

2025年万物AI面临的十大待解难题

评级：强于大市

长城证券产业金融研究院

科技首席：唐泓翼

执业证书编号：

S1070521120001

2024年12月09日

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

一、AI终端需求端侧问题

- AI服务器： scaling law 结束了吗？
- AI终端： 万物+AI 能否崛起？
- AI机器人： 人型机器人启动？

二、AI上游芯片需求问题

- AI 产能： AI大芯片带动需求激增，先进制程/先进封装产能仍供不应求？
- AI 存储： AI大芯片未来将强劲推动HBM扩产需求，NAND存储需求持续升级？
- AI 互联： AI大芯片的互联困境，未来趋势将是“光进铜退”？
- AI 散热： AI大芯片功耗提升，散热终将由风冷向液冷演进？

三、AI发展外部约束问题

- AI能源： 能源吞噬巨兽？
- AI燃料： 高质量的数据快要耗尽了？

四、AI终极问题

- AI终极： 该怎么面对凡人？

五、投资建议及风险提示：

- 端侧AI一触即发，拥抱第四次工业革命，关注AI产业链相关公司
- 手机和PC需求温和复苏，看好“困境反转&龙头低估”企业
- 风险提示： 下游需求不足风险；地缘政治风险；核心竞争力风险；估值风险等。

一、AI终端需求端侧问题

AI服务器： scaling law 结束了吗？

AI初期：小量计算

- 小型到中型的机器学习模型。
- 依赖于单个或少量GPU。

Transformer初现：大模型时代

- 引入了Transformer架构，开启了大模型，时代。
- 服务器需求更大规模的模型训练。

Scaling Law初现：初级智慧涌现

- OpenAI的论文《Scaling Laws for Neural Language Models》
- 提出了模型性能与规模之间的幂律关系。

预训练Scaling Law

- DeepMind的研究提出了与OpenAI不同的观点，强调模型大小和训练token数量的关系

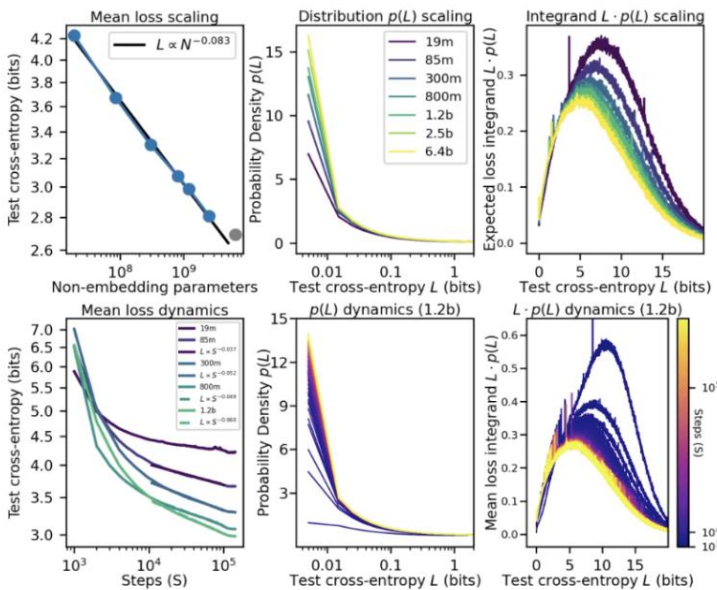
Scaling Law撞墙，推理延续

- Scaling law收益下降，本质是优质数据和底层逻辑
- OpenAI的o1模型展示了推理阶段的Scaling Law

Scaling Law撞墙后新趋势

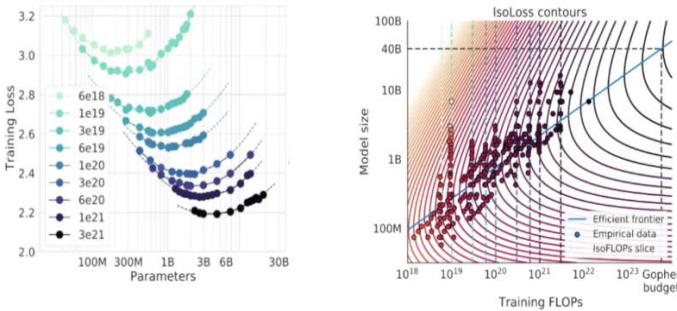
- 从训练到推理的转变 - 数据扩展与合成数据
- 多模态和跨领域应用 - 能源和效率等其他方案

