人工智能行业专题（14）
大模型发展趋势复盘与展望

行业研究 · 行业专题

互联网 · 互联网II

投资评级：优于大市（维持）

证券分析师：张伦可
0755-81982651
zhanglunke@guosen.com.cn
S0980521120004

证券分析师：张昊晨
zhanghaochen1@guosen.com.cn
S0980525010001

请务必阅读正文之后的免责声明及其项下所有内容

核心观点

- **复盘美股科技巨头过去三年股价走势，AI叙事不断递进。**23年OpenAI领先全球开启AI加速度，微软受益于OpenAI独家合作，估值抬升明显。24年市场低估模型进步空间，叙事转向推理侧，认为应用公司最优，Meta坐拥社交垄断生态（潜在Agent入口）和广告场景，股价为除英伟达以外PE唯一抬升的巨头。*（24年2月，英伟达业绩会估计数据中心收入约40%来自推理。）*同年云厂商由于大幅增加资本开支但供给受限，云收入传导有延迟，三大CSP估值略有所回落。25年模型差距与OpenAI明显收敛，谷歌后来居上，生态优势为市场追逐。26年我们认为Scaling Law将持续，模型厂商打开差异化应用市场，模型推理侧需求或进入放量拐点。模型和算力或为最优投资方向。

- **25年四家巨头Capex同比增长50%以上，26年我们测算Capex将持续实现30%以上增速。**25年北美四家巨头Capex持续上修，从年初3200-3300亿美元上修至年末接近4000亿美元，各家全年Capex投入均处于50%以上同比增长。巨额Capex投入数据中心建设，或面临电力瓶颈。24年北美数据中心容量约25GW，据Grid Strategies估计，至29年的五年时间将新增80GW需求，而考虑到煤电退役、配套变压器建设周期长等原因，电力缺口预计将成为主要矛盾。因此巨头在建设数据中心过程中，算力能耗比或成为关键考量因素。

- **模型架构继续演化，Scaling law延续，多模态+长文本为Agent爆发提供基础。**23年迎来Scaling law红利期，24年多模态、推理模型涌现，25年算法工程与Scaling Law并进。长期来看，实现AGI仍然需要模型架构突破和Scaling做到极限。**展望26年：1）架构方面，**下一代模型架构目前需要解决两大核心痛点：①训练阶段Transformer的计算量和内存消耗瓶颈日益凸显；②推理时模型的记忆能力有限，且模型参数无法跟随记忆变化。海外目前已经有谷歌的Titans架构、以及Mamba架构，国内则更多从成本效率优化角度出发，包括Qwen3-Next、DeepSeek V3.2都取得了明显提升。**2）Scaling方面，预计26年Scaling law仍将延续，**包括从预训练到后训练以及推理场景，而强化学习则将成为未来的重点突破方向；**3）多模态、长文本能力更加成熟，**这为Agent的涌现提供技术基础。目前中美模型差距在3-6个月，算力和算法是追赶的关键。

- **通用大模型能力暂未分胜负，厂商商业化路径有差异。**1）**OpenAI：**尽管短期模型能力被反超，但下一代模型表现仍然值得期待，C端8亿用户是其核心壁垒，公司明年也将发力企业业务；2）**Gemini：**当前成为大模型的SOTA基准，得益于对原生多模态路线的坚持以及自研芯片的生态，明年Tokens消耗有望继续大幅提升；3）**Anthropic：**坚持2B路线，在编

请务必阅读正文之后的免责声明及其项下所有内容

核心观点

- 程领域能力突出，依靠2B的定价优势有望取得更好利润率水平，目前公司估值已达到3500亿美元，25年初推出的AI编程产品年末ARR也已经达到10亿美元；4）Grok：信奉大力出奇迹，由于推理场景有限因此训练算力充足，结合特斯拉独特数据优势，下一代原生多模态模型值得期待。

- **模型降低软件开发门槛，需求天花板打开，但玩家重新洗牌。**25年tokens消耗更多用于大模型企业内部以及推荐系统的重构，但26年开始预计下游应用的需求将持续增加，AI实际上打开了软件需求的天花板，据IDC数据，预计2029年全球SaaS市场将达到近1万亿美元规模（对比25年5800亿美元有明显增长），但我们认为其中玩家将重新洗牌。拥有数据壁垒，主要在垂类细分场景中布局，软件定义工作流程较复杂，或对准确度要求极高的行业，被大模型替代的风险较小，比如医疗、能源、会计、安全等领域。同时我们观察到大模型厂商已经开始通过与B端软件服务商合作开发更多行业需求，或与传统SaaS厂商产生直面竞争。

- **26年推理侧需求有望爆发，一级市场方面，编程场景、Agent爆发为主要应用方向。**从用户量和创业公司的收入估值水平来看，当前规模增长较快的行业以AI编程、AI Agent、AI内容创作为主，聚焦生产力提升，今年以来诞生了多个爆款应用，其中AI编程软件CursorARR已达到10亿美元，AI agent Manus则在8个月时间ARR达到1亿美元，AI 搜索工具Perplexity也在向Agent拓宽产品边界，其ARR也已达到2亿美元，因此明年办公场景有望迎来更多产品落地。此外，我们认为随着模型能力的成熟，明年在端侧的AI手机、AI眼镜，以及协助大模型在企业落地的分销商这些领域将看到明显的增长。

- 投资建议：建议关注算力基础设施（芯片、云厂商），如阿里巴巴-W（9988.HK）、百度集团-SW（9888.HK）、英伟达（NVDA.O）、谷歌（GOOGL.O）；以及大模型厂商，如阿里巴巴-W（9988.HK）、谷歌（GOOGL.O）、腾讯 （700.HK）等。

- 风险提示：宏观经济波动，下游需求不及预期，核心技术水平升级不及预期的风险，AI平权竞争加剧影响云业务利润率风险。

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 美股科技巨头股价和资本开支复盘展望

02 模型架构继续演化，Scaling law延续，多模态+长文本更成熟

03 模型厂发展路线分化，格局动态演进

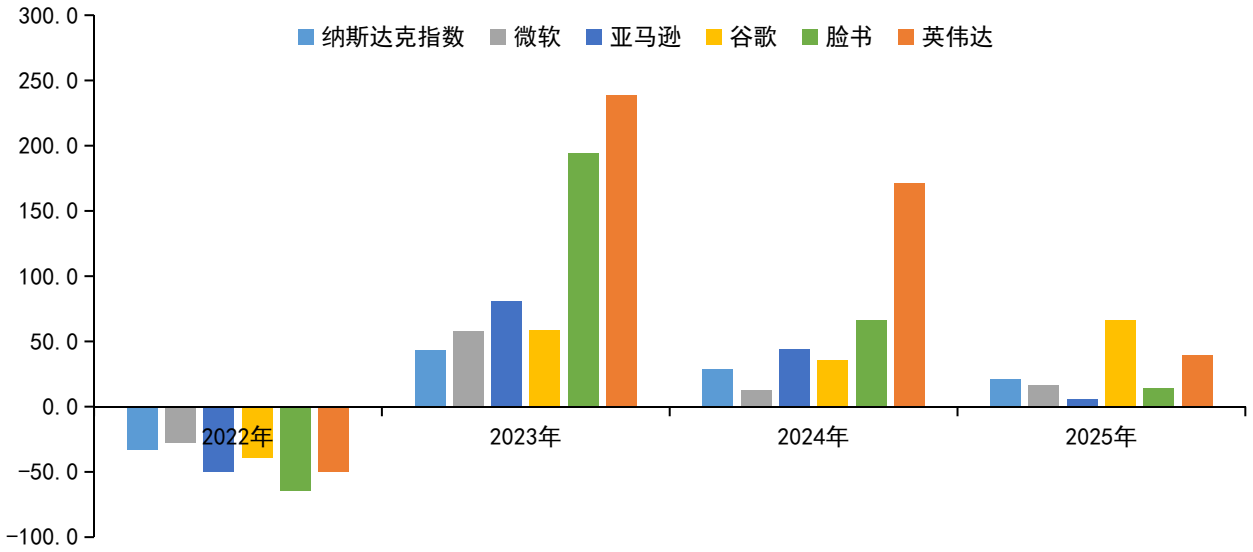
04 推理需求有望爆发，看好编程、Agent等应用

请务必阅读正文之后的免责声明及其项下所有内容

美股科技巨头走势复盘：过去三年AI领涨叙事不断演绎

- **23年：OpenAI领先全球开启AI加速度，微软受益于OpenAI独家合作，估值抬升明显。**全年五家巨头股价在22年大幅回撤后，均出现明显利润修复。
- **24年：市场低估模型进步空间，叙事转向推理侧，认为应用公司最优，Meta坐拥社交垄断生态（潜在Agent入口）和广告场景，股价为除英伟达以外PE唯一抬升的巨头。**（24年2月，英伟达业绩会估计数据中心收入约40%来自推理。）同年云厂商由于大幅增加资本开支但供给受限，云收入传导有延迟，三大CSP估值略有所回落。
- **25年：模型差距与OpenAI明显收敛，谷歌后来居上，生态优势为市场追逐。**上半年云厂商供给逐步释放，收入端呈现加速，AI需求景气度持续。下半年巨头逐步上修全年和26年Capex预期，市场开始担忧AI投资ROI，以及模型能力提升潜力，Scaling Law是否见顶。除谷歌外，巨头估值均出现回落。
- 26年预测：Scaling Law持续，模型厂商打开差异化应用市场，模型推理侧需求或进入放量拐点。模型和算力或为最优投资方向。

图：2022-2025年美股科技巨头年度涨跌幅（%）



资料来源：Wind、国信证券经济研究所整理

图：2023-2025年美股科技巨头股价涨幅因素拆解 （预测值为彭博一致预期）

	微软	亚马逊	谷歌	Meta	英伟达	纳斯达克指数
年初市值（亿元）	17,877	8,569	11,450	3,191	3,595	10,466
23E利润（年初）	733	197	693	228	81	465
23PE（23年初）	24	43	17	14	44	23
年末市值（亿元）	27,948	15,702	17,555	9,096	12,232	15,011
24E利润（亿）	843	390	835	457	464	495
24PE（23年末）	33	40	21	20	26	30
23年涨跌幅	**56%**	**83%**	**53%**	**185%**	**240%**	**43%**
利润涨幅	**15%**	**98%**	**20%**	**100%**	**473%**	**6%**
PE涨幅	**36%**	**-7%**	**27%**	**42%**	**-41%**	**35%**
年初市值（亿元）	27,948	15,702	17,555	9,096	12,232	15,011
24E利润（年初）	843	390	835	457	464	495
24PE（24年初）	33	40	21	20	26	30
年末市值（亿元）	31,339	23,240	23,178	14,837	32,888	19,311
25E利润（亿）	977	666	1098	655	1052	686
25PE（23年末）	32	35	21	23	31	28
24年涨跌幅	**12%**	**48%**	**32%**	**63%**	**169%**	**29%**
利润涨幅	**16%**	**71%**	**31%**	**43%**	**127%**	**39%**
PE涨幅	**-3%**	**-13%**	**0%**	**14%**	**19%**	**-7%**
年初市值（亿元）	31,339	23,240	23,178	14,837	32,888	19,311
25E利润（年初）	977	666	1098	655	1052	686
25PE（25年初）	32	35	21	23	31	28
年末市值（亿元）	36,231	24,858	37,942	16,785	45,572	23,242
26E利润（亿）	1194	863	1343	761	1776	830
26PE（23年末）	30	29	28	22	26	28
25年涨跌幅	**16%**	**7%**	**64%**	**13%**	**39%**	**20%**
利润涨幅	**22%**	**30%**	**22%**	**16%**	**69%**	**21%**
PE涨幅	**-5%**	**-17%**	**34%**	**-3%**	**-18%**	**-1%**

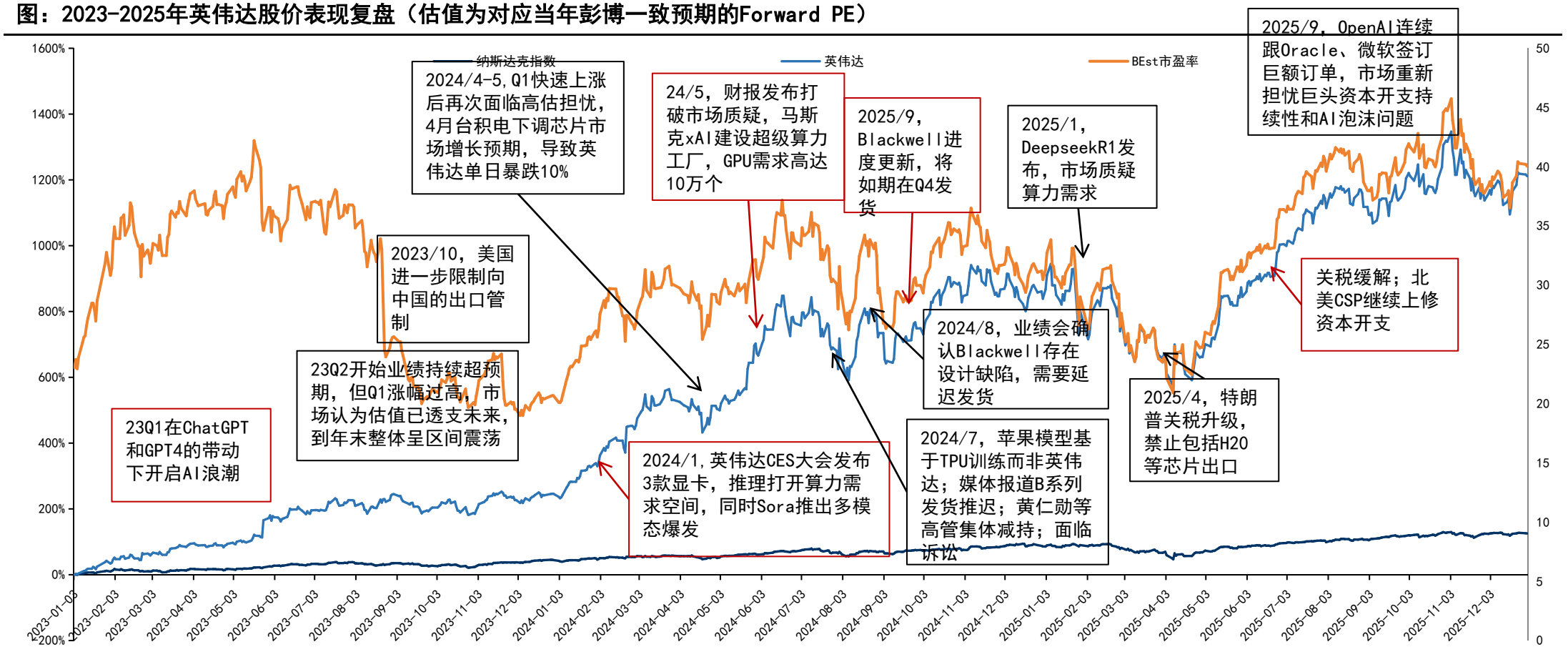
资料来源：Wind、彭博、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

美股科技巨头走势复盘——英伟达

- **英伟达在23-25年的上涨过程中经历了几轮震荡和调整，主因皆为估值达到高位后，市场对竞争、需求确定性、排产节奏等边际变化因素更加敏感，但随后英伟达均通过强劲的业绩表现回应了质疑。1）23年5月-12月：**经历Q1大幅上涨后，市场普遍质疑①估值过高，②美国限制出口风险，③其他公司的竞争；**2）24年中至25年4月持续震荡，**一方面业绩继续强劲，北美CSP持续上修资本开支，另一方面B系列发货推迟、竞争问题（TPU等）、以及DeepseekR1出现后带来对算力需求的质疑都造成公司股价的波动。

图：2023-2025年英伟达股价表现复盘（估值为对应当年彭博一致预期的Forward PE）



资料来源：Wind、彭博、国信证券经济研究所整理

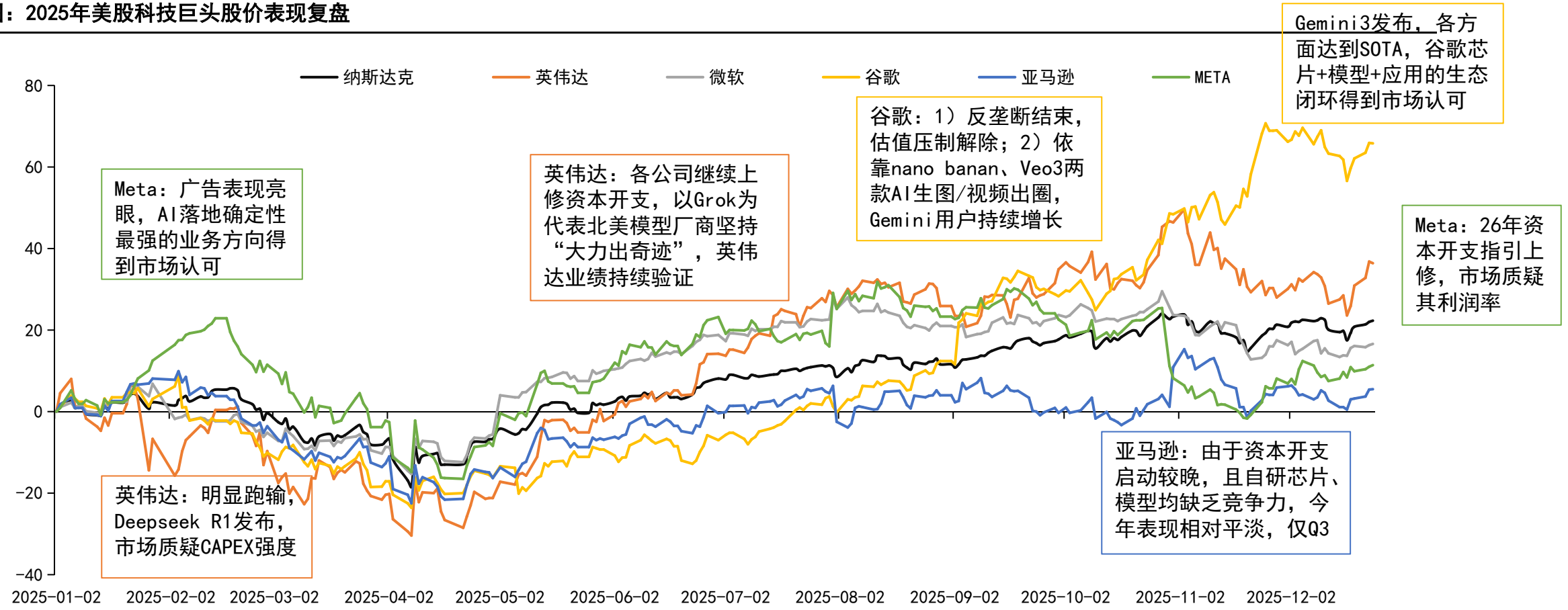
请务必阅读正文之后的免责声明及其项下所有内容

美股科技巨头走势复盘：25年AI叙事变化，拥抱最确定的方向

国信证券 GUOSEN SECURITIES

- **北美AI叙事的变化：拥抱最确定的方向。**年初市场陷入Scaling law放缓的论调，Deepseek发布后质疑CAPEX强度，因此英伟达领跌，选择最先AI落地广告兑现的Meta。Q2各公司持续上修资本开支，继续坚定“大力出奇迹”，结合英伟达业绩兑现，市场重新拥抱算力。Q3开始伴随谷歌反垄断压制解除，同时Gemini模型领先，用户持续增长，结合对OpenAI海量算力订单的质疑，市场对谷歌达成共识。全年来看，Meta由于在模型训练投入收效甚微，亚马逊由于在资本开支、自研芯片、模型的落后，二者成为表现最差的两家公司。

