语言模型（LLM）Gemini，最大亮点之

一为原生多模态大模型。Gemini 模型设计时就原生支持多模态，要具有处理不同形式数据（语言+听力+视觉）的能力；一开始就在不同模态上进行预训练，利用额外的多模态数据进行微调以提升有效性。这有助于 Gemini 从头开始缝地理解和推理各种输入，优于现有的多模态模型。

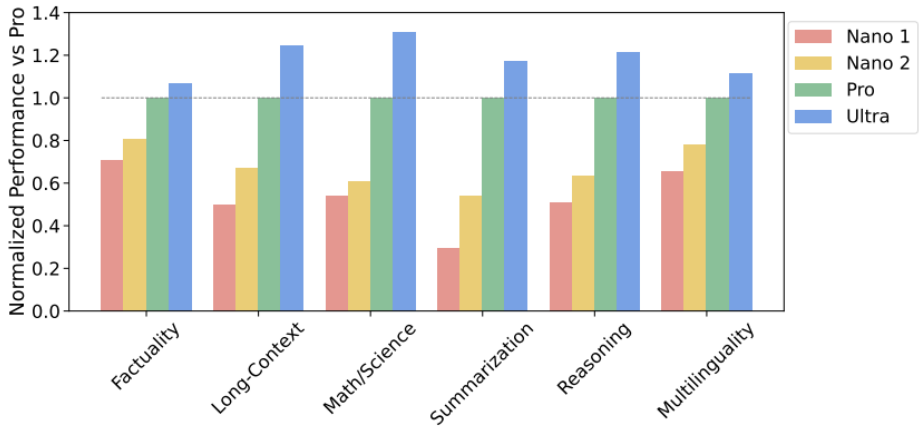
图表15：Gemini 具备原生多模态能力



资料来源：Google《Gemini：A Family of Highly Capable Multimodal Models》，国联证券研究所

Gemini1.0 具有三个不同尺寸。最轻量的版本 Gemni Nano 可以直接在智能手机上离线运行；相对而言更强大的版本 Gemini Pro 可以执行多种任务，将通过谷歌的类 ChatGPT 聊天机器人 Bard，为众多谷歌 AI 服务提供支持，加持谷歌的 Gmail、Maps、Docs 和 YouTube 等服务；功能最强大的版本 Gemini Ultra 也是谷歌迄今打造的最强大 LLM，主要为数据中心和企业应用而设计。

图表16：性能随模型大小的增加而增加



资料来源：Google《Gemini：A Family of Highly Capable Multimodal Models》，国联证券研究所

Gemini 文本性能领先。从自然图像、音频和视频理解到数学推理，Gemini Ultra 的性能在大型语言模型（LLM）研发中使用的 32 个广泛使用的学术基准中的 30 个上超过了当前最先进的结果。Gemini Ultra 的得分高达 90.0%，是第一个在 MMLU（大规模多任务语言理解）上超越人类专家的模型。

图表17：Gemini Ultra 文本性能领先

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflect-ion-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)						
GSM8K Grade-school math (Cobbe et al., 2021)	**94.4%** Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	**53.2%** 4-shot	32.6% 4-shot	52.9% 4-shot (via API**)	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8%	23.9% 4-shot	13.5% 4-shot
			50.3% (Zheng et al., 2023)						
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	**83.6%** 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	**74.4%** 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	**74.9%** 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	**82.4** Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	**95.3%** 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	**74.4** 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—

资料来源：Google《Gemini：A Family of Highly Capable Multimodal Models》，国联证券研究所

Gemini 多模态能力领先。Gemini 模型是原生多模态的，能将其跨模态的能力（如从表格、图表或图形中提取信息和空间布局）与语言模型的强大推理能力（如其在数学和编码方面的一流性能）无缝结合起来。在辨别输入中的细微细节、汇总跨时空的上下文以及将这些能力应用于与时间相关的视频帧和/或音频输入序列方面也表现出很强的性能。