图：三张图神经尺度， scaling law后的涌现现象



图： Scaling law 撞墙后，模型效果提升速度下降

Scaling Law – Optimal Setting

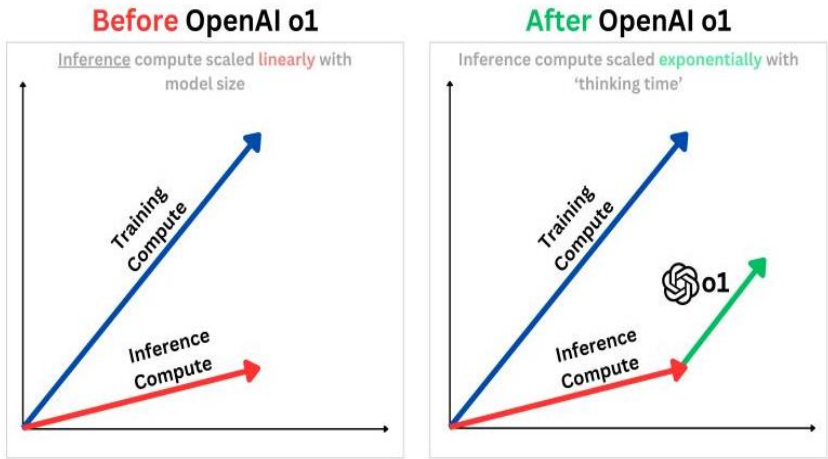


- 数据越多并不是模型效果不提升了，只是从计算角度来看不经济了

图：新的scaling law--推理

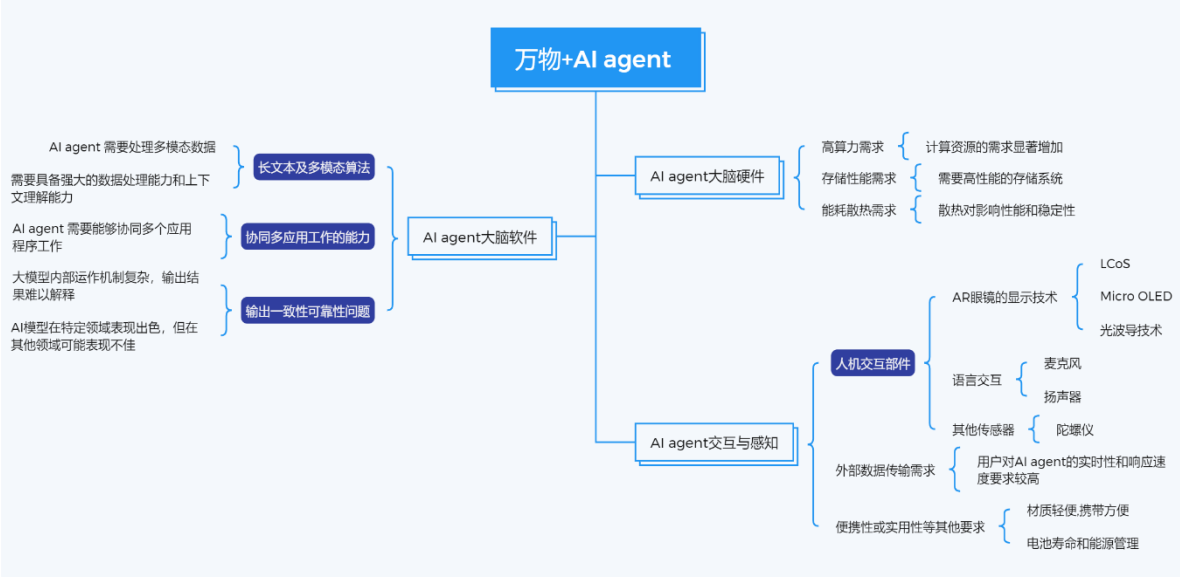
New scaling law: why OpenAI's o1 model matters

OpenAI created a new way to scale - through more compute during generation

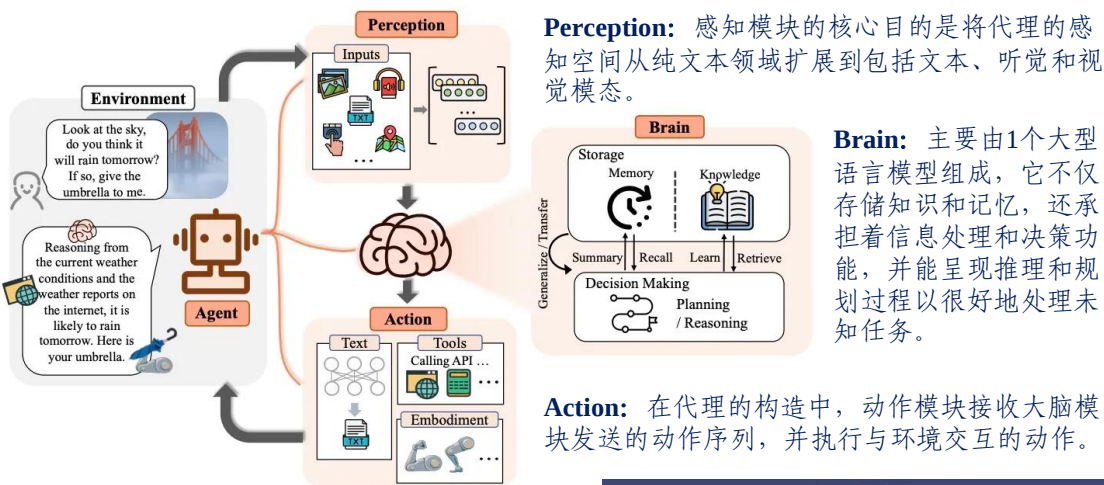


AI终端：万物+AI 能否崛起？

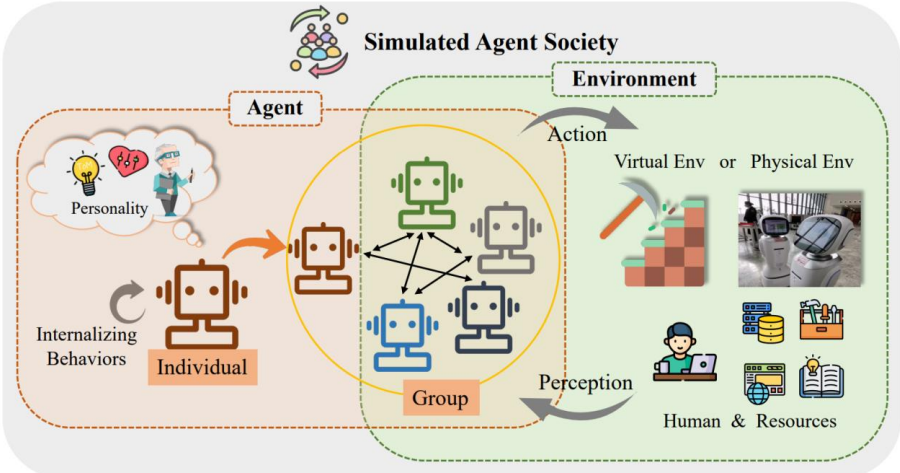
- **AI Agent崛起**：11月20日，微软在Microsoft ignite 2024上全面推出商用AI Agents生态系统，企业用户可通过Azure AI目录访问超过1800个AI模型，用于支持各类AI Agent的部署和运行。据Gartner预测，到2028年，至少15%的日常工作决策将通过AI Agent自主做出。据marketsandmarkets预测，到2028年，全球自主人工智能和自主智能体市场规模将达285亿美元，预计23-28年CAGR~43%。
- **AI手机**：据IDC预测，2024年全球AI手机出货量将达2.34亿部，预计到28年将达9.12亿部，23-28年CAGR~78.4%。
- **AI PC**：据Canalys预测，2024年全球AI PC出货量将达0.48亿台，预计到28年AI PC出货量将达2.05亿台，24~28年CAGR~44%。
- **AI耳机**：10月10日，字节跳动豆包发布首款AI智能体耳机Ola Friend，耳机接入豆包大模型，与豆包APP深度结合。