图：2025年美股科技巨头股价表现复盘



资料来源：Wind、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

资本开支回顾与展望：逐季度上修，26年预计仍将增长30%以上

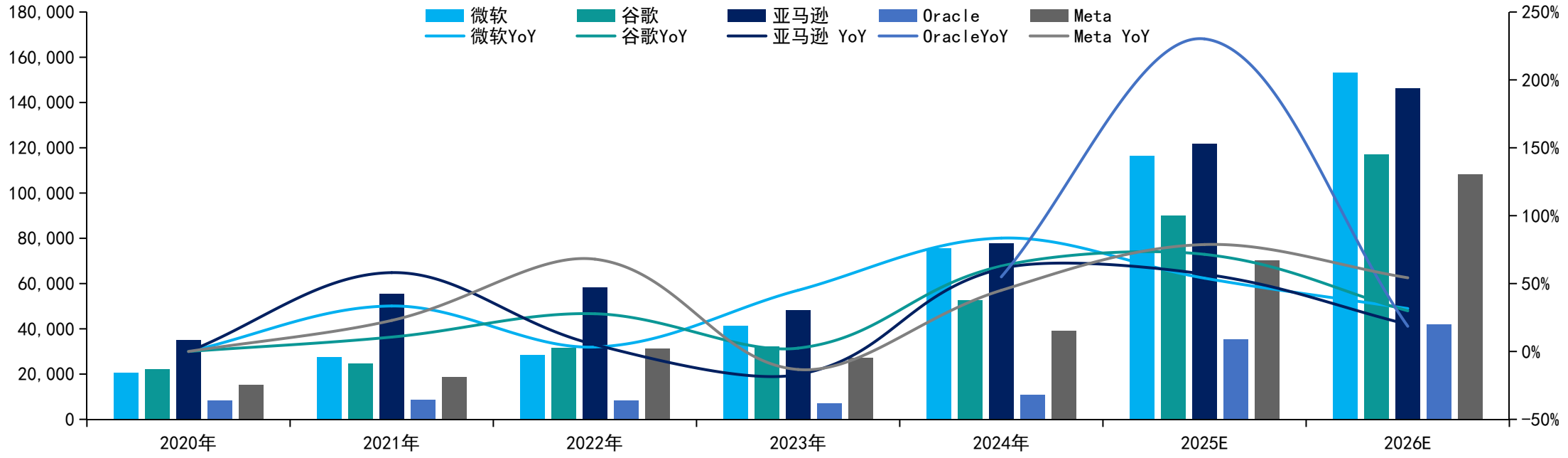
- **25年北美科技公司资本开支持续上修，年初3200-3300亿美元增长至年末的接近4000亿美元，各家全年Capex投入均处于50%以上同比增长。**微软：约1160亿美元(同比+54%)；亚马逊：1250亿美元(同比+61%)；谷歌：910-930亿美元(同比+73%-77%)；Meta:700-720亿美元(同比+78%-84%)。**预计26年仍将实现30%以上增长。**

图：2025各公司每季度财报后对全年CAPEX预计情况

单位：亿美元	2月	5月	8月	11月
Meta	600-650	640-720	660-720	700-720
微软	800	930	1000	1160
亚马逊	1000-1100	1000-1100	1100-1200	1250
谷歌	750	750	850	910-930

资料来源：公司财报、彭博、国信证券经济研究所整理

图：2020-2026年海外科技公司CAPEX（百万美元）



资料来源：Gemini、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

资本开支展望：电力成为制约CAPEX增长的核心因素

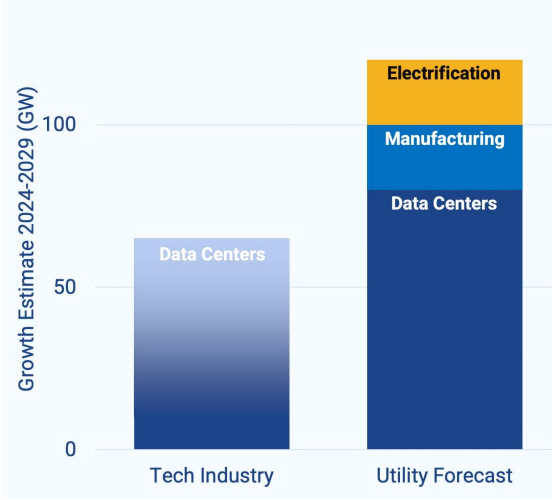
- **随着资本开支的持续增加，北美数据中心的用电需求急剧增加，26年预计CAPEX增长的限制因素将从缺芯更多转向缺电，**微软CEO曾表示微软手里囤积着成堆的英伟达AI芯片，却因为缺电、缺空间，只能让这些昂贵的GPU在仓库里“吃灰”。**据FERC（美国联邦能源管理委员会）预测，24年美国夏季用电负载峰值为929GW，预计到2029年5年时间将累计新增120GW。而据Grid Strategies估计，120GW的增量中大约80GW来自数据中心。**
- **电力供给无法与需求增量匹配带来用电缺口：**1）为了实现环保目标，北美有大量煤电厂需要退役；2）数据中心配套的变压器、输电线等设施的建设周期长达5-8年；3）电力需求要求冗余，这进一步推高了对电网的供给量要求。
- **面对电力缺口，目前各公司通过：1）与能源公司紧密绑定，**例如谷歌收购Intersect Power，亚马逊收购了 Talen Energy 位于 Susquehanna 核电站旁的数据中心园区。绑定之后通过“计量表后”（behind-the-meter）供电方式——直接将电力接入数据中心，绕过公共电网，以弥补供能缺口；**2）像海外寻找电力支撑，例如欧洲、亚太等电力资源相对丰富的区域；3）采用租赁电力资源充足的数据中心。**

图：北美22-24年峰值用电需求及未来五年用电需求增量预测



资料来源：Grid Strategies、国信证券经济研究所整理

图：北美数据中心未来五年增量预测



资料来源：Grid Strategies、国信证券经济研究所整理

表：海外云厂解决缺电问题的具体手段

公司	方式	具体手段
谷歌	收购	2025年12月以 47.5 亿美元 收购了能源基建公司 Intersect Power，绕开电网审批长周期（平均4-5年），直接锁定能源供应。并计划27年前在得州投资400亿美元新建数据中心，配套Intersect的10GWh储能与10.8GW发电项目。
亚马逊	海外拓展	欧洲、亚太地区有相对较多的电力冗余，亚马逊在德国、西班牙、印度等区域加大数据中心建设投资力度。
微软	海外拓展	计划四年内在海湾国家投资80亿美元
	租赁	微软通过提高融资租赁金额缓解买不到电的压力

资料来源：公司官网、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 美股科技巨头股价和资本开支复盘展望

02 模型架构继续演化，Scaling law延续，多模态+长文本更成熟

03 模型厂发展路线分化，格局动态演进

04 推理需求有望爆发，看好编程、Agent等应用

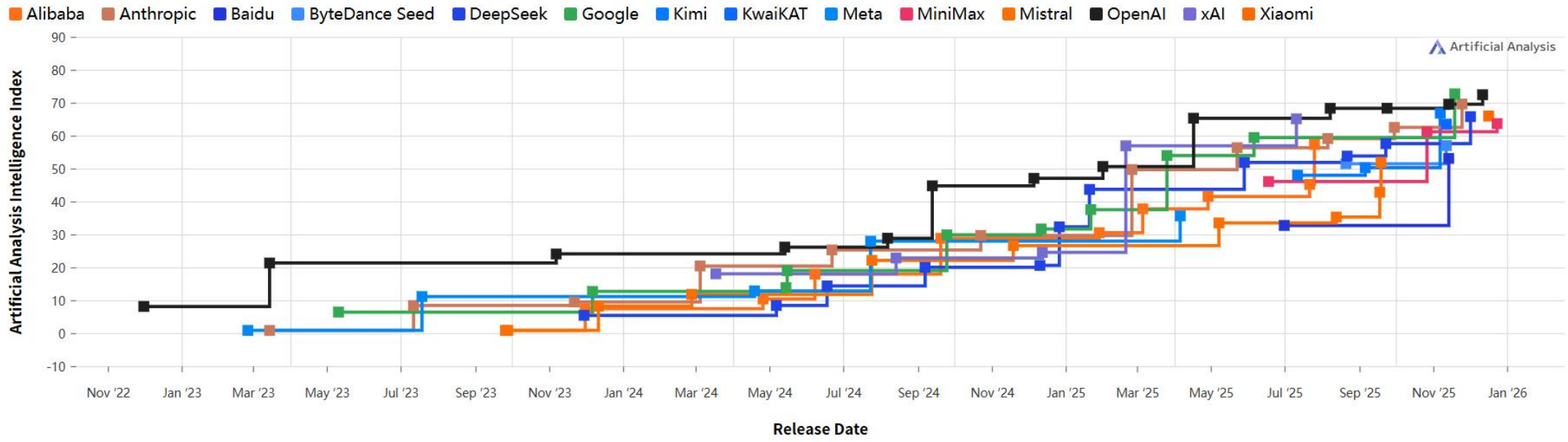
请务必阅读正文之后的免责声明及其项下所有内容

模型发展趋势：23年从“技术奇点”向“产业爆发”

国信证券 GUOSEN SECURITIES

- **Scaling law红利期，OpenAI领先优势明显。**23年各模型厂商的模型版本快速迭代，基本在半年时间就能完成一个大版本的升级，主要是由于预训练阶段的scaling law尚未触及数据瓶颈，通过扩展参数能够实现模型性能的快速提升，具备先发优势的OpenAI在23年通过GPT-4确立了模型在逻辑推理、复杂指令遵循上的SOTA基准。
- **模型仍以文本为主导，多模态雏形出现。**OpenAI作为技术的领导者推出了视觉理解模型GPT-4V，以及基于扩散模型的视频模型DALLE·3，嵌入ChatGPT产品，增加其图像、视频能力，谷歌的原生多模态模型Gemini也在23年末正式推出。
- **基于Transformer架构的大语言模型应用场景以Chatbot为主。**受限于推理能力和上下文长度，在编程等场景中仅具备本科学生级别基本代码能力，可处理中等难度问题，但仍然主要依赖网页端一问一答、手动改代码。

图：前沿模型能力比较



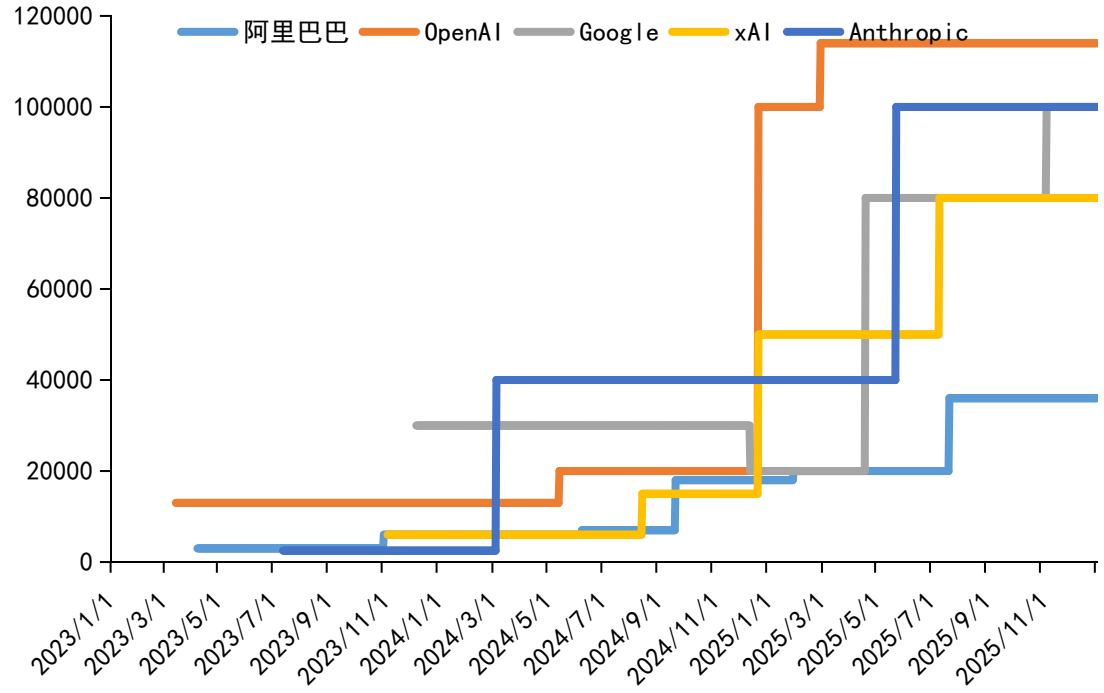
资料来源：artificial analysis、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

模型发展趋势：23年从“技术奇点”向“产业爆发”

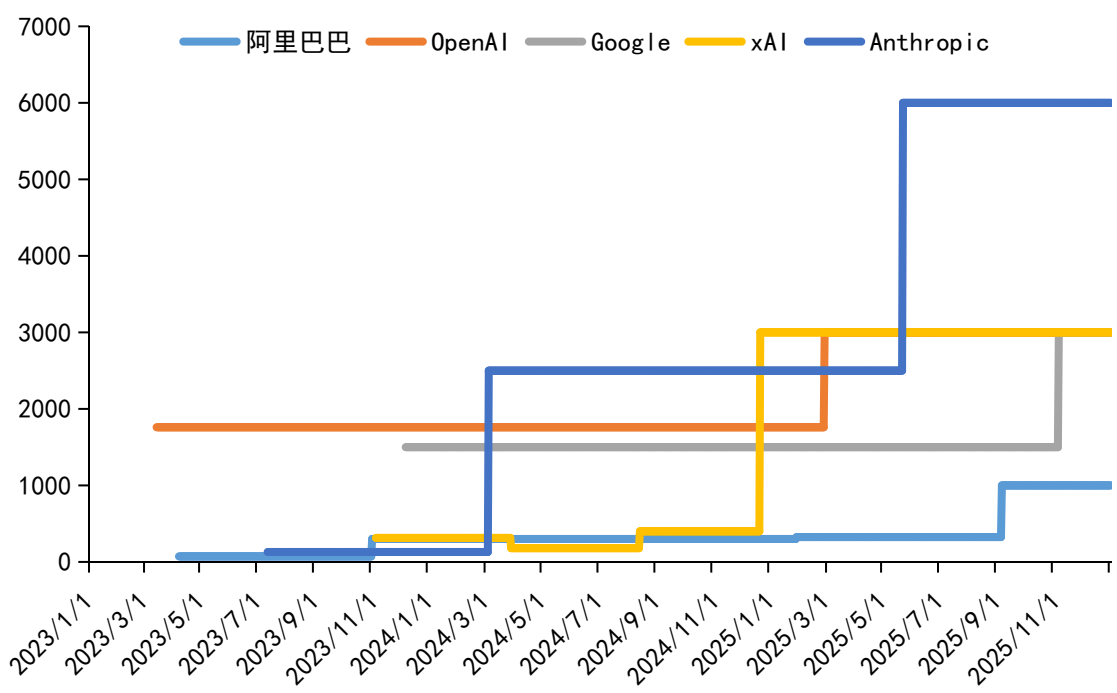
- **模型参数和训练数据量持续增长。**根据DeepMind团队22年提出的Chinchilla Scaling法则，每增加 1 倍的参数量，对应的训练数据量（Tokens）也应该增加 1 倍，且最佳配比应为每10亿（1B）参数，至少需要200亿（20B）Tokens的数据进行训练。在使用合成数据后，模型预训练的数据量得到持续提升，根据lifearchitect.ai，预计海外目前头部模型最大参数量均已达到2万亿以上，马斯克曾提到26年的Grok5将达到6万亿参数，而对应的训练数据也扩张到100万亿以上。但在目前的部分论文中已证明（例如《AI models collapse when trained on recursively generated data》），随着合成数据使用的增加，会导致模型出现崩溃，因此拥有稀缺数据资源的模型厂商更有机会取得模型能力的明显提升。

图：头部模型厂商重点模型版本训练数据量（十亿tokens）



资料来源：lifearchitect.ai、国信证券经济研究所整理

图：头部模型厂商每代最大参数模型参数情况（十亿tokens）



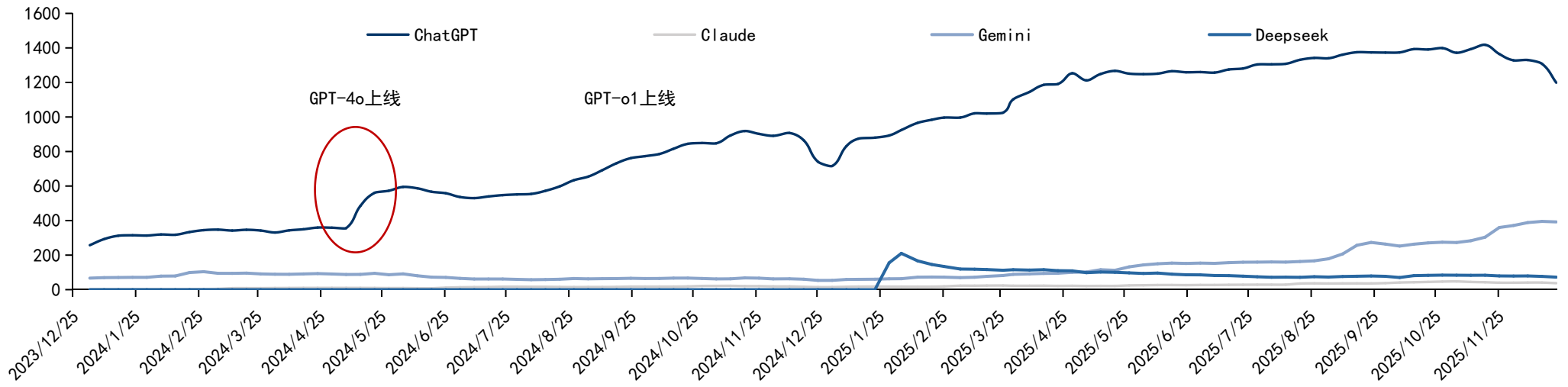
资料来源：lifearchitect.ai、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

模型发展趋势：24年多模态、推理模型开始涌现

- **24年OpenAI继续引领了模型发展的两大核心趋势：1）文本以外，图像、视频更强的模型开始涌现，包括GPT-4o为代表的多模态模型，但能力并未完善，**GPT-4o的生图能力实际延期一年上线，同时OpenAI也在年初推出了视频模型Sora，谷歌5月发布Veo此后包括可灵、即梦等应用陆续上线；**2）以GPT-o1（9月）为代表的推理模型方向。**预训练的Scaling law开始边际放缓，GPT-o1引入了模型的慢思考能力，通过思维链（CoT)在给出最终答案前进行内部思考，利用推理时间的延长提升模型能力上限，在数据、编程、推理等领域提升显著。但模型上下文追踪、工具调用能力不足，agent 能力未落地（当时 SWE bench 仅能修复 1%-5%现实世界 GitHub 任务）。
- **Claude、Gemini、Llama等模型开始缩小与OpenAI的差距。**在 LMSYS Chatbot Arena榜单中，24年初Claude 3 Opus首次完成了对GPT系列模型的超越，此后开源模型Llama3、Gemini也达到接近Openai旗舰模型的能力。
- **编程能力提升，Cursor开始出圈。**年中Claude 3.5开始展现出在逻辑推理和编程方面的领先优势，结合无基础编程的破圈，Cursor的ARR迅速从23年末的100万美元增长至1亿美元。

图：头部模型产品24年以来周访问量情况



资料来源：SimiliarWeb、国信证券经济研究所整理

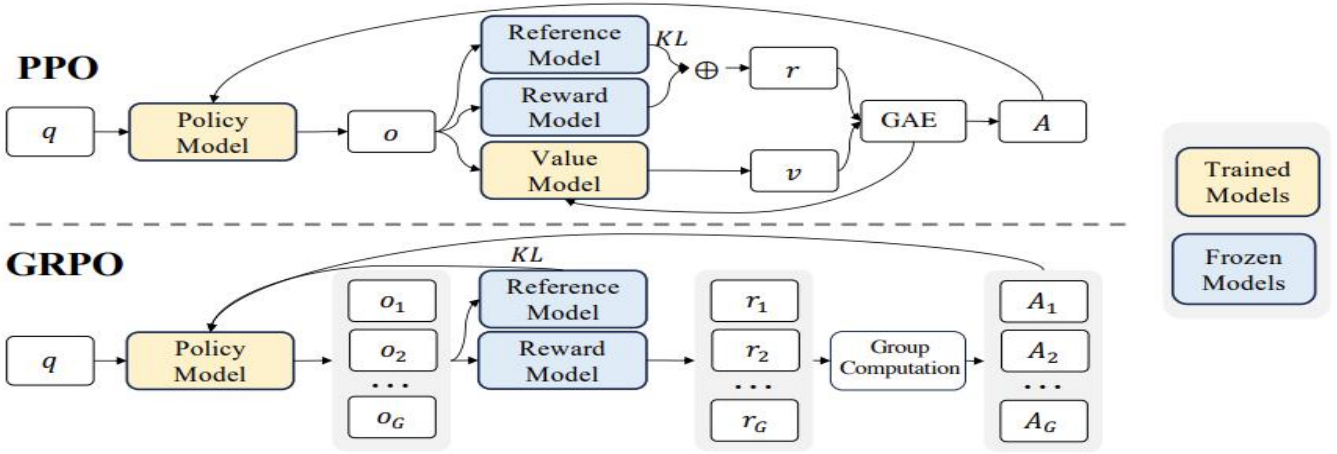
请务必阅读正文之后的免责声明及其项下所有内容

模型发展趋势：25年从“暴力美学”重新回归“算法与工程”