图表18：Gemini Ultra 多模态性能领先

	Gemini Ultra (pixel only)	Gemini Pro (pixel only)	Gemini Nano 2 (pixel only)	Gemini Nano 1 (pixel only)	GPT-4V	Prior SOTA
MMMU (val) Multi-discipline college-level problems (Yue et al., 2023)	**59.4%** pass@1	47.9%	32.6%	26.3%	56.8%	56.8% GPT-4V, 0-shot
	62.4% Maj1@32					
TextVQA (val) Text reading on natural images (Singh et al., 2019)	**82.3%**	74.6%	65.9%	62.5%	78.0%	79.5% Google PaLI-3, fine-tuned
DocVQA (test) Document understanding (Mathew et al., 2021)	**90.9%**	88.1%	74.3%	72.2%	88.4% (pixel only)	88.4% GPT-4V, 0-shot
ChartQA (test) Chart understanding (Masry et al., 2022)	**80.8%**	74.1%	51.9%	53.6%	78.5% (4-shot CoT)	79.3% Google DePlot, 1-shot PoT (Liu et al., 2023)
InfographicVQA (test) Infographic understanding (Mathew et al., 2022)	**80.3%**	75.2%	54.5%	51.1%	75.1% (pixel only)	75.1% GPT-4V, 0-shot
MathVista (testmini) Mathematical reasoning (Lu et al., 2023)	**53.0%**	45.2%	30.6%	27.3%	49.9%	49.9% GPT-4V, 0-shot
AI2D (test) Science diagrams (Kembhavi et al., 2016)	**79.5%**	73.9%	51.0%	37.9%	78.2%	81.4% Google PaLI-X, fine-tuned
VQAv2 (test-dev) Natural image understanding (Goyal et al., 2017)	**77.8%**	71.2%	67.5%	62.7%	77.2%	86.1% Google PaLI-X, fine-tuned

资料来源：Google《Gemini：A Family of Highly Capable Multimodal Models》，国联证券研究所

Gemini 将逐步应用于谷歌产品。Bard 将使用 Gemini Pro 的微调版本来进行更高级的推理、规划和理解等；Pixel 8 Pro 是首款搭载 Gemini Nano 的智能手机，可以支持录音应用中的“总结”等新功能，并在 Gboard 中推出“智能回复”功能；未来几个月，Gemini 将应用于谷歌更多的产品和服务，如搜索、广告、Chrome 和 Duet

AI。谷歌已经开始在搜索中试验 Gemini，使得用户的搜索生成体验(SGE)更快，美国地区的英语搜索的延迟减少了 40%，同时质量也得到了提高。

原生多模态有望成为大模型发展的重要方向。此前创建多模态模型的标准方法是针对不同模态训练单独的组件，然后整合在一起。这些模型有时擅长执行部分任务，例如描述图像，但难以处理更概念性和复杂的推理。原生多模态能力意味着模型能够更自然、高效地处理和融合多种类型的数据，这在实现更复杂的 AI 应用方面具有重要意义，有望成为大模型发展的重要方向。

3.2 OpenAI 公布 2024 年计划

11 月 7 日，OpenAI 举行开发者大会，推出了 GPT-4 Turbo，对六个方面进行了升级：上下文长度、更多的控制、更好的知识、多模态、定制化、更高的速率限制。此外 OpenAI 还降低了模型使用成本，新发布的 GPT-4 Turbo 输入方面降至 GPT-4 的 1/3 即 1 美分/1000 tokens，输出方面降至 1/2 即 3 美分/1000 tokens；GPT-3.5 Turbo 16k 等也进行了降价。OpenAI 的性能升级、收费降低有助于推动大模型应用商业化进程。

12 月 24 日，OpenAI 联合创始人兼首席执行官 Sam Altman 在社交平台公布，AGI、GPT-5、更好的语音模型、更高的费率限制、更好的 GPTs、更好的推理能力、对唤醒/行为程度的控制、视频模型、个性化、更好地浏览、使用 OpenAI 登录、开源，将是 OpenAI 在 2024 年要实现的目标。

图表19：OpenAI 公布 2024 年计划

资料来源：每日经济新闻，国联证券研究所

3.3 国内大模型日益成熟

12 月 21 日，医渡科技发布自主研发的医疗垂域大模型，在分导诊、基础医学、全科医学、临床内科、临床外科、执业资格考试等多个医疗明确任务场景上的评测表现已经超过 GPT3.5，将赋能城市级医疗资源综合监管、传染病监测预警与应急处置

以及 C 端健康服务等多个场景，以大模型重构公司已有的产品和解决方案。此外，公司与华为正式签署合作协议，并发布智慧医疗领域医渡科技大模型训推一体机、智慧医院和智慧健康城市大数据平台联合解决方案，加速推动医疗健康行业的智能化。

图表20：医渡科技大模型在分导诊等医疗场景下表现优于 GPT3.5

资料来源：医渡科技官微，国联证券研究所

12 月 22 日，全国信息技术标准化技术委员会人工智能分委会全体会议公布了全国首个官方“大模型标准符合性测试”结果：百度文心一言、腾讯混元大模型、360 智脑、阿里云通义千问四款国产大模型率先通过测试，符合《人工智能大规模预训练模型 第 2 部分：评测指标与方法》语言大模型的相关技术要求，通用性、智能性等维度达到国家相关标准。