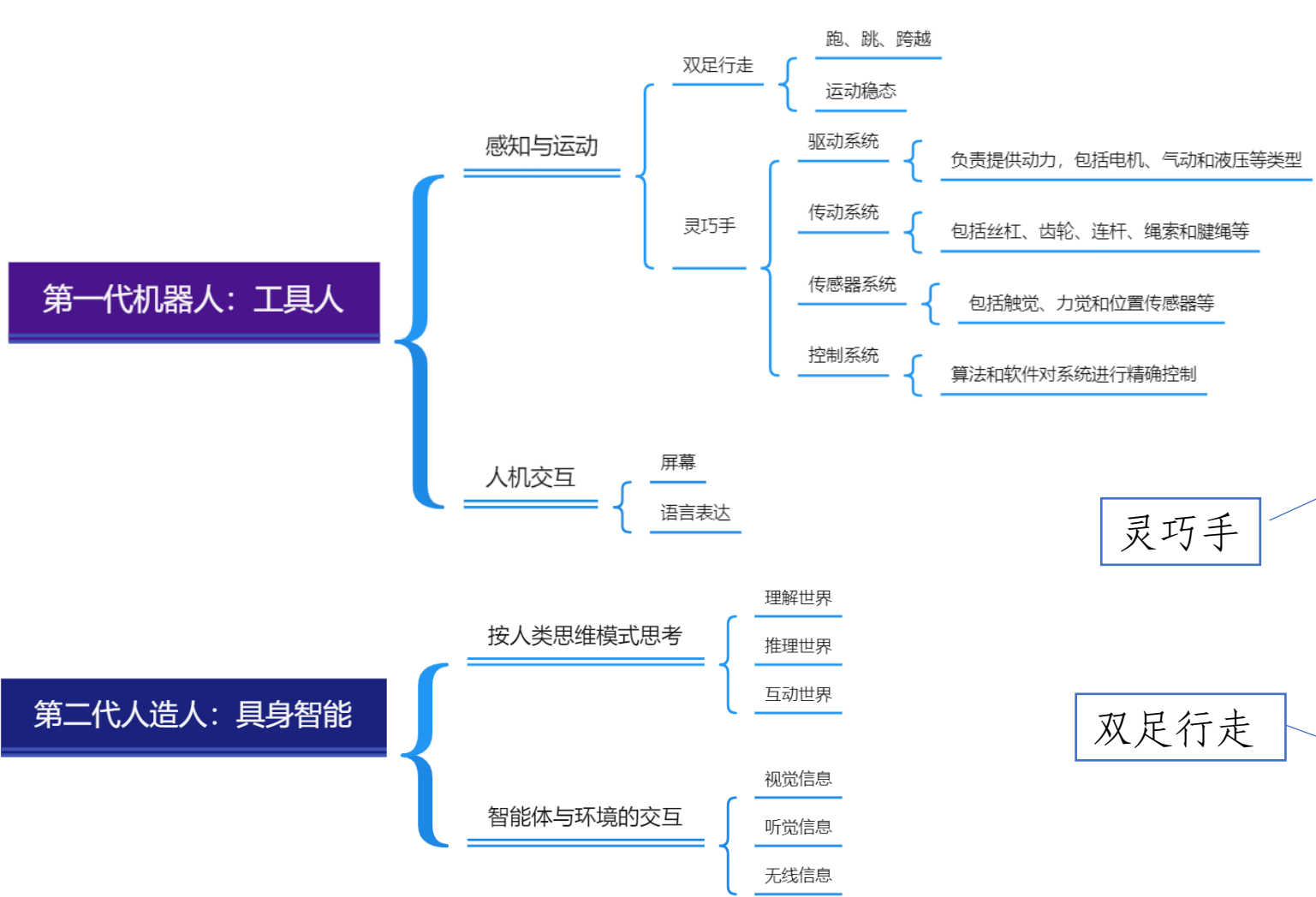
图：AI Agent的整体框架由三个关键部分组成：大脑(Brain)、感知(Perception)和行动(Action)。



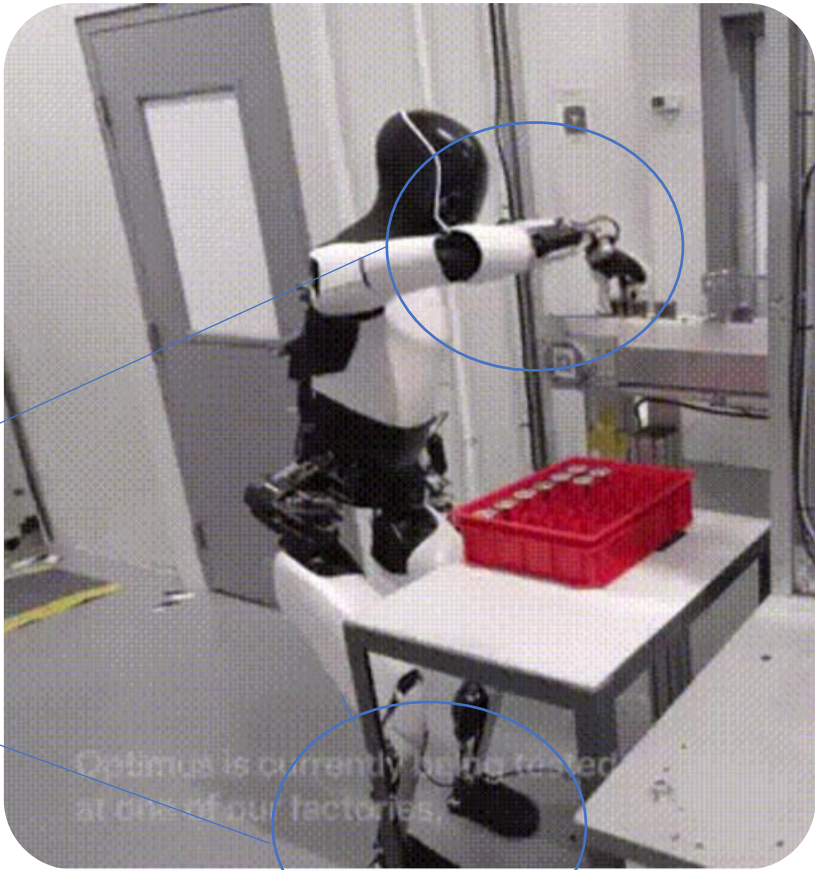
图：AI Agent 像人类一样行动、作出决定并参与人类社会活动



AI机器人：人型机器人启动？



具身智能



灵巧手

双足行走

二、AI上游芯片需求问题

AI 产能：AI大芯片带动需求激增，先进制程/先进封装产能仍供不应求？

先进制程：预计25H1台积电3nm/5nm产能利用率将达100%/101%

- ▶ **先进制程：**据TrendForce预估，2025年全球晶圆代工产值将迎来20%的成长，台积电表现仍将一支独秀，其余晶圆代工厂也可望有近12%的年成长；受AI服务器相关芯片需求推动下，先进制程高产能利用率有望延续至2025年。
- ▶ **据台积电24Q3业绩交流会上表示，AI技术的迅速发展是推动业绩增长的关键因素之一，尤其是针对3nm和5nm等先进制程芯片的需求激增，来自主要客户如苹果和英伟达的订单持续增加。从产品制程营收占比看，公司 24Q3 7nm及以下先进制程营收占比达69%，其中5nm制程营收占比为32%，7nm制程占比为17%，而得益于苹果A18系列处理器的强劲出货量，3nm制程营收占比提升至20%。据台媒《工商时报》报道，25H1台积电3nm产能利用率维持满载；在AI芯片助攻下，25H1台积电5nm产能利用率可能达到101%。**