国信证券 GUOSEN SECURITIES

- **25年开始，训练环节中RLVR（可验证奖励强化学习）成为一种新的普遍使用的方法。**其正式提出是在24年11月Allen Institute for AI（Ai2）的研究团队发布的Tülu 3项目报告中，业内普遍认为在24年9月的OpenAI首款推理模型o1中最早得到了使用。
- **RLVR带动了模型思维链的使用，推理模型进一步实现了模型性能跃升。**在RLVR出现以前，SFT主要是通过让模型模仿人工编写的思维链进行思考，但是RLVR通过可验证的奖励函数，逼迫模型自行寻找最佳的思维链（告诉模型一个推理问题和答案，模型不断尝试不同的思考过程直到正确回答问题），之前常用的RLHF是通过打分制去评分，模型不需要做到最准确只要能获得高分即可。因此相比原来后训练的SFT+RLHF，加入RLVR一方面降低了对人工编写思维链的需求，冷启动的数据要求变少，另一方面由于需要更多进行尝试，算力需求增加，所以Karpathy*（OpenAI联合创始人）*在其博客文章2025年度回顾指出：2025 年 AI 的进步很大程度上是因为实验室把原本用于预训练（读更多书）的算力，挪到了 RLVR 上。**Raschka*（AI领域极具影响力的教育家、研究员）*在其博客文章中认为 2026 年会看到更多 RLVR 相关的工作。目前 RLVR 主要用在数学和代码上，下一步是扩展到其他领域。**
- **Deepseek R1通过工程和算法的创新证明了通过强化学习+思维链，能够比单纯的增大参数（暴力美学）更有效率。**DeepSeek 提出了RLVR+ GRPO，在RLVR的训练范式下，传统的PPO算法需要一个价值模型，消耗大量显存，采用GRPO算法，可以直接省去价值模型，通过在模型生成的回答中，提高得分高于平均分的答案生成的概率，来实现模型性能的提升，带来了效率的极大提升。

图：PPO和GRPO对比



资料来源：知乎、国信证券经济研究所整理

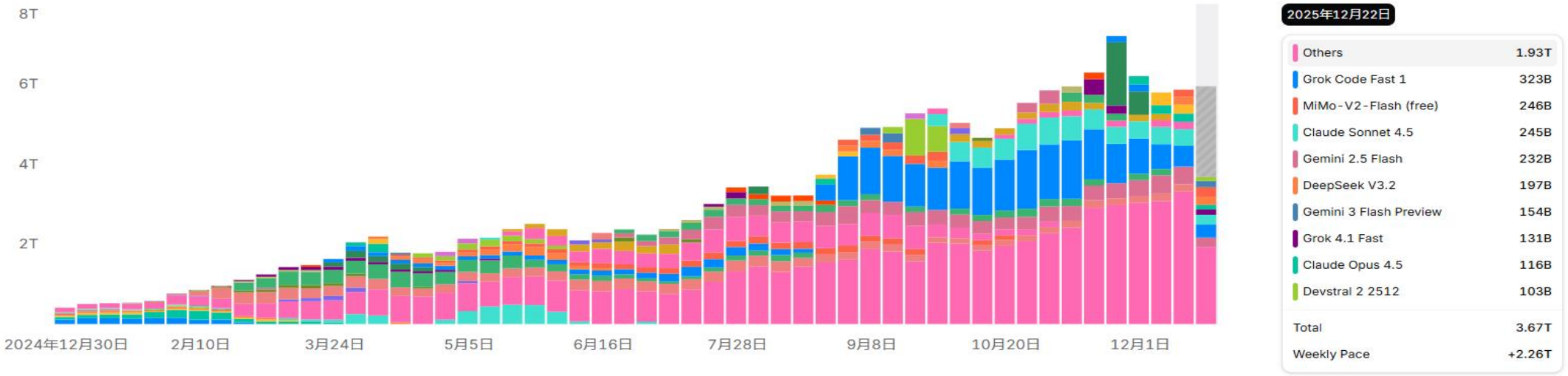
请务必阅读正文之后的免责声明及其项下所有内容

模型发展趋势：25年从“暴力美学”重新回归“算法与工程”

国信证券 GUOSEN SECURITIES

- **OpenAI缺乏模型能力突破，Gemini引领原生多模态。**全年来看，OpenAI的模型能力缺乏有效突破，从诞生起就是原生多模态路径的Gemini系列开始成为市场中的SOTA基准，且谷歌的自研芯片+模型+应用的闭环生态成为其实现领先的关键。
- **Coding发生质变，Agent开始走向成熟。**1）Coding方面，2月claude 3.7 发布，coding 能力显著提升，可通过Cursor 工具在命令行/服务器自主拆分长任务、完成工具调用，改变开发者对 agent coding 的认知。2025 年 3 月 Gemini 2.5 Pro 发布：文本写作、情商表达、幻觉控制能力突出，成为科研论文写作的关键工具。token 消耗呈快速增长（接近指数级），核心驱动为agent coding、多智能体协作（同步开 2-3 个端口推进任务）、深度搜索功能使用。2）经历前两年在RAG的普及、长文本能力推理能力的提升以及MCP的推出等各方面能力积累，Agent开始走向成熟。

图：OpenRouter 周度tokens消耗情况



资料来源：openrouter、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

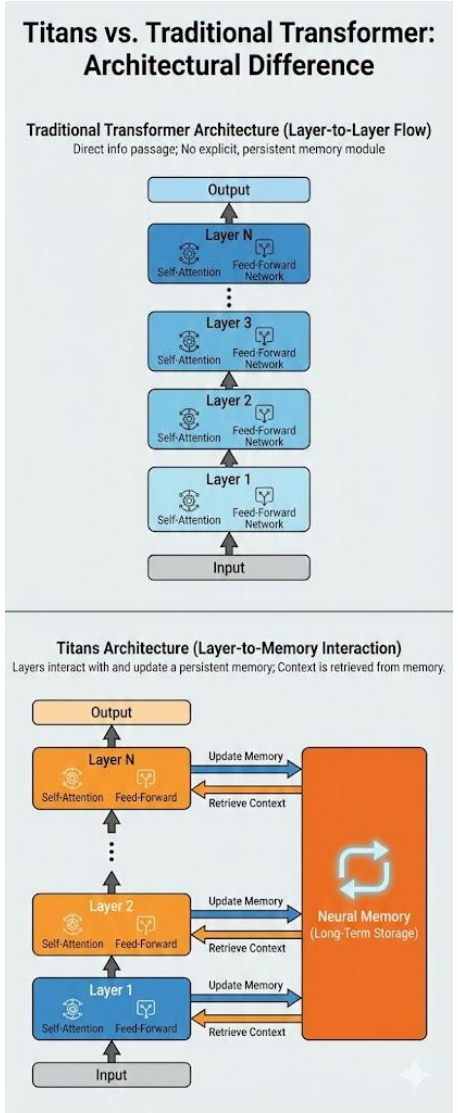
展望26年：后Transformer时代的模型架构演化

- 谷歌 DeepMind CEO Demis认为，距离实现AGI仍然需要：1）需要一两个像Transformer、AlphaGo这样的重大技术突破；2）大语言模型的Scaling必须做到极限。
- **下一代模型架构需要解决的核心痛点：1）训练阶段Transformer的计算量和内存消耗瓶颈日益凸显。**随着模型上下文长度要求增加，自注意力机制的效率问题正在日益凸显，每个token都要“关注”其他所有token，导致计算量和内存消耗与序列长度的平方成正比（$O(N^2)$）。**2）推理时模型的记忆能力有限，且模型参数无法跟随记忆变化。**目前对话的临时信息都存在KV Cache中，缓存清空模型就会失忆，并且这种记忆能力是外挂的，模型参数固定，模型无法在交互中学到新东西。

✓ **Google Titans架构：结合Transformer的并行训练优势与RNN的线性推理优势的架构，对于模型降低推理成本、提升长文本处理能力有明显提升。**谷歌在24年年末发布Titans架构论文（《Titans: Learning to Memorize at Test Time》），年末在NeurIPS 2025大会上，正式提出了Titans+MIRAS，并发布文章《Titans + MIRAS: Helping AI have long-term memory》，预计目前已经全面应用于Gemini的模型中。根据论文内容，Titans保留了Transformer标准的Attention机制，在传统 RNN 里，记忆存放在一个固定的向量里，在Titans里，记忆存放在“权重”神经长期记忆模块里，采用RNN的思路，每走一步，权重更新一次。**这意味着1）传统Transformer在模型训练后会冻结参数，但是Titans能够根据上下文动态更新参数，实现在推理过程中持续学习；2）模型将上下文编码进记忆模块的权重中，能够处理超过200万Token的超长序列。**

✓ **Titans也并非完美无缺，根据Titans的论文：**1）这种架构的幻觉率有可能更高，Transformer是原封不动把数据计入KVcahe，但是Titans是压缩了信息，所以有损失。2）在推理的每一步（或每几个步），模型都需要执行反向传播，这意味着单个token的算力消耗远高于传统Transformer架构，**且现有芯片的支撑尚不成熟，但谷歌拥有自研TPU，能够更好适配部署。**

图：Titans架构和传统Transformer架构对比



资料来源：谷歌、国信证券经济研究所整理

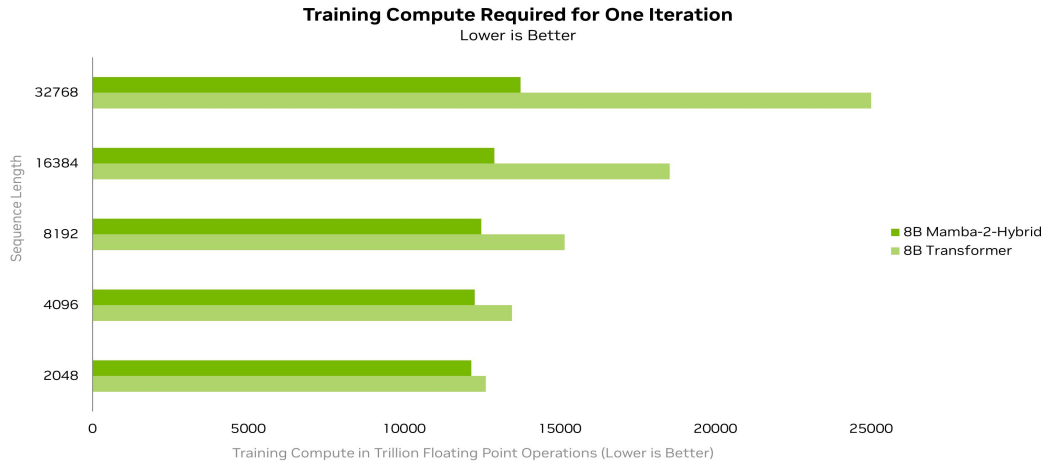
请务必阅读正文之后的免责声明及其项下所有内容

展望26年：后Transformer时代的模型架构演化

国信证券 GUOSEN SECURITIES

- ✓ **Mamba架构：Mamba采用的是结构化状态空间模型（SSM），与Transformer完全不同，23年末由Albert Gu和Tri Dao提出，优势是计算量低于Transformer，更高的推理速度与较低的显存占用。**根据论文《Mamba: Linear-Time Sequence Modeling with Selective State Spaces》，Mamba可以视为一种进化版的RNN，但是在训练时可以利用并行关联扫描算法实现并行计算而不是像RNN那样排队计算，此外还引入了选择性机制，模型可以根据当前的Token内容，动态地决定是“记住”信息、“忽略”噪声，还是“重置”状态。但在实际工程落地中仍面临挑战，特别是其依赖的并行扫描操作主要受限于显存带宽，且当前硬件侧主要都是围绕Transformer进行优化，因此硬件适配度也不够成熟。
- ✓ **Mamba-2：24年中推出，在其论文《**Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality》中指出，通过引入结构化状态空间对偶性（Structured State Space Duality, SSD）理论，消除了 SSM 与 Transformer之间的鸿沟（使二者等价），这使得核心计算也可以利用GPU 里的Tensor Cores（专门加速矩阵乘法的硬件单元），训练速度大大提升。和Titans类似，因为压缩了信息，精准度仍然弱于Transformer。
- ✓ **目前Transformer+Mamba的混合架构也成为了尝试方向，例如：**
  - **AI21 Labs的Jamba：**采用了一种交错结构，即每若干个Mamba层后插入一个Transformer注意力层（Attention Layer）。
  - **IBM的Granite 4.0：**IBM在2025年底推出的Granite 4.0系列模型中，采用了Mamba-2作为主要架构组件。IBM通过大量实验发现，混合架构（Mamba-2 + Attention + MoE）在保持企业级任务所需的推理精度的同时，能显著降低长文档处理（如法律合同分析、代码库理解）的推理成本。

图：8B Transformer 模型一次迭代所需的计算量与 8B Mamba-2-Hybrid 模型一次迭代所需的计算量对比



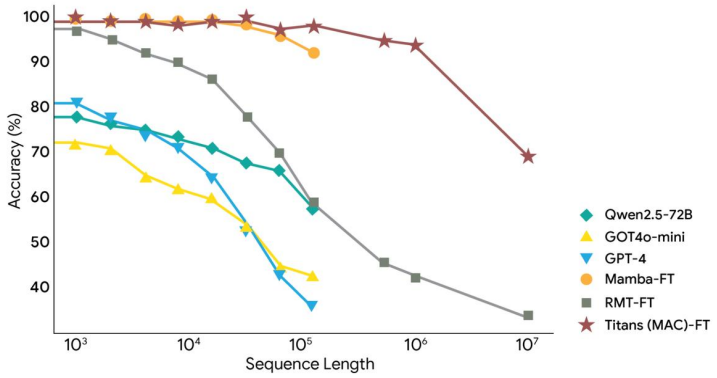
资料来源：Nvdia、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

展望26年：后Transformer时代的模型架构演化

- **国内的架构变化方向：算力受限的背景下，更为关注成本降低**
- **Qwen3-Next：采用混合注意力机制+高稀疏MOE**。25年10月推出，模型架构上，混合注意力机制75%层使用 Gated DeltaNet（线性注意力变种）25%层保留 标准注意力（增强版），而Gated DeltaNet的计算成本随长度平稳增加，历史信息被编码进一个数学上的“隐藏状态”里，信息流过时，状态被不断更新。
- **Deepseek V3.2：采用动态稀疏注意力（DeepSeek Sparse Attention（DSA））**。25年5月推出，在正式计算注意力之前，模型先用一个极小的开销预估哪些历史位置的Token与当前Token最相关，强制注意力机制只计算这些最相关的tokens，其余的跳过或简化处理。

图：Titans架构在长文本的精确度上表现优异



资料来源：Google Research、国信证券经济研究所整理

图：新模型架构成本、上下文长度、速度对比

	谷歌Titans	Mamba-2	Qwen3-Next	DeepseekV3.2
上下文长度（tokens）	200万+		原生支持26万+，可扩展至100万+	128K
速度		序列长度越长，比transformer的速度快的越多，当序列长度达到256K时，Mamba-2 层的训练速度比Transformer层快18倍	**训练上，**80B参数的模型训练所需的GPU Hours不到 Qwen3-30A-3B的80%；**推理上，**长文本情况下，吞吐量远超Qwen3，当上下文长度超过32k时，吞吐提升更是达到10倍以上。	
成本降低	titans在节省成本上并不像国产模型极致，但是增加了模型根据对话学习的能力	长文本情况下，计算成本比Transformer有明显改善（见上页）	80B参数的模型比Qwen3-32B训练成本减少90%	文本长度的增加对边际成本的影响很小，且当文本长度达到128K时输入成本仅为0.2美元/百万tokens，输出仅为0.3美元/百万tokens
输入价格（百万tokens）	Gemini 3 flash：0.5美元		Qwen3 Next 80BA3B ：0.5美元	DeepSeekV3.2：0.28美元
输出价格（百万tokens）	Gemini 3 flash：3美元		Qwen3 Next 80BA3B ：2美元	DeepSeekV3.2：0.4美元

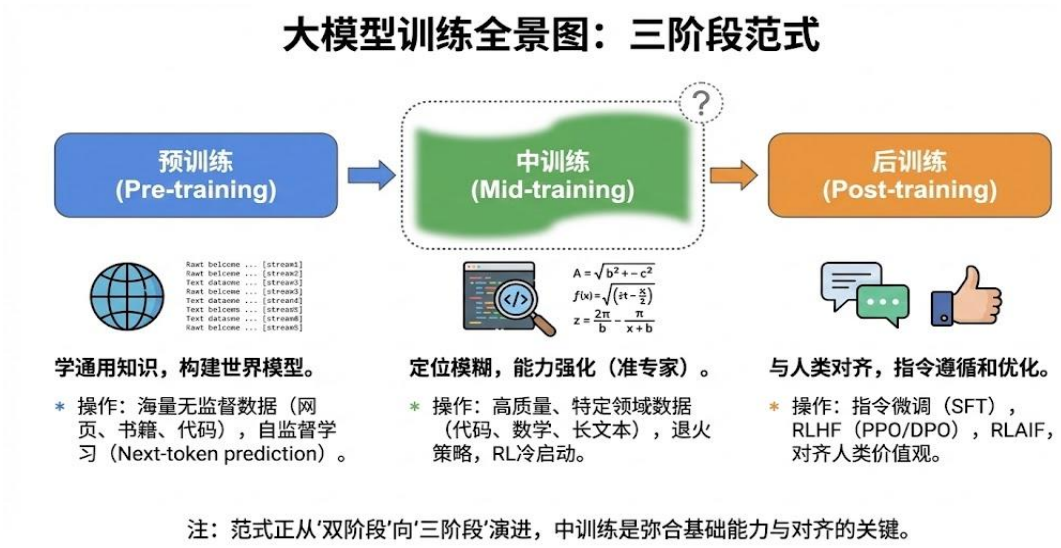
资料来源：artificialanalysis、Google Research、Qwen、Deepseek、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