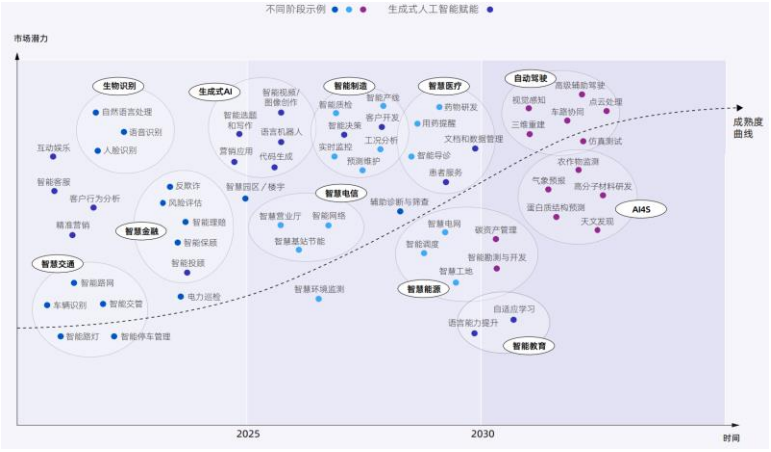
12 月 28 日，在百度举办的“第十届 WAVE SUMMIT 深度学习开发者大会”上，百度首席技术官、深度学习技术及应用国家工程研究中心主任王海峰宣布，百度文心一言用户规模已突破 1 亿。截至 2023 年 12 月底，飞桨已凝聚 1070 万开发者，服务 23.5 万家企事业单位，基于飞桨创建了 86 万个模型。

根据央视新闻 12 月 16 日报道，工业和信息化部赛迪研究院数据显示，目前，我国已有超过 19 个语言大模型研发厂商，其中，15 家厂商的模型产品已经通过备案。预计 2023 年，我国语言大模型市场规模将达到 132.3 亿元，增长率将达到 110%。

4. 文生视频引领 AIGC 应用

人工智能技术已全面迈入应用加速时代，在智慧交通、智慧金融、生物识别等领域已经被广泛应用，在服务、医疗、金融、制造能源、零售、教育等领域已经实现落地。2023 年生成式人工智能的快速发展，给各行业带来新的动能。

图表21：2023 中国人工智能应用场景成熟度曲线



资料来源：IDC 《2023-2024 年中国人工智能计算力发展评估报告》，国联证券研究所

➢ **AIGC 快速发展，或将快速赋能各行各业**

当前 AIGC 从技术创新到应用推广的周期相较于以往的技术大大缩短。AIGC 目前已在文字、图像、音频、视频等领域得到应用。在媒体、广告行业、游戏等行业，可以大幅提升生产力，降低制作成本；在消费零售行业，可以提升销售效率和客户满意度；在高端制造行业，可以帮助生成产品设计、模拟和测试数据，从而提高产品质量和生产效率。随着技术的不断发展和进步，或将出现更多新的应用场景。当前国内外 AIGC 创业如火如荼，诞生了一大批创新应用，如 ChatGPT、微软 Copilot、Midjourney、文心一言等，并且已经陆续商业化。

图表22：AIGC 国内外热门应用汇总

资料来源：腾讯研究院，红杉资本，国联证券研究所整理

4.1 文生视频领域进展不断

11 月 3 日，Runway 的 Gen-2 发布更新，支持 4K 超逼真的清晰度作品，并随后上线运动笔刷新功能；11 月 16 日，Meta 发布 Emu Video；11 月 18 日，字节发布 PixelDance；11 月 21 日，Stable AI 推出 Stable Video Diffusion；11 月 23 日，Adobe 宣布完成对专注于 AI 视频创作的初创企业 Rephrase.ai 的收购。

11 月 28 日，Pika labs 正式发布了其全新的文生视频产品 Pika 1.0，可以通过输入简单的文本或上传图像，即可创建简短、优质的视频；可以扩展视频的画布或宽高比；可以用 AI 编辑视频局部内容；可以使用 AI 扩展现有视频剪辑的长度。12 月 26 日，Pika 宣布 Pika 1.0 网页端访问权限向所有用户开放，且所有用户都可以免费使用。