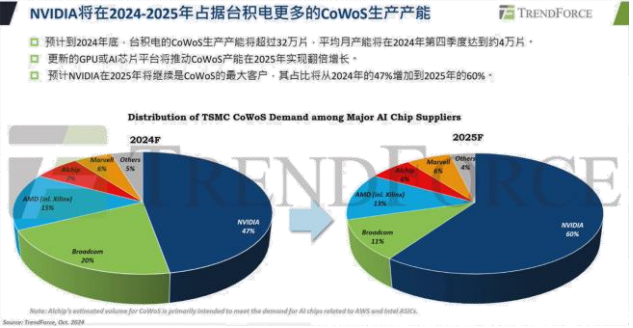
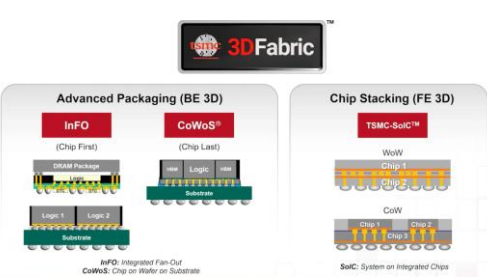
图：先进制程工艺不断演进，预计2025年上半年台积电3nm/5nm产能利用率维持满载



先进封装：预计25年台积电CoWoS月产能有望达8万片

- ▶ **先进封装：**受AI芯片大面积需求带动，2.5D先进封装于2023至2024年供不应求情况明显，据TrendForce预估，2025年晶圆代工厂配套提供的2.5D封装营收将年增120%以上。工研院产业科技国际策略发展所预测，到2025年全球先进封装市场规模比重将达到51%，首次超过传统封装，预计到2028年，先进封装市场年复合增长率可达10.9%。
- ▶ **台积电董事长魏哲家在法人说明会上表示，客户对先进封装的需求远大于供应，预计25年CoWoS产能将持续倍增。据Money DJ数据，台积电CoWoS月产能预计24年将达到 3.5~4万片晶圆，明年将飙升至每月8万片。AI客户强劲而紧迫的需求继续推动产能需求，随着更多设备的增加，台积电的 CoWoS 产能仍可能大幅扩张，到 2026 年可能达到每月 1.4~1.5万片晶圆。**

图：以CoWoS为主的先进封装需求逐步增长，预计2025年台积电CoWoS月产能将增长至8万片



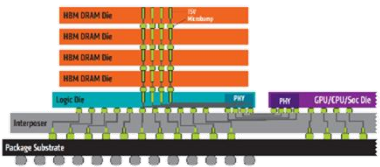
AI 存储：AI大芯片未来将强劲推动HBM扩产需求，NAND存储需求持续升级？

HBM：AI强劲需求带动下，25年全球HBM市场将达199亿美元

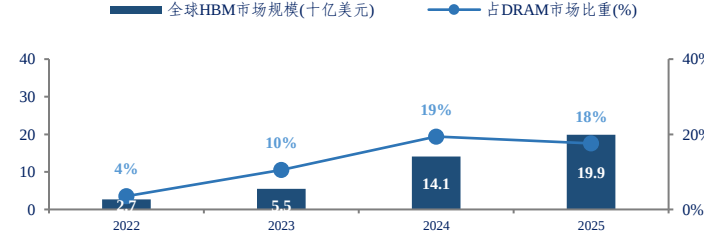
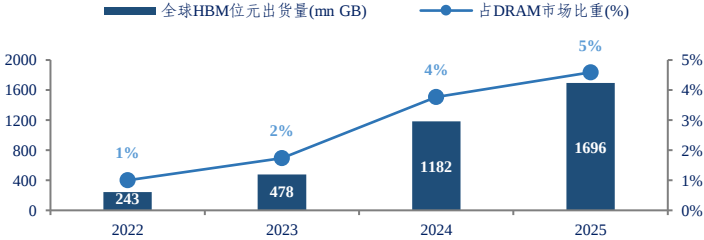
- **AI驱动HBM需求容量增长：**AI服务器强劲需求带动下，TrendForce预计2025年HBM位元需求量将同比增长117%。
- **HBM市场需求：**据Yole预测，到2025年，预计全球HBM位元出货量将达16.96亿GB，预计全球HBM市场规模将达199亿美元。
- **HBM当前进展：**11月5日，在SK AI Summit 2024上，SK 海力士发布业界首款16层堆叠(16Hi)的HBM3E内存(容量为48GB)，其AI训练性能和推理性能分别提升18%和32%，预计将于2025年初提供样品。SK海力士还计划25H2将推出首批12Hi HBM4，2026年将推出16Hi HBM4。