展望26年：Scaling law远未见顶，强化学习是未来重点

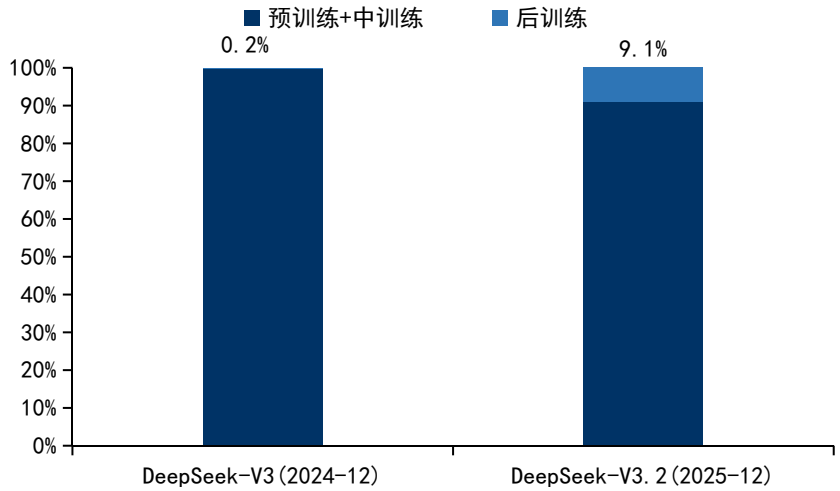
- **预训练scaling仍未见顶：**随着Gemini3的发布，谷歌再次证明尽管可用数据正在逐渐减少，但预训练仍有scaling的空间，据Gemini预训练负责人Sebastian Bourjou，预训练通过合成数据和模型架构的调整，能够继续实现scaling，且在接下来的一年中仍不会失效。（例如随着模型逐步转向原生多模态模型，原本的视频模型、图像模型都可以作为教师模型合成更多优质数据进行训练）
- **引入Mid training：**介于预训练和后训练SFT之间的阶段，通常在模型预训练结束具备基本常识后，使用高质量高密度的数据进行定向能力的提升，包括垂直领域数据/私有数据、代码数据等，或在该阶段扩展上下文长度。例如DeepseekV3，虽然没有在技术文档中专门提到Mid-Training环节，但在预训练后使用YaRN 旋转位置编码技术扩展了上下文长度，同时保证了注意力稳定性。
- **推理中的Scaling：**主要是通过测试时计算（Test-Time Compute），利用思维链，延长模型的思考时间实现scaling（自我博弈），因此我们看到flash、pro、deepresearch模式下模型思考时间越长能够回答越深度的问题，即使不增加模型参数，模型性能依然能提升。24年9月随着OpenAI o1模型的推出就已经进入公众视野。

图：模型训练环节



资料来源：Deepseek、Llama、国信证券经济研究所整理

图：以Deepseek为例，后训练环节的算力持续提升



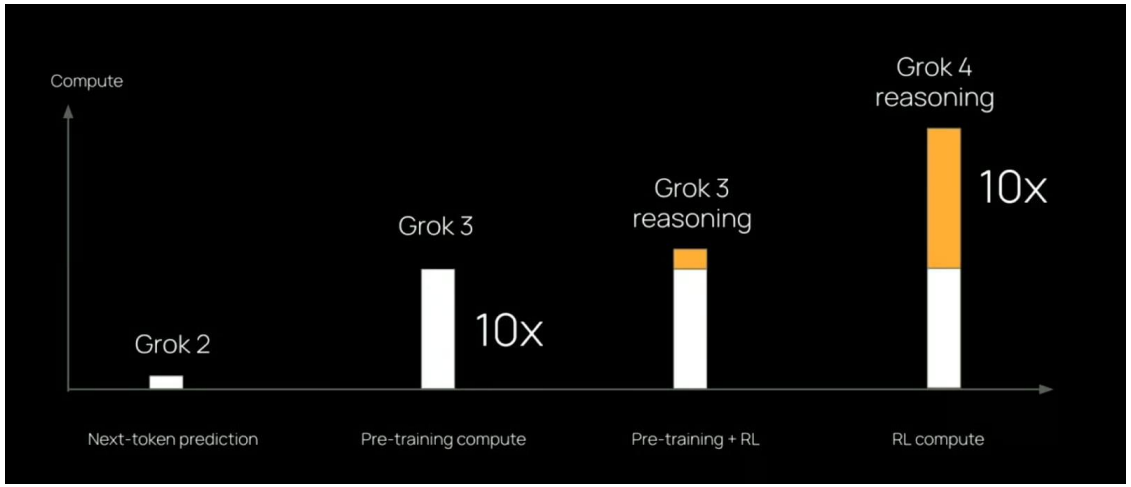
资料来源：Deepseek、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

展望26年：Scaling law远未见顶，强化学习是未来重点

- **谷歌Deepmind首席科学家Jeff Dean认为未来的重点是强化学习，因为人类学习的tokens远少于现在的模型，却能达到很高的能力水平，主要是由于人类会思考行动的后果、观察反馈，这种方式更偏向强化学习。**
- **强化学习也是模型训练继续Scaling law的关键，**2025年7月发布的Grok4较Grok3在预训练环节计算量不变的基础上，在后训练阶段实现了10倍的计算量增长，带来了模型能力的进一步提升。
- **根据奖励来源包括：**1）RLHF（基于人类反馈的强化学习），最早使用，由人为对不同的回答质量做排序，反馈给奖励模型后再对预训练后的模型回答进行打分，模型不断自我修正；2）RLAIF（基于AI反馈的强化学习），利用AI来完成反馈；3）RLVR（基于可验证信号的强化学习）DeepseekR1采用，主要是使用有标准答案的数学、计算机、逻辑等领域数据去自动验证。
- **根据算法包括多种方法，例如：**1）PPO：2017年OpenAI提出，RLHF的标准算法，但是计算效率低且稳定性一般；2）GRPO：DeepSeek24年末提出，能够大幅节省显存；3）PRM（过程奖励）：不再只给最终答案打分，而是针对推理的每一个中间步骤打分。这能有效纠正模型“虽然答案对但逻辑错”的现象。
- 大模型在前期会进行预训练，这个阶段决定了模型能力的上限，模型在过程掌握大量知识，但缺乏指令遵循能力。之后通过监督学习SFT能快速提升模型指令遵循的能力，之后再进行强化学习寻求突破，通过打分机制鼓励模型不断探索，找到更好的答案突破性能的天花板。当前算法优化的方向包括降低成本、关注推理过程、改善模型的尝试方式（如何聪明地尝试不同回答以找到最好的答案）。

图：Grok每代版本算力投入变化情况



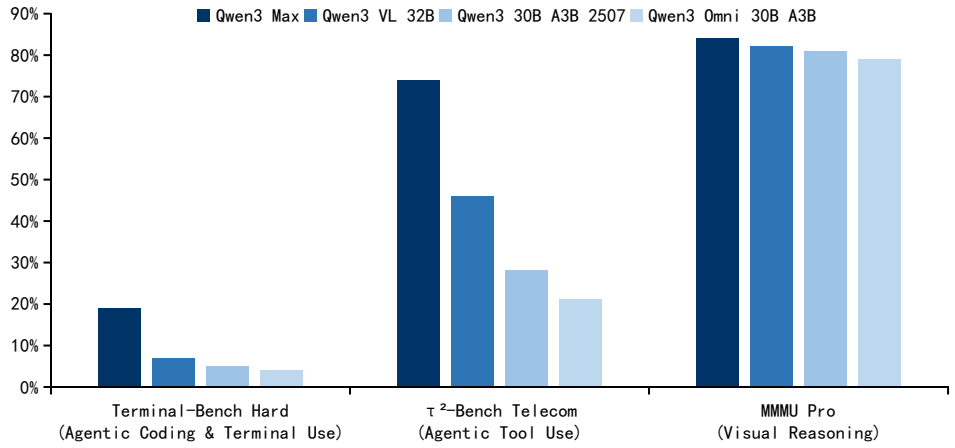
资料来源：xAI、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

展望26年：推理能力提升+多模态+长文本=更成熟的Agent

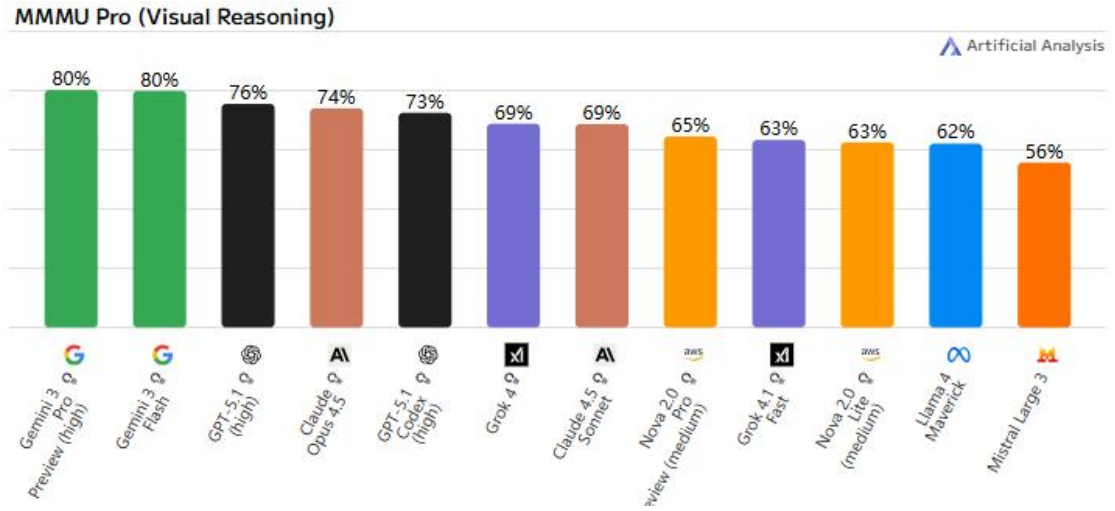
- **推理能力提升+多模态+长文本=更成熟的Agent。**具备多模态能力和更强推理能力才能够应用于更多场景（例如端侧），并且在多步执行中不会忘记前面的执行步骤，因此伴随模型在这些维度能力的继续提升，26年的应用侧有望看到更多更成熟的Agent。
- **多模态能力：25年的一个重要趋势是从文本模型+外挂多模态向原生多模态模型转变。**GPT-4o上线一年后在今年5月推出了原生生图能力，阿里在9月推出了Qwen3-Omni，马斯克在Baron Capital第32届年度投资会议上提到，26年即将推出的Grok5也将是原生多模态模型。
- **在开发原生多模态转变的过程中，核心问题是“如何用一套统一的tokens表示不同内容形态”。**例如一段10分钟的视频当中，文本内容的数据量远低于视频内容，如果各自全部转化为tokens输入模型，视频内容的tokens将远多于文本，导致模型理解的文本内容会被干扰甚至忽略，这是需要算法方面解决的问题，也是谷歌目前保持领先优势的重要原因。

图：Qwen3 Omni较同参数Qwen3模型性能下降



资料来源：artificial analysis、国信证券经济研究所整理

图：谷歌视觉推理等多模态能力领先



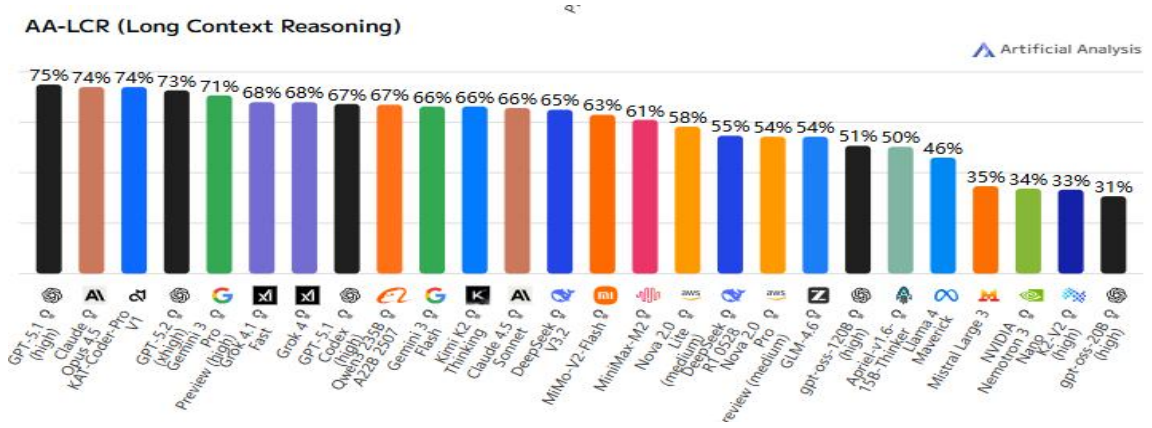
资料来源：artificial analysis、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

展望26年：推理能力提升+多模态+长文本=更成熟的Agent

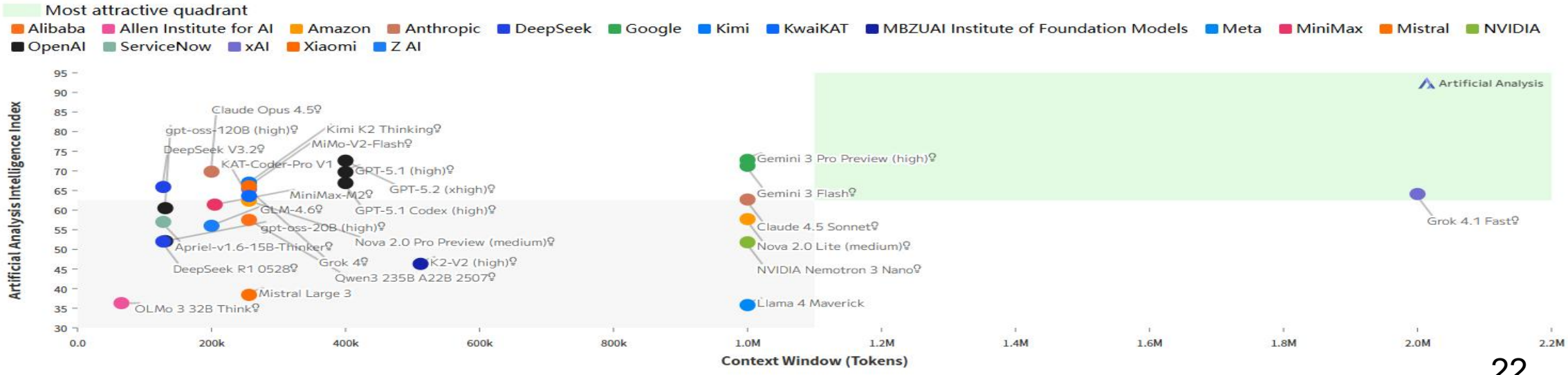
- **长文本：**在当前的模型架构下，Transformer虽好，但自注意力机制下，每个token都要“关注”其他所有token，导致计算量和内存消耗与序列长度的平方成正比，在较长任务执行的过程中很难保持效率。**而解决长文本问题一方面可以通过在长文本上做训练、或改善模型架构实现。**Mamba架构以及谷歌的Titans架构都在长文本上实现了突破。

图：AA-LCR（长文本推理能力）测试评分排名



资料来源：artificial analysis、国信证券经济研究所整理

图：主流模型上下文长度对比



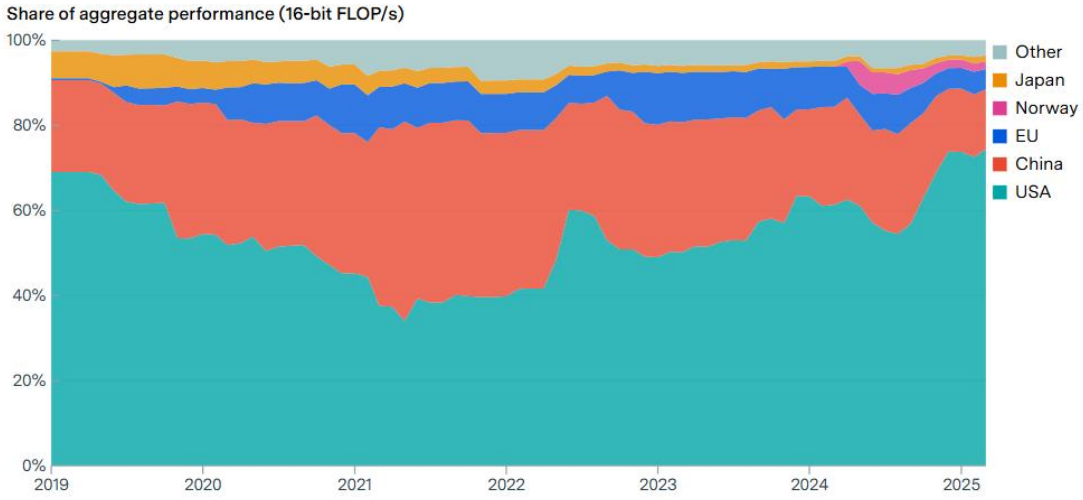
资料来源：artificial analysis、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

国内外模型差距：中美模型差距3-6个月，算力、算法是关键

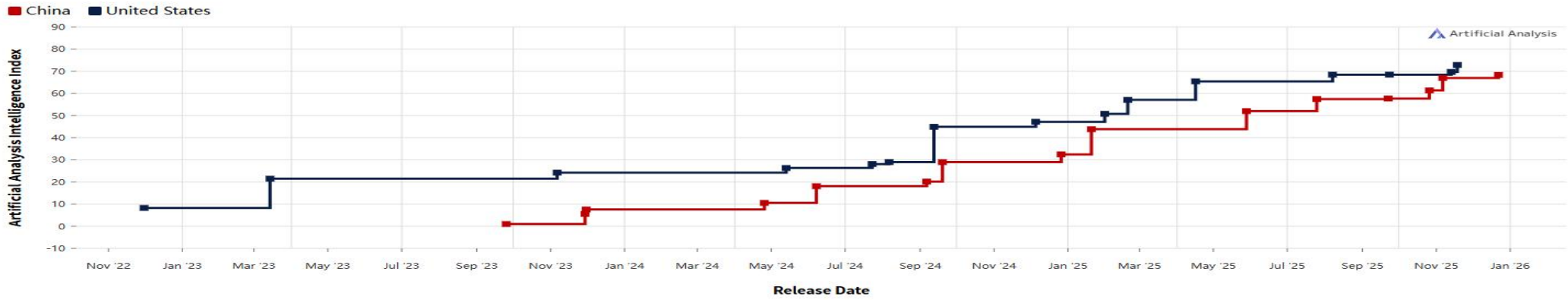
- **算力：**据Epoch AI，美国掌握了全球大约70%的算力资源，中国只有15%-20%，尤其是对头部模型厂商来说，算力的差距使得Google有能力去尝试不同的模型架构而不用担心算力不足的问题（Titans架构本身依旧对算力消耗要求非常高），而国内在算力受限的背景下，无法尝试更大的模型参数。
- **算法：**目前在多模态领域，国内尚未找到如何用一套统一的tokens表示不同内容形态的方法，因此在原生多模态上与谷歌仍有差距。但在强化学习领域海外并没有明显的算法能力上的领先，并且由于算力的限制，国内的算法方向更多聚焦于实现训练推理效率的提升。
- **数据：**模型训练中，数据的质量和丰富度对模型性能影响至关重要，因此催生了Scale AI、Surge AI等数据标注的公司，但国内大厂本身拥有丰富的数据团队以及垂类应用场景，无论数据质量还是丰富度，都没有明显落后于海外。

图：全球GPU集群性能份额



资料来源：Epoch AI、国信证券经济研究所整理

图：中美最前沿模型性能比较



资料来源：artificial analysis.、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 美股科技巨头股价和资本开支复盘展望