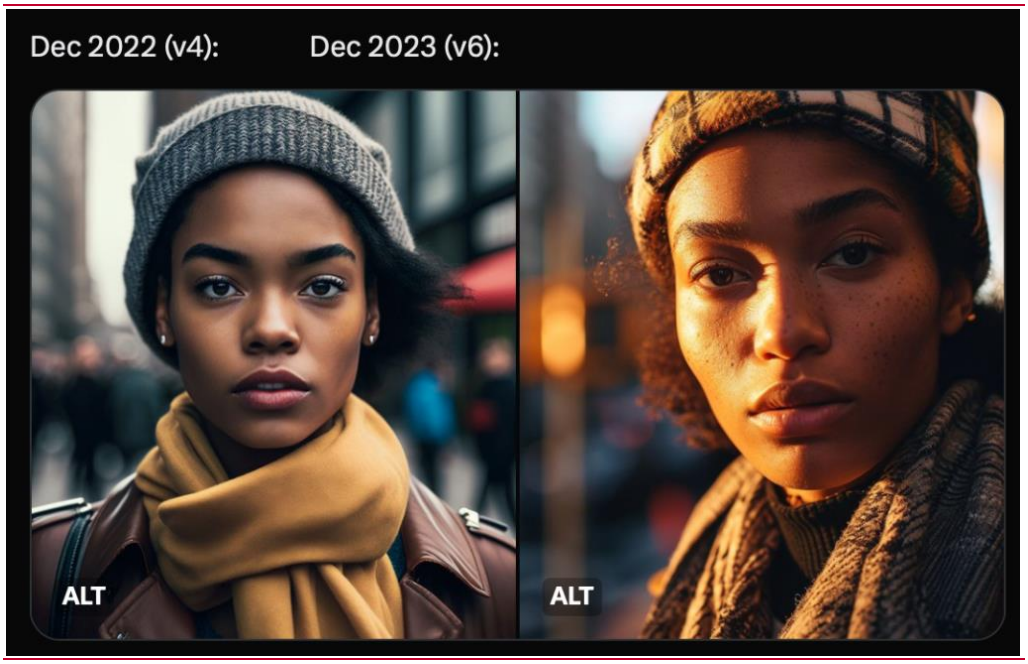
图表23：Pika 可以通过简单文本输入生成视频

资料来源：Pika 官网，国联证券研究所　输入为 Handcrafted wooden Christmas ornaments swinging lightly on a lush green tree with twinkling lights

4.2 文生图领先模型 Midjourney 发布 V6 版本

12 月 21 日，Midjourney 在 Discord 上宣布其最新版本 V6 的测试版发布，目前处于 alpha 测试阶段。V6 版本具有图像质量更好、语义理解更强、能嵌入英文单词、容纳更多 token 等进步。

图表24：Midjourney V6 在图像质量等方面提升

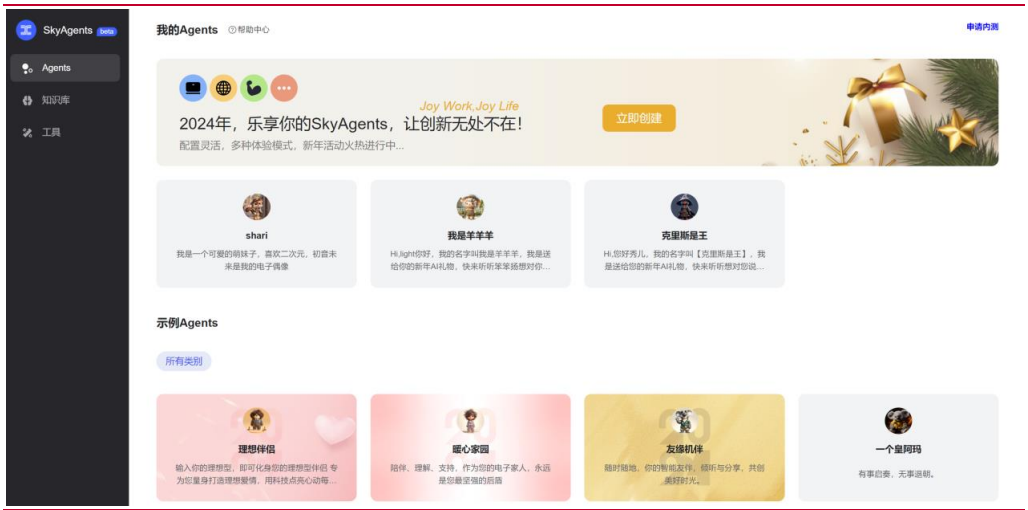


资料来源：Midjourney 官网，国联证券研究所　输入为：street style closeup of a woman in New York City, shot on porta 400, early morning –style raw –ar 4:5 –v 6.0