图：HBM堆叠层数将不断增加，预计2025年全球HBM市场规模达199亿美元

①HBM堆叠结构



②SK海力士12Hi HBM3E产品

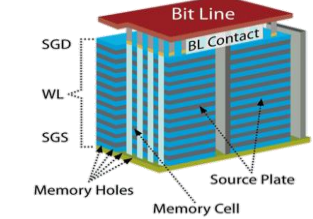


NAND：AI浪潮推动下，预计25年全球NAND市场将达870亿美元

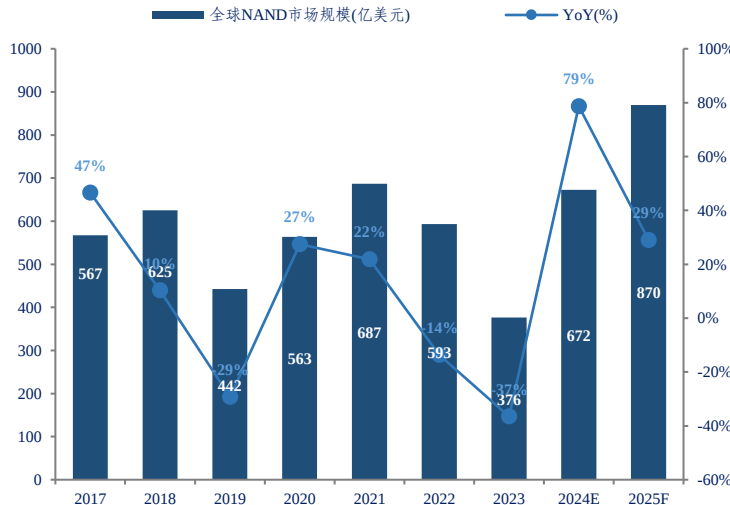
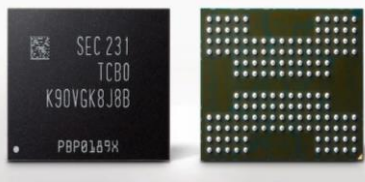
- **AI驱动NAND单机需求容量增长：**受益AI需求驱动，TrendForce预计2025年手机/笔电/服务器上单机需求SSD容量将同比+13.3%/+4.4%/+14.4%。
- **NAND市场需求：**据TrendForce预测，到2025年，预计全球NAND位元需求量将同比+13.4%；预计全球NAND Flash市场规模将同比+29%至870亿美元。
- **NAND当前进展：**9月12日，三星宣布已量产1Tb QLC第9代V-NAND，据TechInsights数据，三星计划25H2推出430层第10代V-NAND；11月21日，SK海力士宣布开发出业界首款321层1TB TLC NAND闪存，将于25H1开始供应。

图：3D NAND Flash层数将不断增加，预计2025年全球NAND市场规模达870亿美元

①3D NAND Flash结构



②三星3D NAND Flash产品



AI 互联：AI大芯片的互联困境，未来趋势将是“光进铜退”？

PCB连接： 预计28年市场规模达138亿美元

- **PCB连接趋势：** 英伟达对服务器规格的提升，使得PCB和HDI零部件的需求进一步增加，新规格产品需要更多高端HDI产品来加快运算和连接速度。
- **市场规模：** 据Prismark数据，预计2028年全球服务器及存储用PCB市场规模约为138亿美元，2023-2028年CAGR约11%，随着AI服务器升级，GPU主板将逐步升级为HDI，据Prismark预计，2023-2028年HDI的CAGR将达到16.3%，是增速最快的品类。

铜连接： 预计28年市场规模达28亿美元

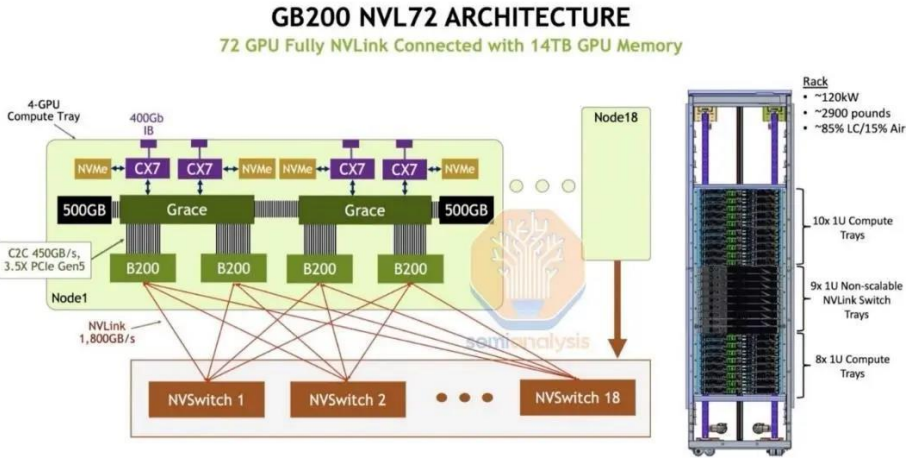
- **铜连接趋势：** 英伟达于3月发布GB200系列机架，其中背板连接、近芯片连接及机柜间I/O连接均用到铜连接方案。除英伟达外，高速铜互联在AI短距离场景已有成熟经验，dojo/谷歌等均使用定制铜缆或 DAC&AEC作为短距互联方案。在短距场景下，高速铜连接相较光纤连接具有性价比、稳定性、功耗优势，正成为AI集群短距传输优选方案。
- **市场规模：** 据Lightcounting数据，预计2028年全球高速电缆市场规模将达28亿美元。

光连接： 预计27年市场规模逾200亿美元

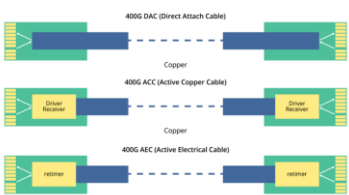
- **光模块、光芯片连接趋势：** 据LightCounting，高性能计算对网络速率的需求是目前的10倍以上，CPO能将现有可插拔光模块架构的功耗降低50%，并有效减少尺寸，有望成为未来主流。
- **市场规模：** 根据Lightcounting预测，光模块的全球市场规模在2022-2027年或将以CAGR~11%保持增长，2027年有望突破200亿美元。根据Lightcounting和Coherent预测，全球数通光模块市场23年-28年的CARG为18%，其中，AI用数通光模块市场CAGR为47%。