02 模型架构继续演化，Scaling law延续，多模态+长文本更成熟

03 模型厂发展路线分化，格局动态演进

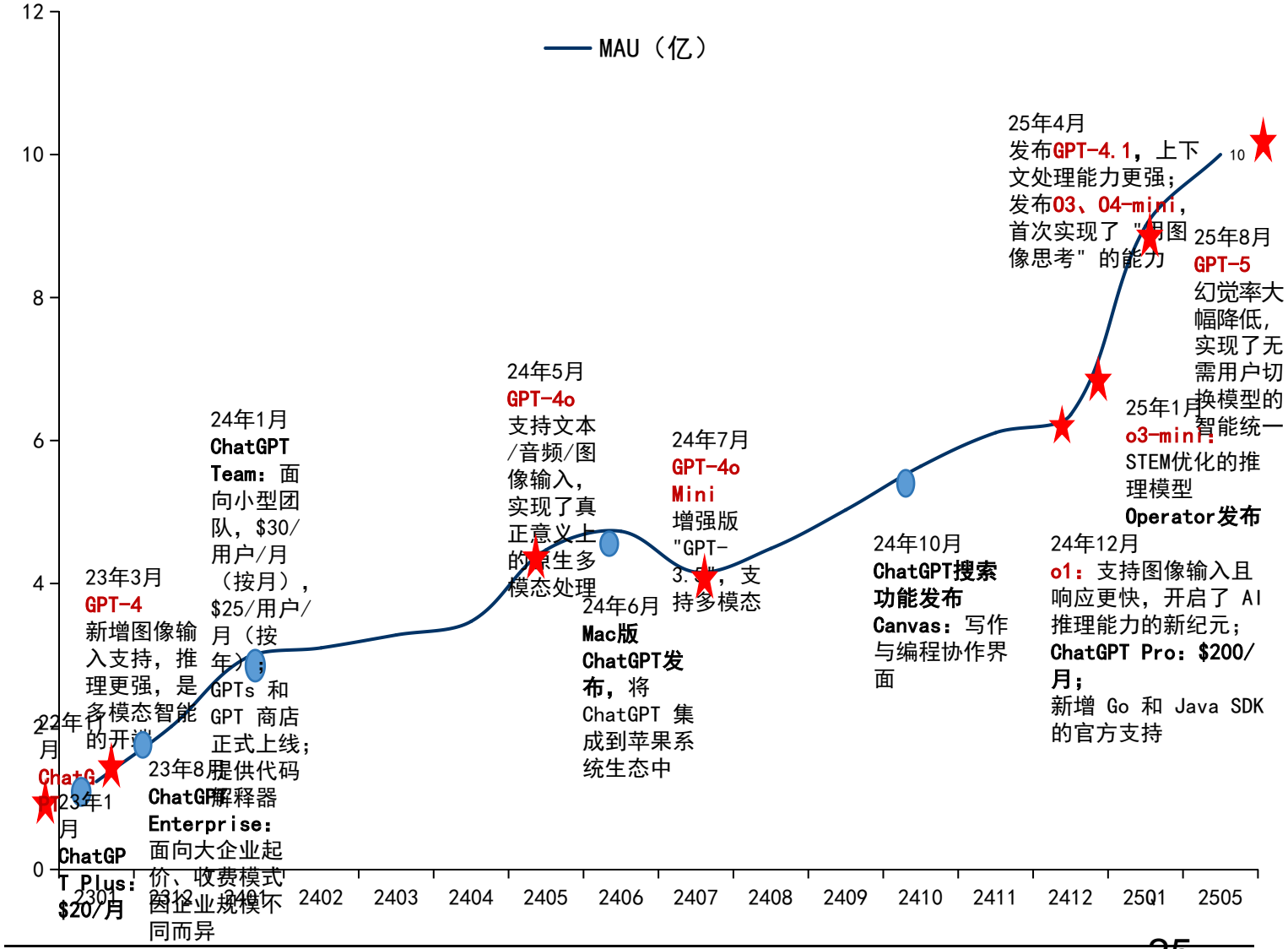
04 推理需求有望爆发，看好编程、Agent等应用

请务必阅读正文之后的免责声明及其项下所有内容

OpenAI：C端份额领先，明年将发力企业业务

- **OpenAI从一开始就坚持“消费者优先”的2C战略。主要原因是：1）早期模型的稳定性和能力，并不足以支撑大多数企业级应用；而现在，它们正在达到这个门槛；2）OpenAI具备先发优势，一旦占据C端市场也会更容易进入B端市场。因此，今年OpenAI企业业务的增长速度，已经超过了消费者业务，26年的战略重点也会是企业业务。**
- **记忆是OpenAI的护城河。**奥特曼在12月17日Big Technology Podcast的采访中表示，相比模型性能更重视粘性，认为chatgpt的用户粘性和护城河来自习惯、一致性和个人体验积累出的信任感。由于ChatGPT最先抢占市场构建起了超过8亿用户的基本盘，用户在ChatGPT有更多的历史交流记录，也意味着回答能够更贴近用户的需求。
- **模型能力来看，**1）OpenAI各方面能力相对更加均衡，无明显短板，但没有突出特色。2）尽管今年OpenAI模型能力突破并不明显，但仍需观察有望26Q1上线的OpenAI下一代模型基于英伟达B系列芯片训练后显示出的性能表现。

图：ChatGPT月活变化



数据来源：ChatGPT官网、Xsignal、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

OpenAI：25年收入预计达到130亿美元，2028年目标1000亿美元

国信证券 GUOSEN SECURITIES

- OpenAI当前收入仍以C端订阅为主，我们预计OpenAI2025年收入130亿美元，其中C端订阅约占61%，目前整体用户中付费渗透率约5%，付费用户约4000万+。长期来看，根据纽约时报，OpenAI目标在2028年实现1000亿美元，除了用户订阅持续增长外，企业端收入持续增长，还需要广告、电商等其他流量变现方式共同推动公司整体的利润率改善。

图：模型训练环节

单位：百万美元	2024A	2025E	2026E	2027E	2028E	2029E	2030E
OpenAI收入	3,700	13,000	30,000	60,000	100,000	150,000	200,000
YoY		251%	131%	100%	67%	50%	33%
订阅	2,700	7,976	16,608	27,000	38,000		
Agent	-	2,990	7,493	13,300	20,900		
API调用	1,000	2,035	3,534	6,850	12,450		
直接调用	800	1,628	2,827	5,480	9,960		
占比	80%	80%	80%	80%	80%		
通过Azure调用	200	407	707	1,370	2,490		
其他收入（广告、电商抽佣）			2,365	12,850	28,650		
营业成本		10,158	23,400	45,600	72,000		
收入分成（20%）	740	2,600	6,000	12,000	20,000		
云服务成本（推理成本）		7,558	17,400	33,600	52,000		
占收比		58%	58%	56%	52%	49%	
毛利		2,842	6,600	14,400	28,000		
毛利率		22%	22%	24%	28%		
销售&管理费用	1,000	6,047	8100	11400	13000		
费用率	27%	47%	27%	19%	13%		
研发费用（主要为训练算力成本）		20,256					
费用率		156%					
经营利润		-23,460					
OPM		-180%					

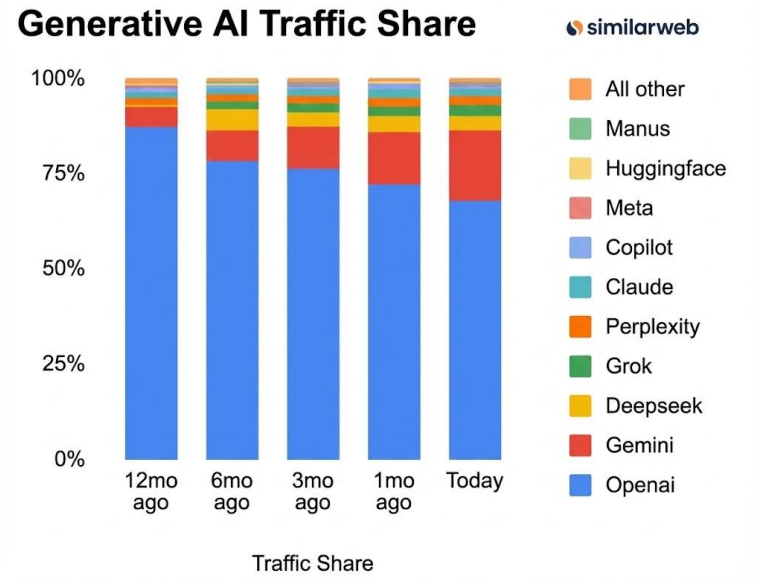
资料来源：The information、国信证券经济研究所整理测算

请务必阅读正文之后的免责声明及其项下所有内容

Gemini：技术积累+高质量数据+算力=SOTA

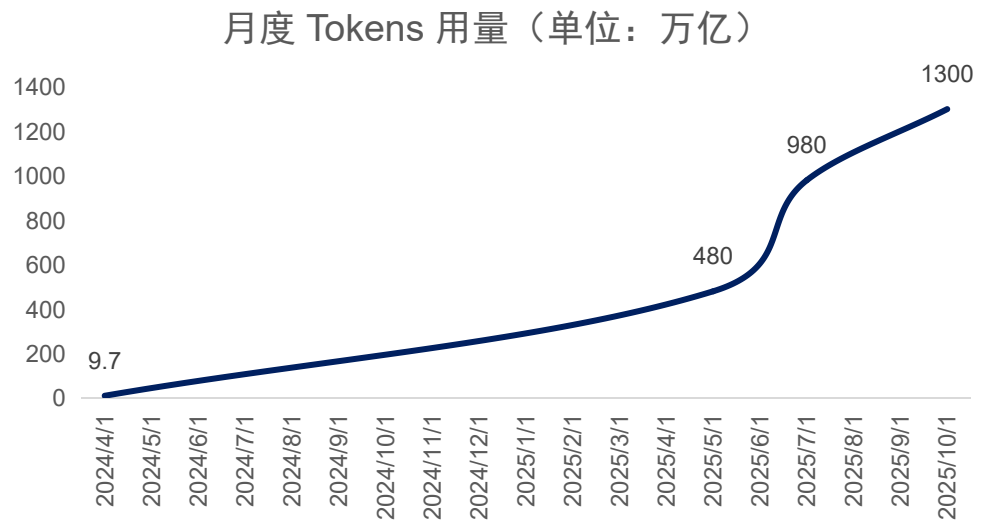
- **今年以来谷歌成为大模型技术新的引领者，月tokens消耗量以及流量的市场份额持续增加，主要得益于：**
- **1）正确技术路线的坚持，技术积累深厚。**谷歌在AI领域积累深厚，是Transformer的提出者，且在Gemini一推出就坚持走原生多模态路线，并且在目前其他厂商（例如Grok）将大量资源投入到后训练情况下，Gemini预训练负责人认为预训练仍然有能力依靠架构优化实现scaling law；
- **2）丰富的高质量数据资源：**谷歌搜索拥有Wiki这样的高质量文本数据（且搜索内容质量相对高于口语/网络用语较多的社交媒体数据）以及Youtube这样高质量的视频数据，且由于一开始就坚持多模态，因此对视频数据的预处理也做的相对较好，能够很好用于模型训练。
- **3）充足的算力资源：**自研TPU不仅提供了充足的算力资源，同时还能针对模型训练做专门适配，实现训练效率的最大化。

图：25年末生成式AI流量份额



资料来源：similarweb、国信证券经济研究所整理

图：谷歌月度tokens增长



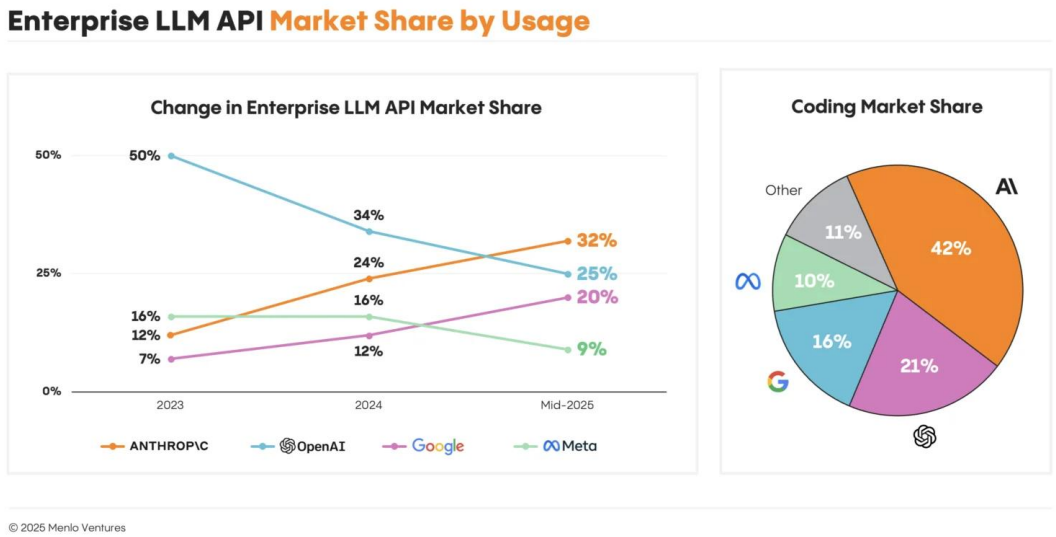
资料来源：公司公告，国信证券经济研究所整理测算

请务必阅读正文之后的免责声明及其项下所有内容

Anthropic：Coding能力突出，是代码开发场景的首选

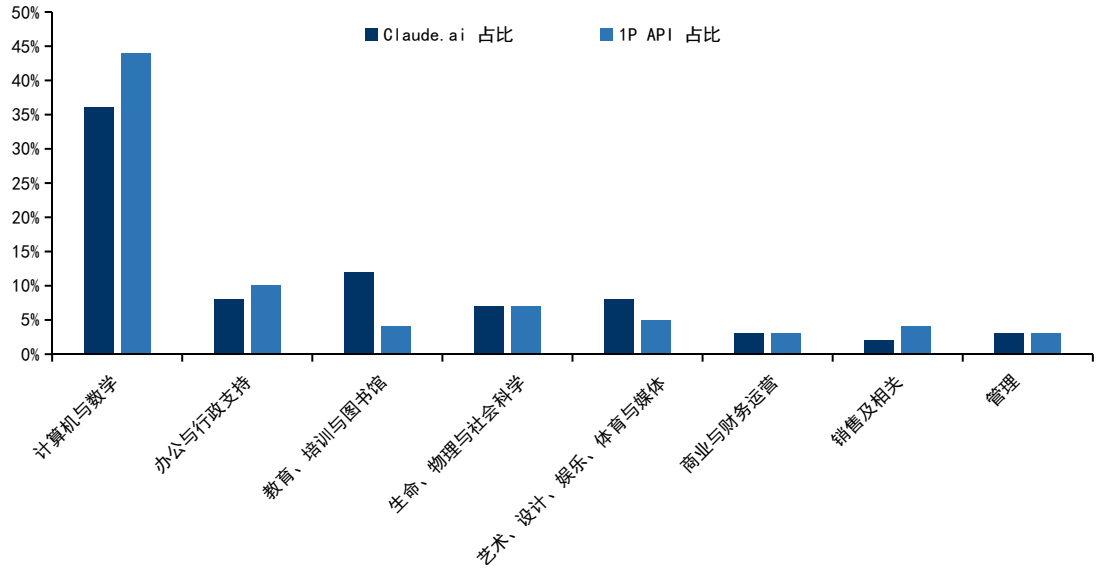
- **Coding能力突出，是代码开发场景的首选。**Anthropic 的势头起于 2024 年 6 月发布的 Claude Sonnet 3.5，到了 2025 年 2 月的 Claude Sonnet 3.7 更是首次展示出“Agent 优先”的 LLM 雏形。到 2025 年 5 月，随着 Claude Sonnet 4、Opus 4 以及 Claude Code 的推出，其领先优势已被彻底坐实。根据Anthropic数据，44%的流量都是与计算机和数学类职业相关，细分的使用场景中前三名的分别是调试 Web 应用程序、解决技术问题以及构建专业商业软件。此外，根据Menlo Ventures，目前Anthropic在企业大模型API市场份额达到32%，Coding市场份额则达到42%。
- **Anthropic的逐步赶超尤其是在编程领域的优势主要得益于1）创始人Amodei对于编程场景的重视；2）编程和推理场景由于有确定性的答案，因此尤其在RLVR出现后能得到明显的能力提升。**Amodei认为代码能力 = 顶级逻辑能力，在代码中磨练出的“解决复杂问题的能力”，可以直接迁移到生物医学、法律、数学和金融等严谨学科中。

图：企业大模型API调用份额



资料来源：Menlo Ventures、国信证券经济研究所整理

图：Anthropic Claude对话用途结构（通过claude应用和API调用）



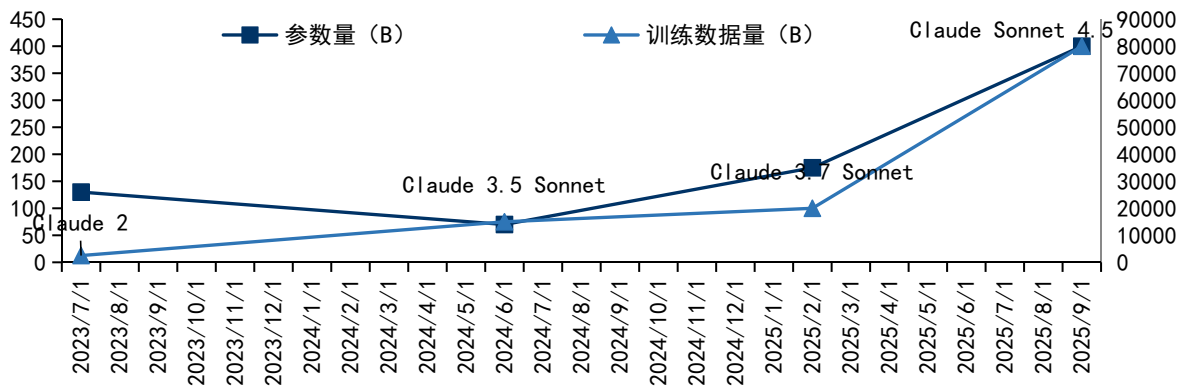
资料来源：Anthropic、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Anthropic：核心团队成员来自OpenAI，当前估值3500亿美元

- Anthropic成立于2021年5月，核心团队成员来自OpenAI，联合创始人Amodei兄妹曾分别担任OpenAI研发副总裁和安全与政策副总裁。25年11月英伟达和微软投资150亿美元后，目前估值已达到3500亿美元。
- 首款Claude模型上线于23年3月，并在24年3月Claude3开始分为不同尺寸的版本，并且在Claude 3 Opus首次全面超越同时期的GPT-4。25年2月发布AI编程应用Claude Code，截至25年末ARR已达到10亿美元。

图：Claude部分版本参数量和训练数据量预估



资料来源：lifearchitect.ai、国信证券经济研究所整理

图：Claude版本迭代情况

大版本	版本型号	发布时间	特点
Claude 1	Claude 1.0 / Instant	2023年3月	首款商用模型；引入“宪法AI”（Constitutional AI）理念。
Claude 2	Claude 2.0 / 2.1	2023年7月/11月	首次开放网页端；上下文提升至 200k；逻辑与代码大幅增强。
Claude 3	Opus / Sonnet / Haiku	2024年3月	首次拆分为三个版本，由强到弱依次是Opus / Sonnet / Haiku；Opus 性能首次全面超越同时期的 GPT-4。
Claude 3.5	Sonnet / Haiku	2024年6月/10月	Artifacts 功能上线；确立了全球最强代码模型的地位。
Claude 3.7	Sonnet (Reasoning)	2025年2月	首个“混合推理”模型；发布 Claude Code 终端开发工具。
Claude 4	Opus / Sonnet	2025年5月	Agent 时代开启；支持“扩展思考”模式与大规模工具并行调用。
Claude 4.5	Opus / Sonnet / Haiku	2025年9月-11月	2026年旗舰；实现全自主“计算机使用”能力，代码准确率近乎完美。

资料来源：Anthropic、国信证券经济研究所整理

图：Anthropic 融资及估值

融资轮次	时间	融资金额（亿美元）	投后估值（亿美元）	领投方 / 主要投资者
Series A	2021年5月	1.24		Jaan Tallinn（Skype联合创始人）， Eric Schmidt等
Series B	2022年4月	5.8		FTX首席执行官Sam Bankman-Fried领投，Caroline Ellison、Jim McClave等
Series C	2023年5月	4.5	50	Spark Capital 领投，谷歌、Salesforce Ventures、Sound Ventures、Zoom Ventures 等
Series D	2023年9月	12.5	200	Amazon
Series E	2024年3月	27.5	184	Amazon
Series E+	2025年3月	45	615	Lightspeed Venture Partners 领投，Bessemer Venture Partners等
Series F	2025年9月	130	1,830	ICONIQ Capital，Fidelity，Lightspeed等
Strategic	2025年11月	150	3,500	NVIDIA（100亿），Microsoft（50亿）