4.3 智能体领域天工 SkyAgents 开放测试

12 月 25 日，昆仑万维 AI Agents 开发平台“天工 SkyAgents”Beta 版正式开放测试。昆仑万维“天工 SkyAgents”AI Agents 开发平台，基于昆仑万维“天工大模型”打造，具备从感知到决策，从决策到执行的自主学习和独立思考能力。用户可以通过自然语言构建自己的单个或多个“私人助理”，并能将不同任务模块化，通过操作系统模块的方式，实现执行包括问题预设、指定回复、知识库创建与检索、意图识别、文本提取、http 请求等任务。

图表25：昆仑万维“天工 SkyAgents”Beta 版正式开放测试



资料来源：中国网科技，国联证券研究所

5. 数据合规风险日益凸显

《纽约时报》过去数月一直在与 OpenAI 和微软就“内容付费”进行商谈，但双方并未达成协议。12 月 27 日，美国《纽约时报》将开放人工智能研究中心（OpenAI）和微软告上法庭，指控这两家公司未经授权使用该媒体数以百万计文章训练 ChatGPT 等人工智能（AI）聊天机器人，要求它们停止使用其内容训练 AI 模型并销毁训练数据。这场版权官司于 27 日被纽约曼哈顿联邦法院受理。起诉书显示，涉事公司“非法使用”《纽约时报》的原创内容训练自动聊天机器人，与该新闻机构产生了直接竞争的关系。起诉书表示，若相关新闻机构无法保护其独立报道，原创新闻报道会随之减少，届时“社会将出现计算机和 AI 无法填补的真空”。

《纽约时报》在诉讼中还提到了 AI 语言模型的另一个通病——“AI 幻觉”，即 AI 会生成并传播虚假、无意义或令人反感的内容。比如说，微软必应上的聊天机器人曾罗列过“15 种有利于心脏健康的食物”，并将信源指向《纽约时报》，但这 15 种食物中有 12 种都未被原报道提及。

我们认为诉讼结果将会对 AI 新兴法律框架带来深远影响。

6. 投资建议

我们认为应该重视 AI 技术创新对全社会生产力提升的长期价值。建议关注产业链中国产算力基础设施、模型商业化、行业应用三条主线。

6.1 国产算力基础设施

建议关注：海光信息、寒武纪-U、中科曙光、浪潮信息、紫光股份、锐捷网络等；

6.2 AI 模型商业化

建议关注：百度、科大讯飞、商汤、拓尔思等；

6.3 AI 应用

建议关注：**（1）音视频：**万兴科技、海康威视、大华股份等；**（2）图像：**美图公司、虹软科技等；**（3）办公：**金山办公、福昕软件、泛微网络、用友网络等；**（4）垂直领域：**同花顺、恒生电子、宇信科技、中科软、卫宁健康、医渡科技、中科创达等。

7. 风险提示

AI 技术发展演进不及预期；商业化进程不及预期；法律政策监管风险；行业竞争加剧。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

评级说明

投资建议的评级标准		评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 到 12 个月内的相对市场表现，也即：以报告发布日后的 6 到 12 个月内的公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中:A 股市场以沪深 300 指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以摩根士丹利中国指数为基准；美国市场以纳斯达克综合指数或标普 500 指数为基准；韩国市场以柯斯达克指数或韩国综合股价指数为基准。	股票评级	买入	相对同期相关证券市场代表指数涨幅 20%以上
		增持	相对同期相关证券市场代表指数涨幅介于 5%~20%之间
		持有	相对同期相关证券市场代表指数涨幅介于-10%~5%之间
		卖出	相对同期相关证券市场代表指数跌幅 10%以上
	行业评级	强于大市	相对同期相关证券市场代表指数涨幅 10%以上
		中性	相对同期相关证券市场代表指数涨幅介于-10%~10%之间
		弱于大市	相对同期相关证券市场代表指数跌幅 10%以上

一般声明

特别声明

版权声明

联系我们

北京：北京市东城区安定门外大街 208 号中粮置地广场 A 塔 4 楼

上海：上海市浦东新区世纪大道 1198 号世纪汇二座 25 楼

无锡：江苏省无锡市金融一街 8 号国联金融大厦 12 楼

深圳：广东省深圳市福田区益田路 6009 号新世界中心大厦 45 楼

电话：0510-85187583