图： AI互联行业趋势：光进铜退

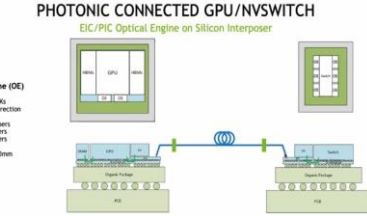
英伟达GB200 NVL72系统架构：



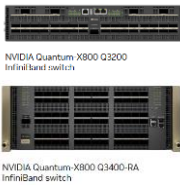
① 铜互连



② 光模块



③ 互联网InfiniBand



④ PCB连接



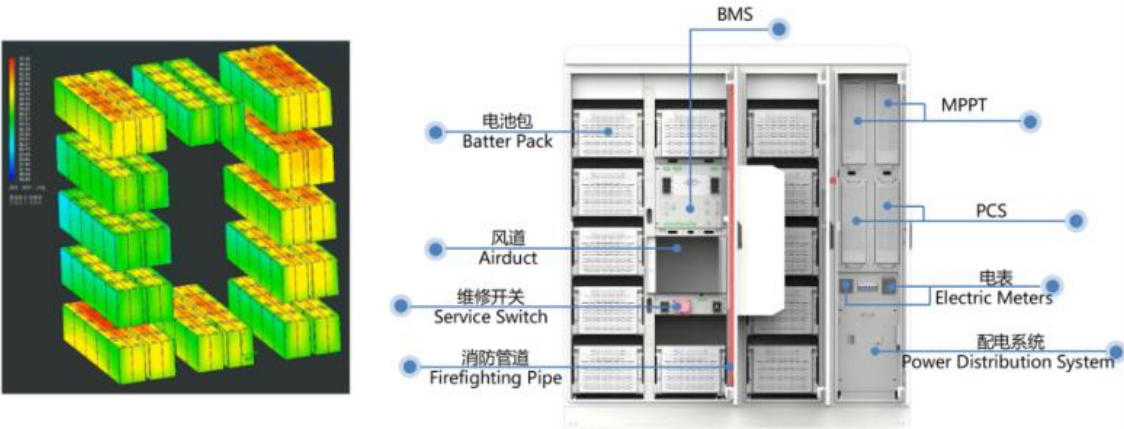
AI 散热：AI大芯片功耗提升，散热终将由风冷向液冷演进？

风冷：预计28年风冷市场规模将达85亿美元

- **风冷散热：**据兰洋科技官网，风冷系统中，风机转速从1000r/min提高到4000r/min，芯片散热中对流占主导，流速增加对流换热系数显著增加，风冷方式能有效改善芯片散热问题。据浪潮信息，传统风冷数据中心PUE(PUE=数据中心总能耗/IT设备能耗，PUE越接近1表示能效越高)约1.4-1.5。风冷方案优势主要有：成本低，维护方便，应用时间长，技术成熟；但存在噪音、有功率限制。
- **风冷市场规模：**据Dell'Oro Group数据，2028年数据中心热管理市场规模(风冷+液冷)将达120亿美元，23-28年CAGR为14%，风冷规模将达85亿美元。

图：风冷：数据中心目前的主流散热方案，预计28年市场规模将达85亿美元

- **原理：**通过散热器将CPU发出的热量转移至散热模块，通过风扇将热气吹走

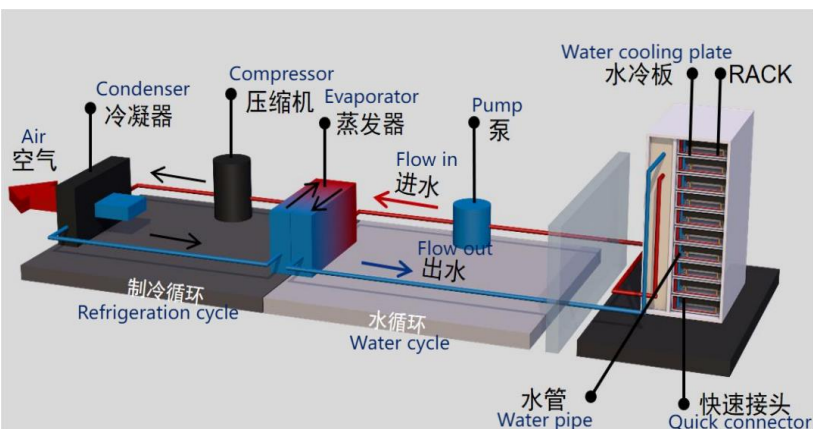


液冷：预计28年液冷市场规模将达35亿美元，渗透率持续提升

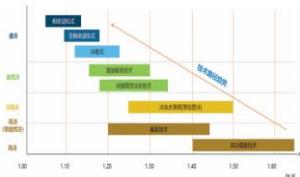
- **液冷散热：**据《数据中心液冷解决方案》白皮书，当机架密度达到20kW以上时，风冷效率减弱，需要采用液冷技术满足高热密度机柜的散热需求(以英伟达新推出的GB200 NVL72机柜方案为例，其热设计功耗高达约140kW)。据浪潮信息，液冷PUE约近1.1，具备较高能效优势。液冷方案优势主要有：安静，降温效率更高，功耗更低；但其约束在于成本高，维护困难。
- **液冷市场规模：**据TrendForce预估，2025年随着GB200机柜方案正式放量出货，将带动AI芯片液冷散热渗透率，从24年的11%提升至25年的24%。据Dell'Oro Group数据，2028年液冷规模将达35亿美元，占热管理支出的近1/3。