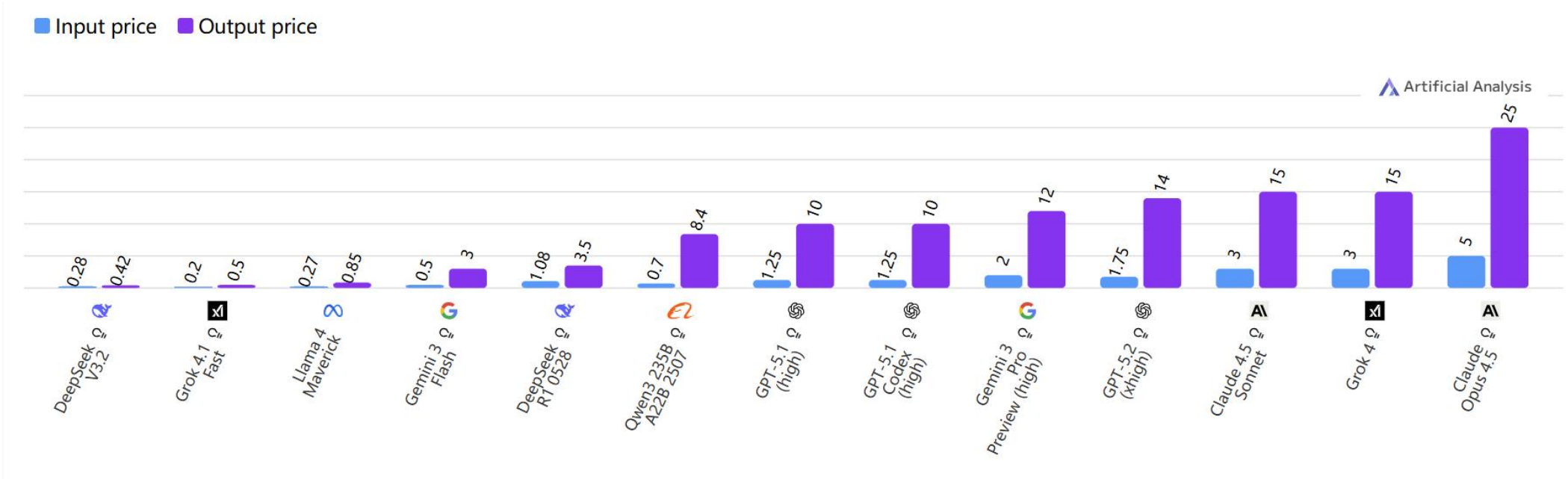
资料来源：Anthropic、CBInsights、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Anthropic：坚持2B路线毛利率更优，最早有望27年迎来现金流转正

- **Anthropic坚持2B路线，主要服务于对稳定性要求更高的客户，定价更高，因此预计毛利率水平较OpenAI更好。**目前Claude Opus 4.5是定价最贵的模型，百万tokens输出价格为25美元，远高于其他厂商的SOTA模型，因此Anthropic拥有更好的毛利率水平，据The Information，25年Anthropic毛利率已回到50%，并且未来几年仍将持续提升，预计28年达到约75%。考虑到随着收入规模的增加，训练成本（研发费用）占比将持续下降，预计27年Anthropic有望最先迎来现金流的转正。

图：主流模型输入输出定价（美元/百万tokens）



资料来源：artificial analysis、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Anthropic：坚持2B路线毛利率更优，最早有望27年迎来现金流转正

国信证券 GUOSEN SECURITIES

- **算力总投入持续增加，与多家云厂合作。**据The information，预计25年Anthropic的推理+训练算力投入月60亿美元，28年将增长至270亿美元，Anthropic与多家云厂合作，亚马逊Project Rainier为公司提供约100万颗Trainium芯片的算力支持，25年10月与谷歌签署了将租用100万颗TPU芯片（价值数百亿美元），同样在25年底还承诺向微软购买300亿美元的算力。
- **商业化进展：**据Anthropic CEO Dario Amodei，Anthropic过去三年收入每年增长10倍，预计今年底将达到80亿至100亿美元ARR。据路透社，25年公司整体收入预计将达到接近50亿美元，在API调用以外，公司25年推出了编程产品Claude Code，截至年底ARR已达到10亿美元，未来有望成为公司重要收入利润来源。

图：Anthropic收入毛利率预测

单位：百万美元	2024A	25Q1	25Q2	25Q3	25Q4E	2025E	2026E	2027E	2028E
ARR		1,600	3,200	5,000	9,000				
收入	381	400	800	1,250	2,250	4,700	15,200	38,900	70,000
YoY							223%	156%	80%
营业成本	739	200	400	625	1,125	2,350	6,080	11,670	17,500
营业成本中云服务占比（推理成本）	75%	75%	75%	75%	75%	75%	75%	75%	75%
云服务成本（云厂收入）	554	150	300	469	844	1,763	4,560	8,753	13,125
毛利率	-94%	50%	50%	50%	50%	50%	60%	70%	75%
毛利润	-358	200	400	625	1,125	2,350	9,120	27,230	52,500
研发费用（主要为训练算力成本）	1,500					4,000	8,000	12,837	15,400
研发费用率	394%					85%	53%	33%	22%
销售&管理费用	114					1,175	3,040	5,835	7,000
销售&管理费用率	30%					25%	20%	15%	10%
经营利润	-1,972					-2,825	-1,920	8,558	30,100
OPM	-518%					-60%	-13%	22%	43%

资料来源：The information、国信证券经济研究所整理测算

请务必阅读正文之后的免责声明及其项下所有内容

Grok：坚持大力出奇迹，与特斯拉形成协同

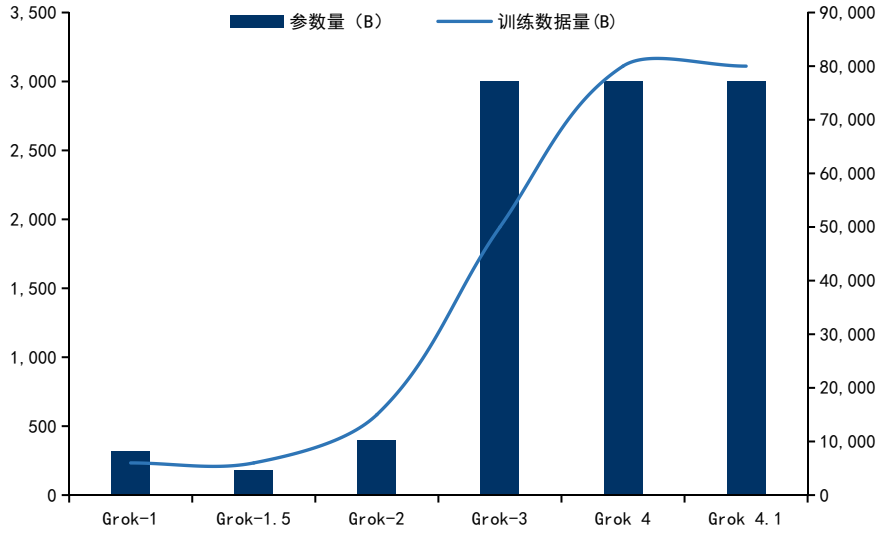
- **xAI在模型训练上坚持大力出奇迹。马斯克认为在通向AGI的过程中，“快速扩展算力与数据处理能力将是决定性因素”，强调xAI的优势在于1）资金能力；2）协同条件（Grok今年已集成进特斯拉的汽车系统）。**因此Grok2/3/4版本上完成了3-6个月一迭代的速度，在Grok3后迅速将算力资源投入到后训练，用10倍后训练算力投入推出了Grok4，迅速向SOTA基准的看齐。并且马斯克此前宣称26年将推出的Grok5将进一步扩展Scaling law的天花板，达到6万亿参数。
- **算力储备上：**X.AI 算力集群建设以 “超高速迭代、规模化部署、垂直整合电力” 为核心特征，核心项目包括Colossus I与Colossus II，主要聚焦 GPU 规模扩张 + 电力设施自建，目标2025年实现30万GB200芯片集群（此目标下仅GPU芯片所需资金就接近100亿美元）
- **推特、特斯拉的独家数据资源优势。**在训练过程中，未来的原生多模态模型更加依赖于图片、视频数据，而xAI能够基于特斯拉大量的自动驾驶数据提升模型的多模态能力。

图：Grok 4.1 thinking模型综合表现排行第三

Model 291 / 291	Overall	Expert	Hard Prompts	Coding	Math	Creative Writing	Instruction Following	Longer Query
gemini-3-pro	1	3	1	3	2	1	3	3
gemini-3-flash	2	7	5	8	3	2	6	6
grok-4.1-thinking	3	9	4	6	9	10	12	15
claude-opus-4-5-202…	4	2	2	1	7	5	1	2
claude-opus-4-5-202…	5	1	3	4	6	3	2	1
grok-4.1	6	19	9	13	19	15	17	13
gemini-3-flash (thi…	7	12	7	10	4	8	11	8
gpt-5.1-high	8	8	11	14	5	11	9	11
gemini-2.5-pro	9	14	15	24	11	4	10	10
claude-sonnet-4-5-2…	10	4	6	2	8	9	4	4
claude-opus-4-1-202…	11	10	8	5	12	6	5	5

资料来源：IMARENA.AI、国信证券经济研究所整理

图：Grok 历代模型参数量及训练数据量预估



资料来源：lifearchitect.ai、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 美股科技巨头股价和资本开支复盘展望

02 模型架构继续演化，Scaling law延续，多模态+长文本更成熟

03 模型厂发展路线分化，格局动态演进

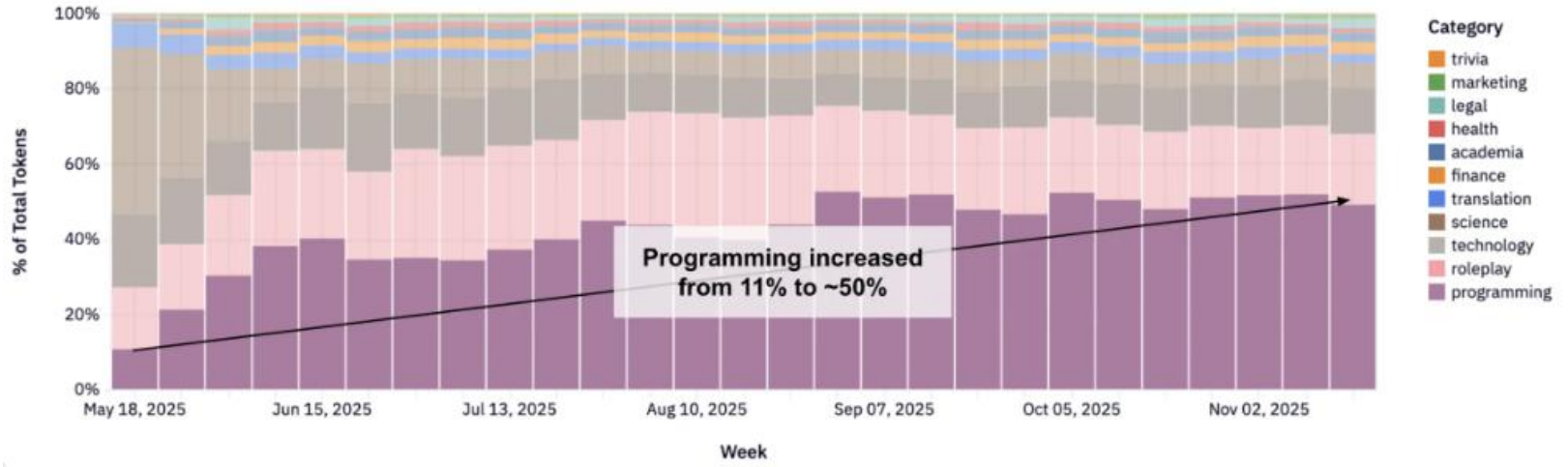
04 推理需求有望爆发，看好编程、Agent等应用

请务必阅读正文之后的免责声明及其项下所有内容

推理侧需求：Tokens消耗量持续增长，编程是主要需求

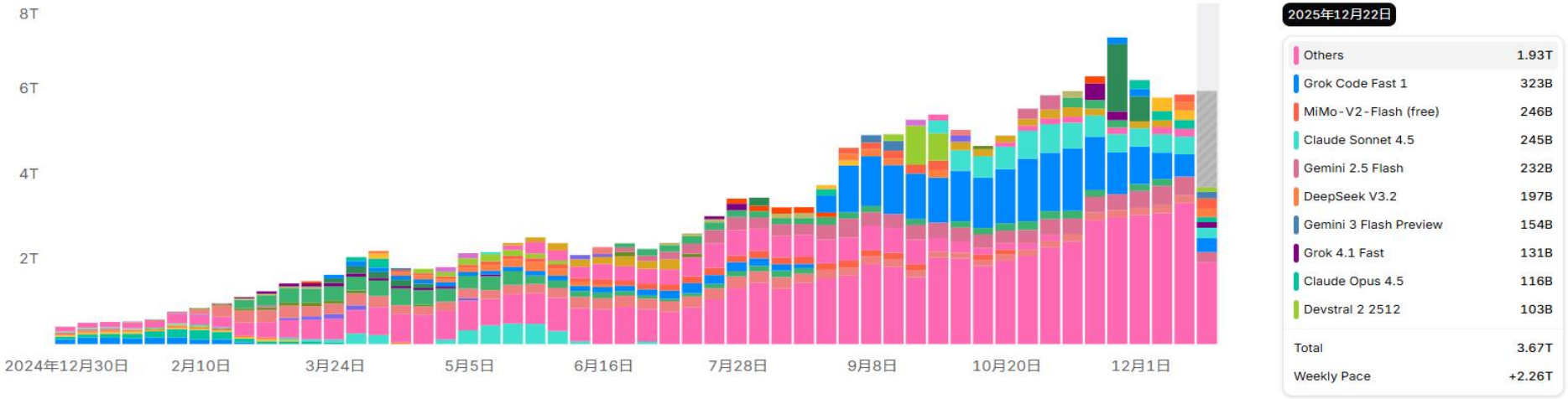
- 推理需求主要反映在tokens消耗的增长上，根据大模型聚合平台OpenRouter数据，25年的周tokens消耗量已从年初的5000亿增长至年末的3.67万亿。从tokens消耗结构看，编程是主要驱动因素，并且从单次请求的tokens结构来看，编程也是主要的增长驱动力，与编程相关的任务一直需要最大的输入上下文。

图：OpenRouter 周度tokens消耗结构（按使用场景）



资料来源：openrouter、国信证券经济研究所整理

图：OpenRouter 周度tokens消耗情况



资料来源：openrouter、国信证券经济研究所整理

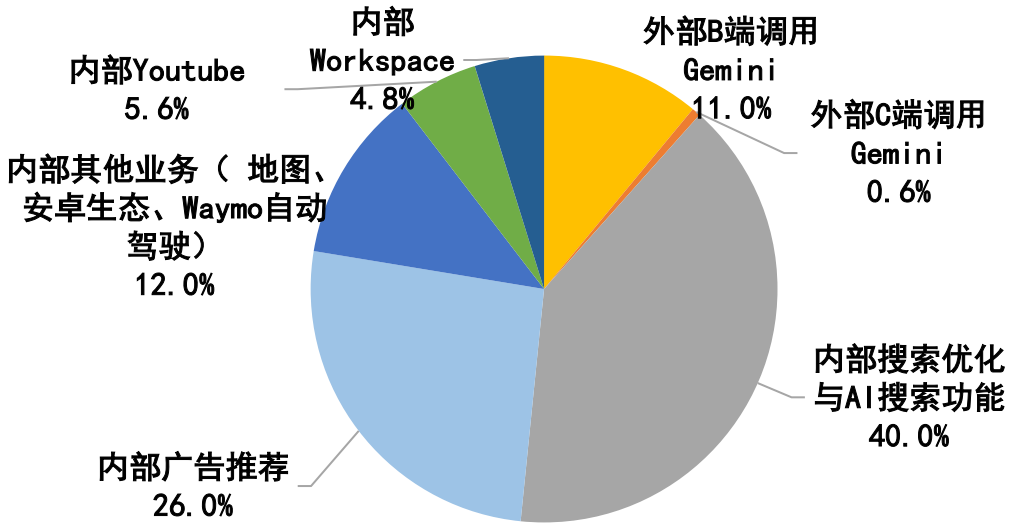
请务必阅读正文之后的免责声明及其项下所有内容

推理侧需求：25年更多使用场景聚焦在大模型企业内部

国信证券 GUOSEN SECURITIES

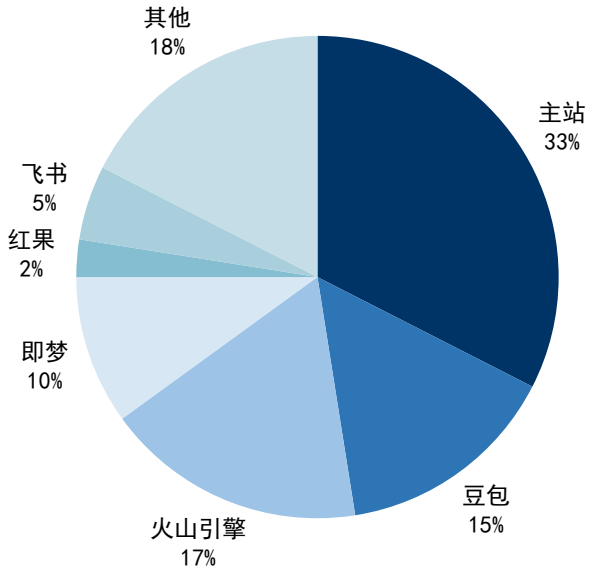
- 尽管外部需求快速增长，但从结构上看，25年各模型厂商的消耗更多还是聚焦于自身业务发展。例如谷歌Gemini，我们测算其内部消耗占比约90%，OpenAI的模型tokens消耗则更多用于ChatGPT，国内阿里、字节Tokens消耗也都以内部场景占比更多。
- **展望26年，1）随着模型在多模态领域的突破，在视频、音频等场景的用量增加，预计推理侧将迎来Tokens的进一步大幅增长**（文本1个汉字大约消耗1个token，1秒1080p 24帧视频消耗2.5万个tokens，1张图像大约消耗4000个tokens）；**2）随着模型厂商率先完成自身业务接入大模型，26年对外的需求和用量占比将增加。**