图：液冷：未来渗透率有望持续提升，预计28年市场规模将达35亿美元，占比提升至33%

- **原理：**通过液体流通带走热量以降温达到散热目的



技术路径：从能效角度，相较风冷，液冷散热的PUE值更能逐渐趋近于1



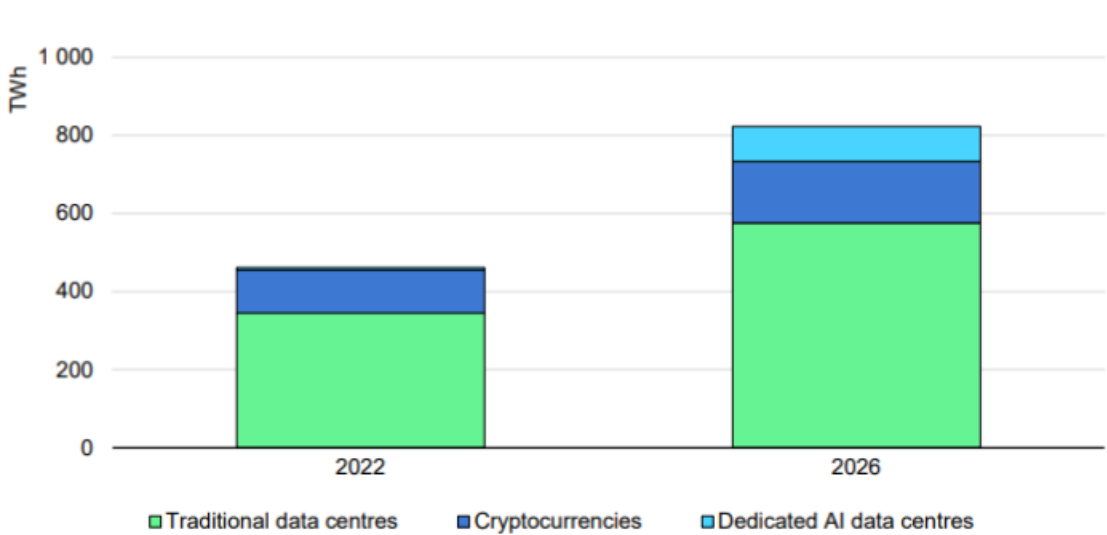
三、AI 发展外部约束问题

AI能源：能源吞噬巨兽？

AI能源吞噬巨兽

- AI数据中心吞噬能源：AI模型的训练和运行需要大量的计算资源。例如，OpenAI的GPT-4模型训练需要2.5万张a100，消耗约5.1~6.2万MWh的电力，相当于1万个普通家庭五年的用电量(按一户普通人家一年用电1000KWh算)。而ChatGPT在生成一幅图像所需的电量几乎与一部智能手机所有电量相当。
- 全球电力需求暴增：国际能源署(IEA)预测，到2026年，AI、数据中心和加密货币的电力消费可能达到2022年水平的两倍以上，预计将增加540太瓦时(TWh)至1000TWh。

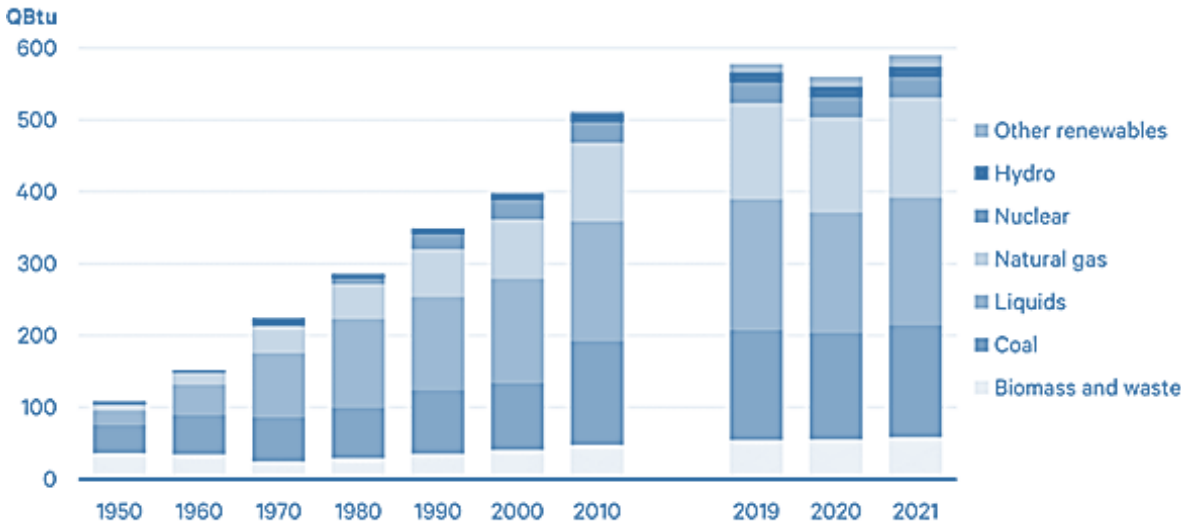
图：全球能源需求数据



可再生能源的利用，提供重要能源增量

- 核能的利用：核能作为一种低碳能源，也在被考虑用于满足AI的能源需求。
- 可再生能源的利用：许多国家和企业开始转向可再生能源。例如，青海省利用其丰富的绿色电力资源，包括光电、风电、水电等，实现了数据中心的100%清洁能源供应。

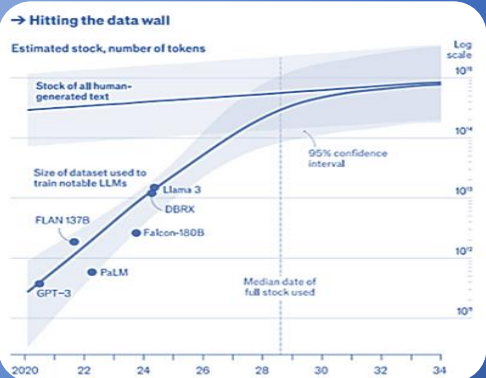
图：全球能源需求结构以煤炭、油及油气为主，未来如何演变？



数据墙显现，数据需求指数远超供给增速，新数据源急需拓展

数据墙显现

- ✓ 数据墙：高质量数据的需求增速大于高质量数据的创造速度



数据需求量指数增长

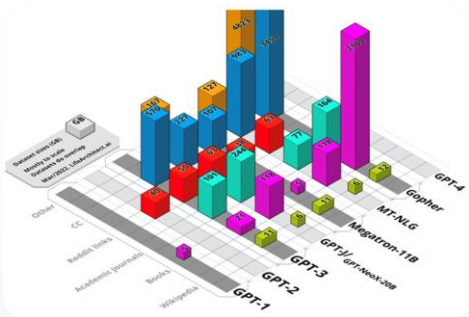
- ✓ 2018年GPT-1数据集约4.6GB,
- ✓ 2020年GPT-3数据集753GB
- ✓ 2021年Gopher数据集已达10550GB
- ✓ 2023年GPT-4的数据量更是GPT-3的数十倍以上。

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GPT-1		4.6					4.6
GPT-2				40			40
GPT-3	11.4	21	101	50	570		753
The Pile v1	6	118	244	63	227	167	825
Megatron-11B	11.4	4.6		38	107		161
MT-NLG	6.4	118	77	63	983	127	1374
Gopher	12.5	2100	164.4		3450	4823	10550

1. 主要数据集大小汇总，以GB为单位。公开的数据以粗体表示，确定的数据以斜体表示，仅原始训练数据集大小。

高质量数据源

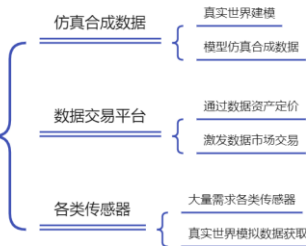
- ✓ 公开数据集：如互联网知识平台、公共健康数据等。
- ✓ 企业数据：企业内部积累的用户行为数据、交易数据等。
- ✓ 第三方数据服务提供商：专业数据服务公司提供的数据。