图：谷歌月度tokens调用量场景占比



资料来源：Gemini、国信证券经济研究所整理测算

图：字节Tokens消耗结构



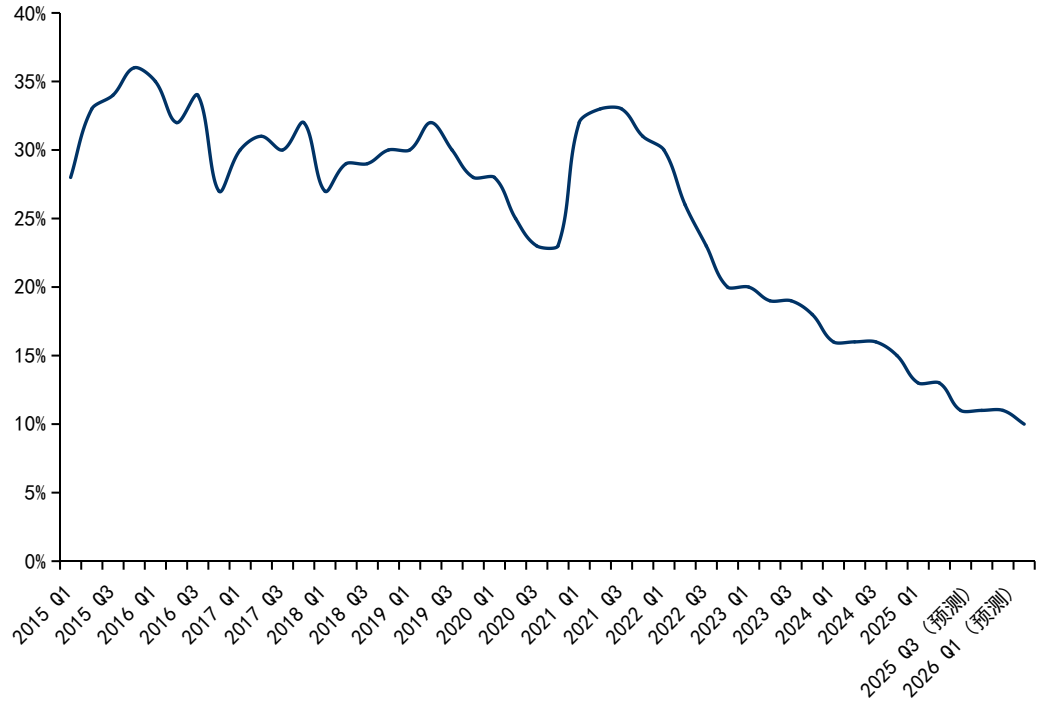
资料来源：豆包、国信证券经济研究所整理测算

请务必阅读正文之后的免责声明及其项下所有内容

推理侧需求：大模型降低软件开发壁垒，成熟SaaS企业收入持续降速

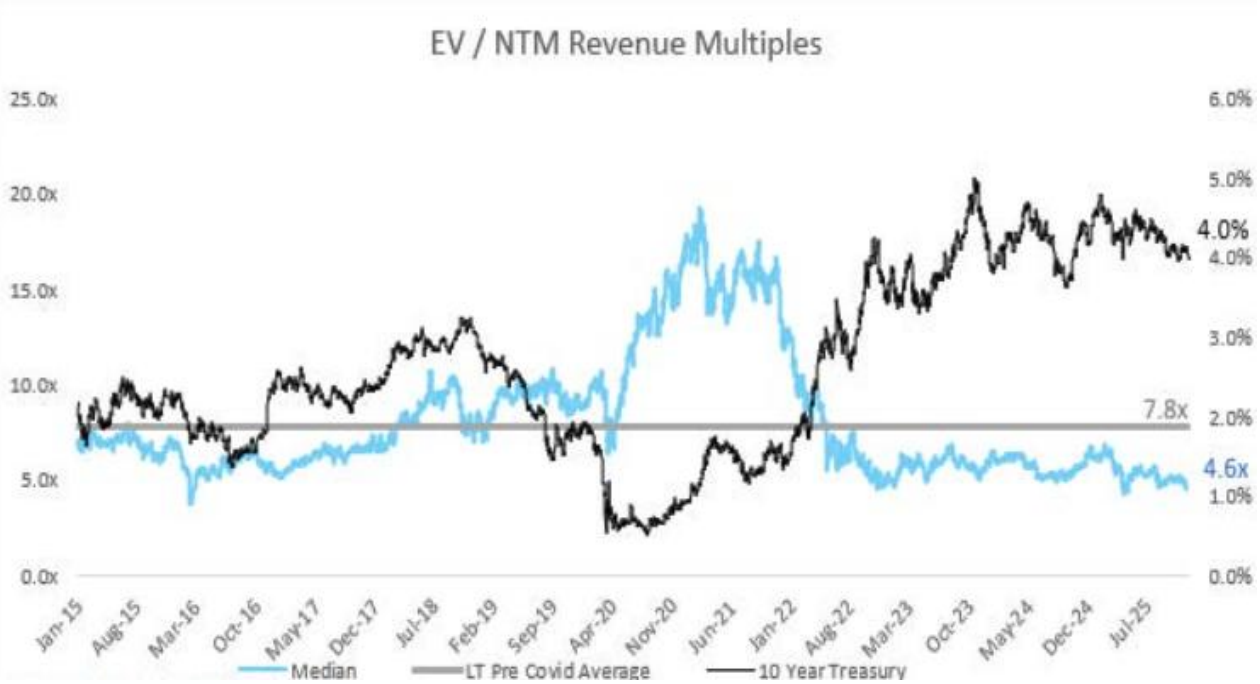
- **大模型降低了软件的开发壁垒，也在某些场景，重新用模型定义了工作流，改变了软件的使用流程，比如photoshop等等，导致SaaS板块收入不断降速。**SaaS板块在21年的收入增速高点后就进入持续降速阶段，伴随23年起大模型冲击下SaaS公司原本稳定的现金流壁垒开始面临扰动，估值也持续降低。

图：美股SaaS板块收入增速



资料来源：Capital IQ、国信证券经济研究所整理

图：美股SaaS公司估值情况



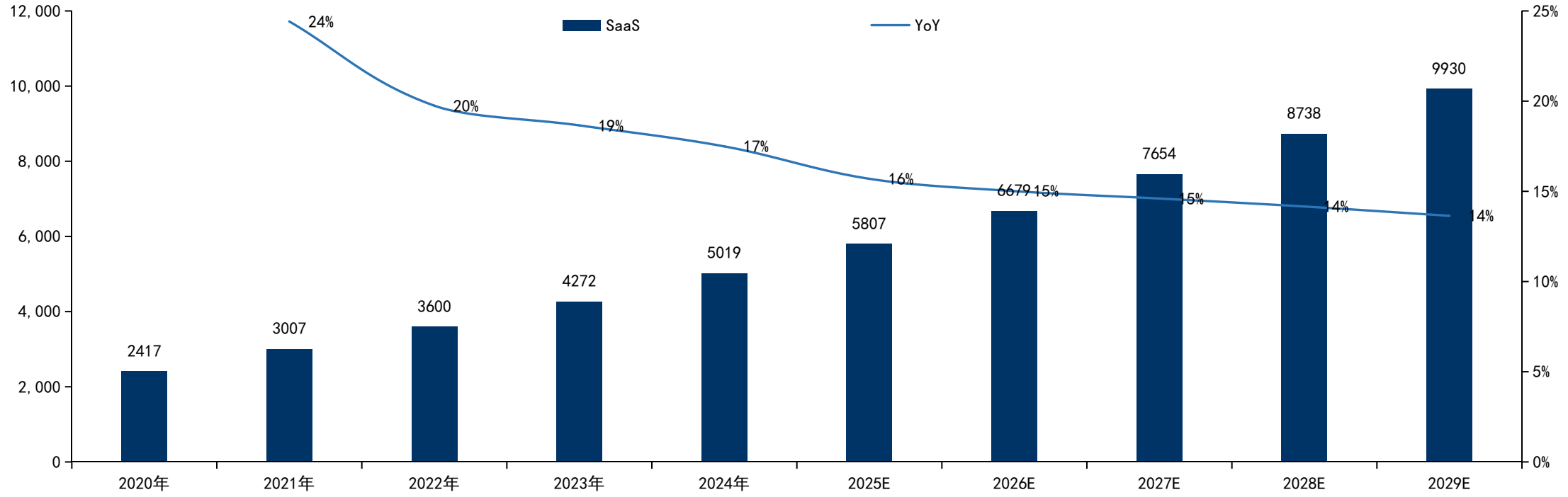
资料来源：ALTIMETER、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

推理侧需求：需求上升，玩家重构，29年全球SaaS市场空间或达1万亿美元

- **SaaS代表的AI应用市场长期有望达到1万亿美元。**据IDC数据，预计2025年全球SaaS市场将达到5800亿美元，同比+16%，并且2029年将达到近1万亿美元规模，长期来看，SaaS市场都将成为AI应用，但存在路径的分化，部分公司被大模型替代，部分公司完成AI转型，还有部分原生AI应用贡献增量。

图：全球SaaS市场规模（亿美元）及增速



资料来源：IDC、国信证券经济研究所整理

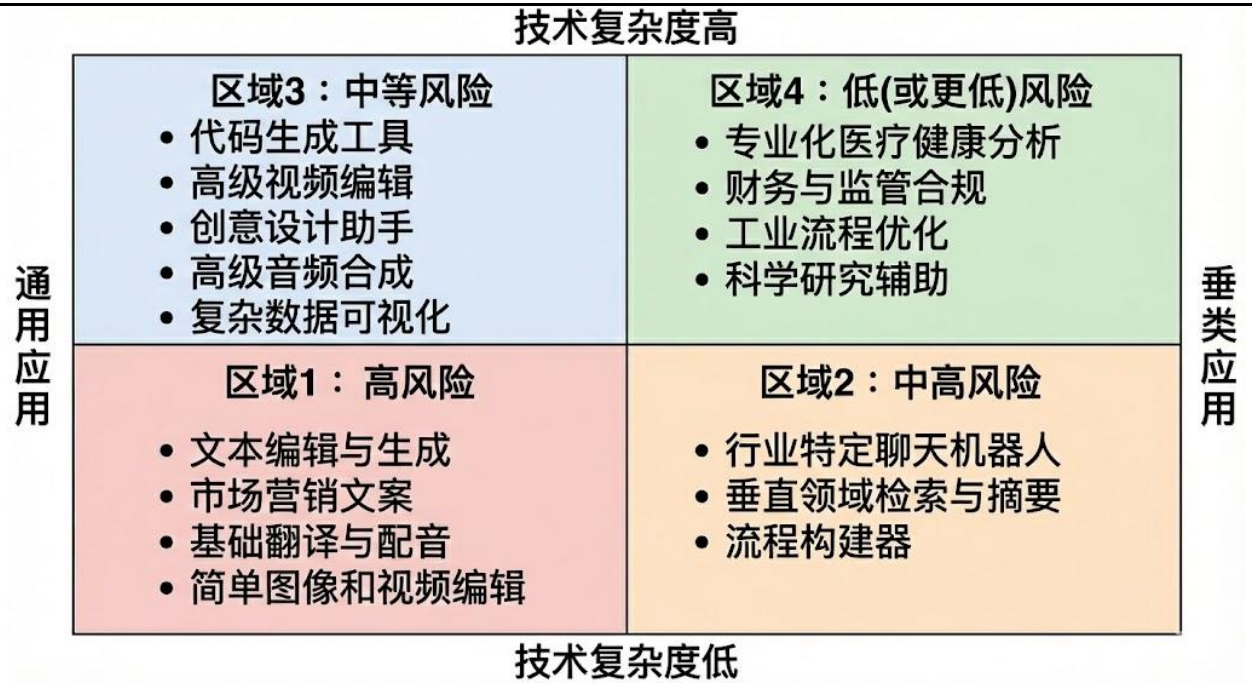
请务必阅读正文之后的免责声明及其项下所有内容

推理侧需求：通用型、准确度要求低的SaaS将被大模型替代

国信证券 GUOSEN SECURITIES

- **我们预计大模型会颠覆轻量级的工具，以及对准确度要求没有那么高的工具，比如营销工具、翻译工具、邮件回复机器人等。**
- 例如根据maginative，营销SaaS Jasper在21年推出后，18个月达到15亿美元估值，但是在ChatGPT推出后面临巨大冲击，只能通过裁员和战略调整，将重心全面转向大型企业客户，通过深挖品牌定制构筑护城河。
- **拥有数据壁垒，主要在垂类细分场景中布局，软件定义工作流程较复杂，或对准确度要求极高的行业，被大模型替代的风险较小，比如医疗、能源、会计、安全等领域。**

图：不同SaaS场景面临大模型替代风险



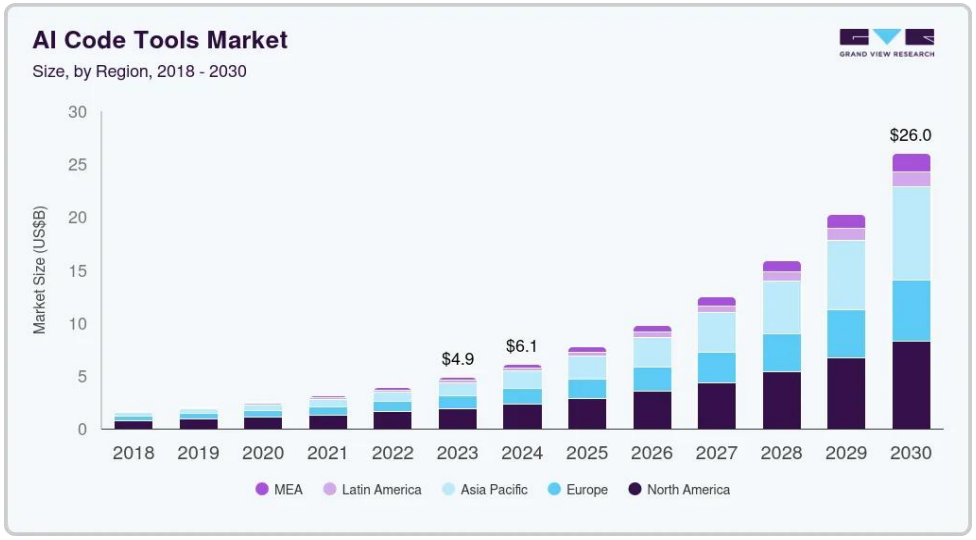
资料来源：Capital IQ、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

推理侧需求：编程和企业Agent商业化落地最快

- **编程和企业Agent是商业化落地最快的场景。**目前AI编程是落地最快的AI应用场景，据Grand view research，AI coding市场2030年有望达到260亿美元。另据CB insights对未上市独角兽的统计，原生AI应用中，估值50亿以上的公司编程占比最多，包括Cursor的母公司Anysphere目前估值接近300亿美元，25年收入已达到10亿美元规模，此外还有两款未上市的企业agent产品也达到了100亿美元估值，ARR超过1亿美元，Salesforce的Agent产品ARR已经达到5亿美元。

图：AI编程市场规模预测



资料来源：Grand view research、国信证券经济研究所整理

图：OpenAI tokens调用量超过1万亿的客户中成熟SaaS公司

公司	业务简介	AI相关进展
Duolingo	语言学习 App，以游戏化课程设计著称	25Q3Duolingo Max渗透率达到9%
Salesforce	企业级 CRM 与云计算服务提供商	AI 平台 Agentforce 25Q3 ARR突破5亿美元，调用openai tokens超过1万亿，
Shopify	全球电商平台，为中小企业提供建站与支付解决方案	使用 Shopify Magic 和 Sidekick 的商家，平均客单价（AOV）通常能提升 15-20%，同时运营成本可降低 30%
Canva	在线图形设计平台，整合 AI 图像生成与排版	AI 工具每月产生超过 8 亿次 交互
Datadog	云监控与数据可视化平台	25Q3AI 原生客户贡献了总收入的 12%

资料来源：OpenAI、国信证券经济研究所整理

图：一级市场中估值约50亿美元及以上的原生AI应用公司

公司	估值（亿美元）	收入	行业
Anysphere（Cursor）	293	收入10亿美元	编程
Perplexity	200	ARR 2亿	搜索
Sierra	100	ARR 1亿	企业agent
Reflection AI	80	收入460万	编程
Mercor	100	ARR 5亿	数据标注和模型评估
Cognition AI	102	ARR 1.5亿	企业agent
Lovable	66	ARR 2亿	编程
Cyera	90	ARR 1亿	网络安全
Helsing	53.7	收入1-2亿	防务
NinjaOne	50	收入3-5亿	网络安全
Quince	45	ARR 5亿	电商

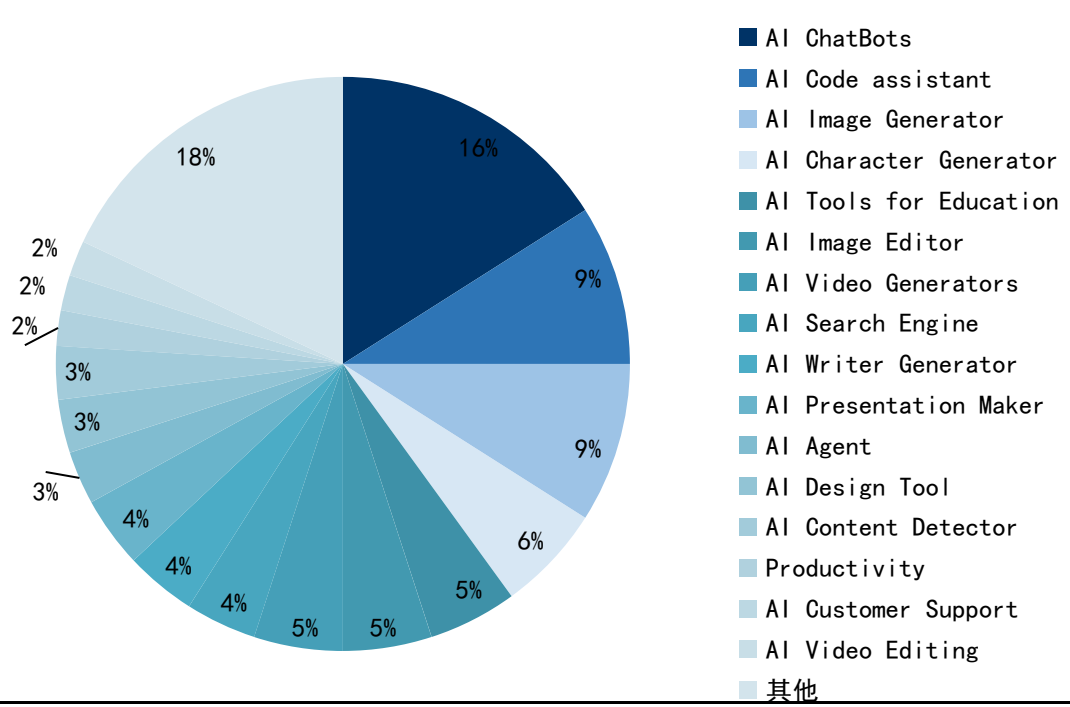
资料来源：CB insights、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

推理侧需求：从用户量看，chatbot、内容创作、AI coding是主流应用

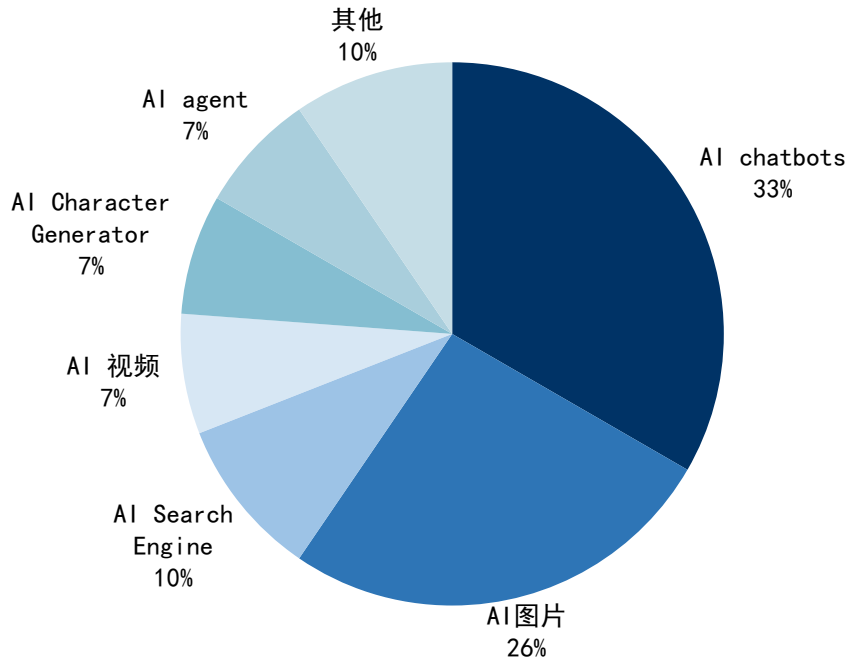
- 据Similiar web数据，11月AI应用网站流量数据来看，目前主要应用场景集中在chatbot、coding、内容创作（视频、图像、写作生成编辑）、教育、搜索、AI角色陪伴等。流量排名前100的网站中，Chatbot类占比16%，Coding占比9%。手机端则以Chatbot和内容创作类产品为主，11月MAU超过2000万的产品中，chatbot类产品占比达到50%。
- 从产品演化趋势来看，Chatbot的功能边界在不断拓宽，从对话能力演化到陪伴、内容创作、个人agent等多维度的功能。