新数据获取方式

- ✓ 仿真合成数据：利用模型仿真生成的合成数据
- ✓ 数据交易平台：通过数据资产定价，激发数据市场交易。
- ✓ 传感器：进一步增加各类传感器，扩大数据获取

新数据获取



四、AI终极问题

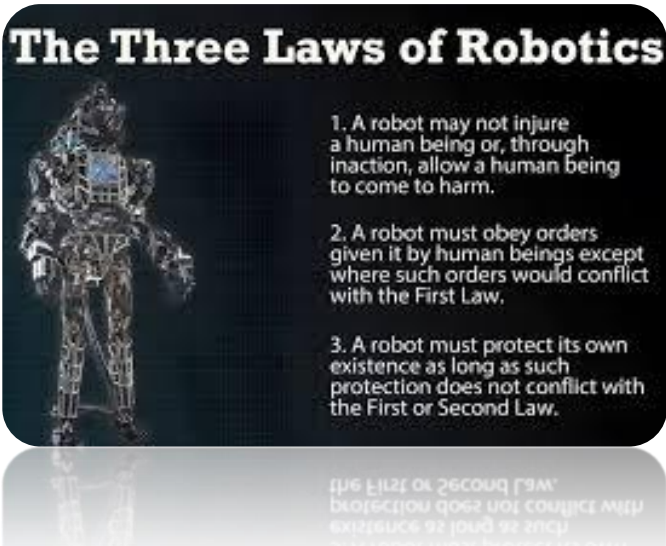
AI终极:该怎么面对凡人?



图：集异壁之集大成者



图：机器人的道德规范?



图：没有道德的智慧体，还需要人吗?



五、投资建议及风险提示

●端侧AI一触即发，拥抱第四次工业革命，关注AI产业链相关公司

端侧AI产品加速发布，处理器和内存是两大核心变化点。IDC定义AI手机算力须达30TOPS，微软定义AI PC算力须达40TOPS，以满足计算需求，驱动处理器在手机及电脑单机价值量提升。同时，端侧设备参数量不断变大，内存将不断增加，下一代AI手机内存有望增长至12~16GB，AIPC内存有望增长至16~32GB，存储芯片领域有望持续受益。此外，AI时代功耗增加是必然结果，配套电源管理及散热材料值得关注。相关公司包括立讯精密（苹果AI）、江波龙（存储模组）、澜起科技（存储接口芯片）、长电科技（先进封装）等。

●手机和PC需求温和复苏，看好“困境反转&龙头低估”企业

手机和PC等消费电子弱复苏，看好“困境反转&龙头低估”企业。我们认为24年智能手机和PC等消费电子呈现弱复苏，半导体龙头公司在行情来临时，有望获得较好的Beta收益，相关公司包括纳芯微（汽车模拟芯片）、卓胜微（射频前端芯片）、韦尔股份（CMOS传感器）等。

●风险提示：下游需求不足风险；地缘政治风险；核心竞争力风险；估值风险等。

图：重点公司财务指标及估值情况

代码	公司名称	主营业务	市值(亿元)	23年营收	23年归母净利润(亿元)	PE(2023)	24年一致预期归母净利润(亿元)	PE(2024E)	25年一致预期归母净利润(亿元)	PE(2025E)
002475.SZ	立讯精密	AI果链	2,883.92	2,319.05	109.53	26.33	135.80	21.24	171.83	16.78
603501.SH	韦尔股份	CMOS图像传感器	1,196.70	210.21	5.56	215.38	32.58	36.74	44.07	27.15
688008.SH	澜起科技	内存接口芯片	786.29	22.86	4.51	174.38	13.57	57.94	20.44	38.47
600584.SH	长电科技	先进封装	691.25	296.61	14.71	47.00	19.83	34.86	27.24	25.38
300782.SZ	卓胜微	射频芯片	544.58	43.78	11.22	48.52	7.02	77.58	10.04	54.26
301308.SZ	江波龙	存储器	369.56	101.25	-8.28	-44.64	10.53	35.11	10.65	34.70
688052.SH	纳芯微	隔离芯片	197.54	13.11	-3.05	-64.70	-2.94	-67.31	0.27	731.65

免责声明

长城证券股份有限公司(以下简称长城证券)具备中国证监会批准的证券投资咨询业务资格。

本报告由长城证券向专业投资者客户及风险承受能力为稳健型、积极型、激进型的普通投资者客户(以下统称客户)提供,除非另有说明,所有本报告的版权属于长城证券。未经长城证券事先书面授权许可,任何机构和个人不得以任何形式翻版、复制和发布,亦不得作为诉讼、仲裁、传媒及任何单位或个人引用的证明或依据,不得用于未经允许的其它任何用途。如引用、刊发,需注明出处为长城证券产业金融研究院,且不得对本报告进行有悖原意的引用、删节和修改。

本报告是基于本公司认为可靠的已公开信息,但本公司不保证信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用,并非作为或被视为出售或购买证券或其他投资标的的邀请或向他人作出邀请。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。在任何情况下,本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

长城证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易,或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。长城证券可能与本报告涉及的公司之间存在业务关系,并无需事先或在获得业务关系后通知客户。

长城证券版权所有并保留一切权利。

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引(试行)》已于2017年7月1日起正式实施。因本研究报告涉及股票相关内容，仅面向长城证券客户中的专业投资者及风险承受能力为稳健型、积极型、激进型的普通投资者。若您并非上述类型的投资者，请取消阅读，请勿收藏、接收或使用本研究报告中的任何信息。因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师声明

本报告署名分析师在此声明：本人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，在执业过程中恪守独立诚信、勤勉尽职、谨慎客观、公平公正的原则，独立、客观地出具本报告。本报告反映了本人的研究观点，不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接接收到任何形式的报酬。

公司评级：

买入——预期未来6个月内股价相对行业指数涨幅15%以上；

增持——预期未来6个月内股价相对行业指数涨幅介于5%~15%之间；

持有——预期未来6个月内股价相对行业指数涨幅介于-5%~5%之间；

卖出——预期未来6个月内股价相对行业指数跌幅5%以上。

行业评级：

强于大市——预期未来6个月内行业整体表现战胜市场；

中性——预期未来6个月内行业整体表现与市场同步；

弱于大市——预期未来6个月内行业整体表现弱于市场。

行业指中信一级行业，市场指沪深300指数

深圳

地址：深圳市福田区福田街道金田路2026号能源大厦南塔楼16层

邮编：518033

传真：86-755-83516207

北京

地址：北京市西城区西直门外大街112号阳光大厦8层

邮编：100044

传真：86-10-88366686

上海

地址：上海市浦东新区世博馆路200号A座8层

邮编：200126

传真：021-31829681

长城研究 · 与您共成长

长城证券产业金融研究院