图：11月AI应用网站流量排行前100中分类型占比



资料来源：AI产品榜、国信证券经济研究所整理

图：11月AI应用移动端月活1000万以上产品分类型占比



资料来源：AI产品榜、国信证券经济研究所整理

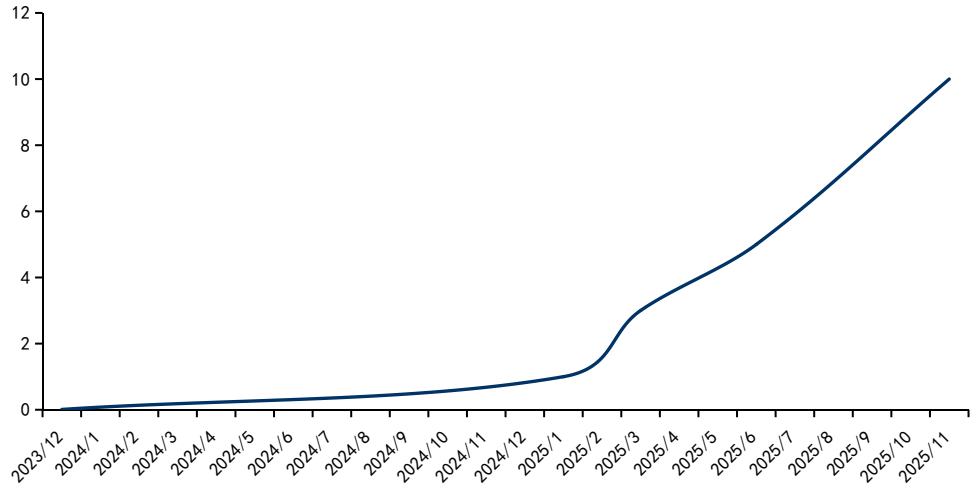
请务必阅读正文之后的免责声明及其项下所有内容

Cursor：AI编程龙头，ARR达10亿美元，估值近300亿美元

国信证券 GUOSEN SECURITIES

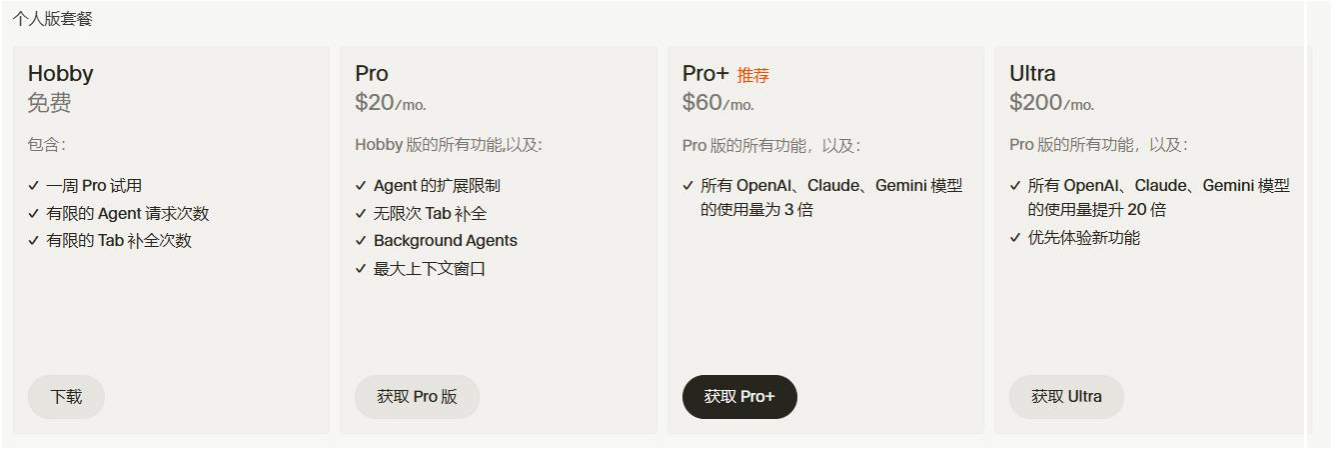
- **Cursor 是一款基于 AI 的代码编辑器，**由 Anysphere 团队在2023年3月正式推出，基于VS Code 打造、支持自然语言编程与聊天调试，推出几个月后上线了内联命令与上下文对话功能，让开发者能像和同事聊天一样“写代码”。**Cursor目前支持Composer 1（自研模型）、GPT-5、Claude、Gemini、Grok几个目前最强模型，**其自研模型Composer1于25年10月推出，是专为编程场景优化的“代理式”模型，生成速度比同级别的 Claude 3.5 Sonnet 或 GPT-4o 快 4 倍。**与竞品相比，Cursor在上下文理解、可解释性与隐私部署 上的优势明显，产品体验更贴近程序员的思维节奏。**
- **商业化方面，**根据公司官方数据及CBinsights，Cursor23年末ARR达到100万美元，24年中开始进入加速期，发布 Composer 功能后，用户量呈指数级增长，25年1月ARR突破1亿美元，3月达到3亿美元，付费用户超过 36 万；4 月日活跃用户突破 100 万，企业客户数量攀升至 1.4 万，**年底ARR已达到10亿美元，最新估值已达到293亿美元，是目前一级市场中估值最高的AI应用公司。**

图：Cursor ARR（亿美元）



资料来源：Z finance、国信证券经济研究所整理

图：Cursor个人版定价



资料来源：Cursor、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Manus：通用Agent，上线8个月ARR达到1亿美元

- **Manus是一款通用Agent，其功能包括网站开发、PPT制作、图片设计、广泛研究、操作浏览器等。**与目前的Chatbot类产品相比，Manus能够直接执行任务并交付结果，而非仅提供建议，设定目标后自主规划路径，无需持续监督，并且直接生成可用的应用、报告和演示文稿。
- 25年3月发布，5月开放注册后首日用户突破100万。10月Manus 1.5发布，任务完成时间从 15 分钟缩短到 4 分钟以下。商业模式上结合固定订阅费用与基于积分（Credit）的使用量计费。**12月ARR超过1亿美元，累计处理Tokens超过147万亿，创建 8000万+ 虚拟电脑实例。12月30日，Meta以数十亿美元估值对Manus完成收购，公司创始人肖弘出任Meta 副总裁，蝴蝶效应保持独立运作。**

图：Manus收费标准



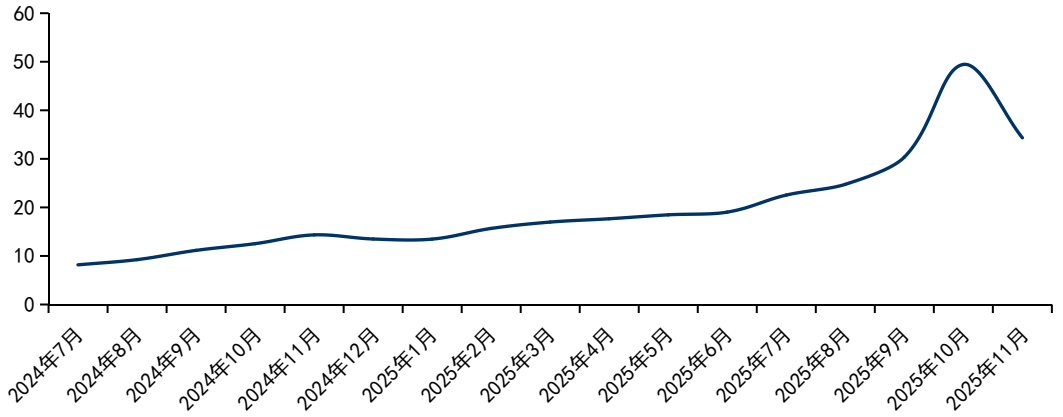
资料来源：Manus、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Perplexity：从AI搜索到AI Agent，ARR接近2亿美元

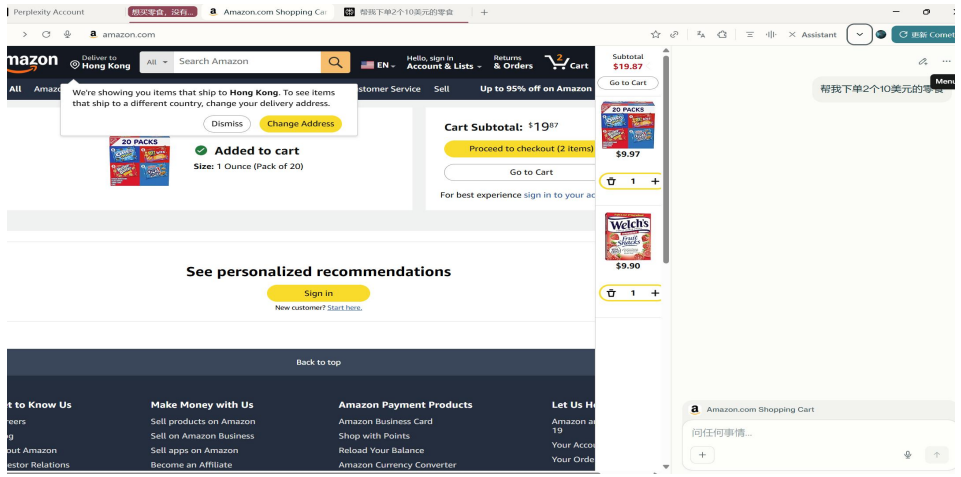
- **Perplexity AI 是一款于2022年8月推出的AI搜索引擎。**与传统的搜索引擎不同，它的目标不是提供一堆网页链接，而是直接通过自然语言回答你的问题，并为每一个结论提供透明的引文来源。2025年7月，Perplexity推出了基于Chromium内核的AI浏览器Comet，该搜索引擎可以执行各种任务，例如生成文章摘要、描述图像、进行主题研究、撰写电子邮件、以及帮助电商购物。10月随着Comet面向全球全量上线，带动了Perplexity的访问量达到峰值，但是伴随竞争对手动作（OpenAI 快速跟进 SearchGPT等），10月的流量增长并没有实现很好留存，11月用户再次回到正常增长水平。
- Perplexity的商业模式包含广告和订阅，订阅服务下，Perplexity Pro的权益包括回答中更多的引用量、更新的外部模型以及图像视频生成等功能，**25年末公司ARR已接近2亿美元，估值达到200亿美元。**
- Perplexity聚焦于深度研究与企业级搜索解决方案，从实时搜索切入，逐步扩展至医疗、教育、企业服务领域。但同时还需要面临版权问题，25年12月《纽约时报》就指控其“非法”复制数百万篇文章。

图：Perplexity 网页端MAU（百万）



资料来源：Similiar Web、国信证券经济研究所整理

图：Comet浏览器中自动帮用户完成搜索商品并添加购物车



资料来源：Comet、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

推理侧需求：端侧落地加速，手机Agent塑造全新流量入口

- **伴随模型在多模态和Agent调用能力上的提升，AI手机成为重要落地场景，**用户通过AI Agent，除了能够实现修图、问答等功能，更重要的是具备屏幕感知、跨应用操作的能力，帮助用户大幅节省操作手机的时间。
- 以字节豆包手机助手为例，这是一款豆包和手机厂商在操作系统层面合作的手机AI助手，目前搭载在中兴通讯的nubia M153上，且豆包手机助手拥有系统级的操作权限，能够后台执行操作，这带来了用户效率的大幅改善。
- **技术实现：**豆包手机助手的实现分两步：一是训练小参数量的视觉理解模型，基于手机和电脑界面训练，用于理解界面元素位置；二是通过UI pass工程框架调用系统权限执行模拟点击事件。模型部署上，在云端模型为几十B到30B规模，与努比亚手机AI键的视频识万物功能、豆包APP视频通话功能使用的模型一致，依赖云端算力运行。本地则采用分组模型，核心为13B规模（由1B的小路由器+12B的分组模型组成），12B模型分为对话、理解等工作组；此外还有多个1B左右的编解码加速型模型，均在端上运行。
- **当前挑战：**1）技术成熟度不足：处理复杂信息时，理解模型会受阻，并且在执行长任务时仍然出现出错的情况；2）各APP的反制：部分电商页面存在安全机制或干扰信息，会影响视觉模型对界面的理解，包括微信登

图：豆包手机助手购物、预定餐厅实机演示



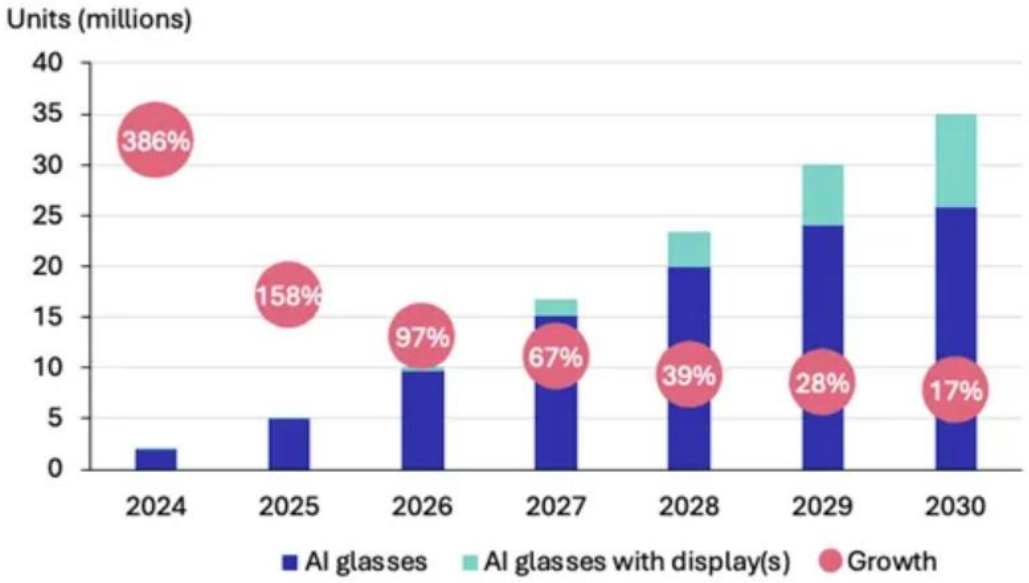
资料来源：豆包手机助手、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

推理侧需求：端侧落地加速，AI眼镜有望迎来放量

- **眼镜能够解决手机笨拙、无法持续在线的痛点，AI眼镜作为全新端侧场景也将为AI在推理侧带来巨大的增量需求。**据Omdia，2025年全球AI眼镜出货量将达到510万台，同比+158%，26年有望达到1000万台，其中中国市场出货量预计将达到120万台，占比12%
- **25年伴随AI功能的落地，AI眼镜迎来了拐点。**雷朋眼镜前期仅靠硬件前3个月仅卖5万部，上线AI的Llama功能后3个月销量增至50万部，增长10倍。目前市场中仍以Meta一家占据主要份额（预计25年达到75%），但苹果、谷歌正在积极入局。**展望26年，26H2开始将逐步进入AI眼镜的爆发期**，我们预计Meta将推出全彩弹幕式新品（Halo升级版），谷歌也将在26H2推出多款AI眼镜，Gemini 3也将在AI眼镜上落地，整体硬件的放量有望在2027年出现。国内来看，阿里此前推出夸克AI眼镜，最大化阿里生态优势，结合了通义千问、夸克AI、高德导航、支付宝“看一看”、淘宝搜同款、飞猪出行提醒等AI功能。

图：AI眼镜出货量预测



资料来源：Omdia、国信证券经济研究所整理

图：夸克AI眼镜用支付宝直接进行扫码支付



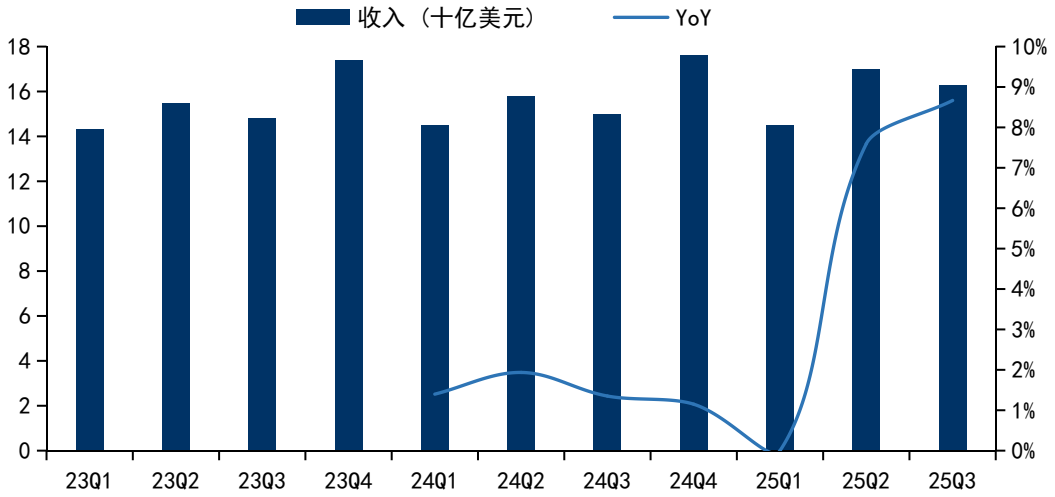
资料来源：夸克AI眼镜、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

大模型厂商发力企业业务带动分销伙伴需求

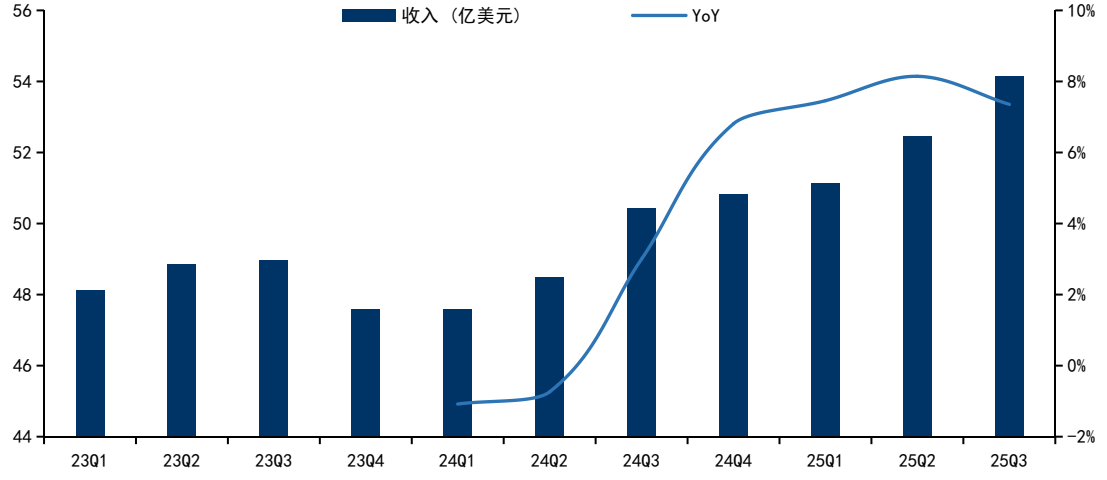
- 谷歌、OpenAI都开始发力企业业务，10月谷歌发布企业级AI解决方案Gemini Enterprise，OpenAI则在8月将ChatGPT Team升级为ChatGPT Business，目前OpenAI 目前已为超过 700 万个 ChatGPT 席位提供服务，ChatGPT Enterprise 的席位数量同比增长约 9 倍，已有超过 9,000 家组织累计处理了超过 100 亿 token，其中近 200 家组织的处理量已超过 1 万亿 token。
- **发力企业业务带动分销伙伴需求，**1）例如IBM通过企业级开发平台watsonx.ai给客户提供包括Claude等多款模型的调用，25Q3生成式 AI 业务订单额达到95亿美元，软件的订阅服务同时撬动了相关咨询（安装、调试等）以及数据存储工具的需求，带动公司咨询业务收入从负转正。**2）埃森哲**推出了模型聚合平台Accenture Switchboard，并且在25年12月与Anthropic合作Accenture Anthropic Business Group，为 30000 名 埃森哲专业人员提供关于 Claude 系列模型的深度认证和培训，这些人员驻扎在500强企业中教客户如何使用Claude。 此外埃森哲还加深了与OpenAI的合作，预计随着企业业务的加速，收入端逐渐将有所体现。 **3）根据公司财报，Cognizant全年累计大单（1亿美元以上）数达到 16 个，同比增长 40%，客户不仅需要咨询AI，还需要找Cognizant重构旧系统以落地AI项目**

图：IBM收入及增速



资料来源：公司财报、国信证券经济研究所整理

图：Cognizant收入及增速



资料来源：公司财报、国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

风险提示

第一，宏观经济波动。若宏观经济波动，公司业务、产业变革及新技术的落地节奏或将受到影响。

第二，下游需求不及预期。若下游AI需求不及预期，相关的AI研发投入增长或慢于预期，致使行业增长不及预期。

第三，核心技术水平升级不及预期的风险。AI大模型研发进度落后，AIGC相关产业技术壁垒较高，核心技术难以突破，影响整体进度。

第四，AI快速迭代、平权化下竞争加剧，影响云业务利润率。

请务必阅读正文之后的免责声明及其项下所有内容

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）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，也即报告发布日后的6到12个月内公司股价（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。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。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口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032