谷歌 (GOOGL US)

搜索王者站在十字路口，能否抢回 AI 主导权？

华泰研究 首次覆盖

2025 年 1 月 04 日 | 美国 互联网

首次覆盖，给予“买入”评级，目标价 235 美元。谷歌主营业务为搜索广告、YouTube 及云计算。我们认为谷歌拥有完备 AI 生态，凭借大规模搜索流量、自研 AI 芯片 TPU、Gemini 大模型，以及全链条云业务部署，将受益于 AI 应用下半场浪潮，并有望抢回 AI 的主导权，引领估值持续抬升。

深耕 AI 多年，凭借 Gemini、TPU、搜索和云生态，抢回 AI 主导权正当时

自 ChatGPT 空降后，市场普遍认为谷歌 AI 技术在走下坡。但我们认为尽管 OpenAI 凭借微软加持抢占市场，谷歌在 AI 研究根深蒂固。谷歌早在 2016 年已洞悉降低 AI 计算 TCO 的重要性，自研 AI 芯片 TPU 并经历多次迭代，对比其他科技巨头具备先发优势。谷歌也在 2017 年发布大模型奠基算法 Transformer，随后在 2018 年发布蛋白结构预测系统 AlphaFold，发明者在 24 年荣获诺贝尔化学奖。凭借 TPU 和 Gemini 2 新大模型，以及庞大的搜索生态数据，叠加全链条云布局，我们认为谷歌抢回 AI 主导权正当时。

广告业务高搜索渗透率或能维持，主要关注 AI 协同主营业务进展

我们预计 Services 业务 FY24/25/26 营收为 3046/3332/3632 亿美元。谷歌搜索引擎在全球拥有高市占率，我们认为将继续维持。Perplexity 等已将聊天机器人与搜索有机结合，因此将强劲 AI 实力有效融入到谷歌核心产品已成为未来挑战之一。近期谷歌推出 Gemini Deep Search 有所尝试，我们看好谷歌在打磨新搜索商业模式之后，凭借其 AI 技术积淀和庞大的应用生态，在 AI 时代继续保持搜索王者地位。若后续观察到 DeepMind 管理团队融入搜索等核心生态，我们认为或为谷歌释放更积极的整合信号。

IaaS-PaaS-SaaS 全链条布局，AI 入云份额持续提升，与头部差距缩小

我们预计谷歌云业务 FY24/25/26 营收为 434/553/694 亿美元。24Q3 谷歌云营收增速为 35%，对比 AWS 和微软的 19%和 34%，持续巩固其全球第三大云地位，市占率从 19Q4 的 6%攀升到 24Q3 的 12%。谷歌云的 IaaS 市占率更是前五大提供商中增速最高。我们认为，目前谷歌云已将市场策略从构建技术转变为产品和解决方案，叠加 AIaaS，持续赢得中大型企业青睐。

首次覆盖给予“买入”评级，目标价 235 美元，25 年 26.0x PE

我们认为谷歌当前被市场低估，Forward PE 为 20.7x，低于五年历史均值 20.9x，也低于广告和科技可比均值 23.3x 和 39.0x。历史上谷歌近两次估值上升分别由 2018 年企业上云(峰值 23.0x)和 2020 年疫情催生居家办公(峰值 26.9x)驱动，但公司在 2023 年 AI 浪潮并未跟上(峰值 20.9x)。我们认为谷歌在 AI 技术壁垒逐步兑现后将迎来价值重估，应重返历史估值高位，我们给予 25 年 26.0x PE，预计 24/25/26 年净利润为 947/1106/1287 亿美元。

风险提示：AI 技术落地不及预期、行业竞争激烈、反垄断监管变化等。

经营预测指标与估值

会计年度	2022	2023	2024E	2025E	2026E
营业收入 (美元百万)	282,836	307,394	350,355	391,565	436,710
+/-%	9.78	8.68	13.98	11.76	11.53
归属母公司净利润 (美元百万)	59,972	73,795	94,692	110,645	128,651
+/-%	(21.12)	23.05	28.32	16.85	16.27
EPS (美元，最新摊薄)	4.56	5.80	7.74	9.04	10.51
ROE (%)	23.62	27.36	31.26	32.10	33.18
PE (倍)	41.27	32.42	24.49	20.96	18.02
PB (倍)	9.66	8.44	7.19	6.32	5.68
EV EBITDA (倍)	28.21	24.82	19.28	16.44	14.25

资料来源：公司公告、华泰研究预测

投资评级(首评)：	买入
目标价(美元)：	235.00

研究员 何翩翩
SAC No. S0570523020002 purdyho@htsc.com
SFC No. ASI353 +(852) 3658 6000

研究员 夏路路
SAC No. S0570523100002 xialulu@htsc.com
SFC No. BTP154 +(852) 3658 6000

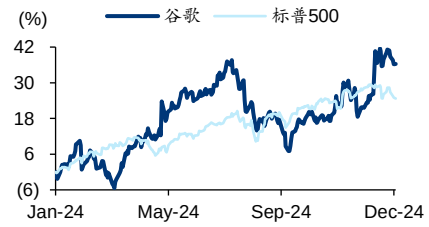
研究员 丁骄琬
SAC No. S0570523040003 dingjiaowan@htsc.com
SFC No. BPJ942 +(86) 21 2897 2228

联系人 易楚妍
SAC No. S0570124070123 yichuyan@htsc.com
+(86) 21 2897 2228

基本数据

目标价 (美元)	235.00
收盘价 (美元 截至 1 月 2 日)	189.43
市值 (美元百万)	2,318,813
6 个月平均日成交额 (美元百万)	4,542
52 周价格范围 (美元)	130.19-201.42
BVPS (美元)	22.74

股价走势图



资料来源：S&P

盈利预测

利润表

会计年度 (美元百万)	2022	2023	2024E	2025E	2026E
营业收入	282,836	307,394	350,355	391,565	436,710
销售成本	(126,203)	(133,332)	(146,648)	(162,779)	(180,685)
毛利润	156,633	174,062	203,707	228,786	256,025
销售及分销成本	(26,567)	(27,917)	(28,351)	(30,193)	(32,307)
管理费用	(15,724)	(16,425)	(14,809)	(15,920)	(16,556)
其他收入/支出	(39,500)	(45,427)	(49,292)	(54,074)	(60,060)
财务成本净额	1,817	3,557	2,200	3,585	6,248
应占联营公司利润及亏损	(337.00)	(628.00)	(628.00)	(628.00)	(628.00)
税前利润	71,328	85,717	111,322	130,052	151,216
税费开支	(11,356)	(11,922)	(16,630)	(19,407)	(22,565)
少数股东损益	0.00	0.00	0.00	0.00	0.00
归母净利润	59,972	73,795	94,692	110,645	128,651
折旧和摊销	(13,475)	(11,946)	(12,373)	(14,264)	(16,216)
EBITDA	82,986	94,106	121,495	140,730	161,184
EPS (美元，基本)	4.59	5.84	7.74	9.04	10.51

资产负债表

会计年度 (美元百万)	2022	2023	2024E	2025E	2026E
存货	2,670	0.00	4,281	470.96	4,804
应收账款和票据	132,141	134,832	137,109	146,383	149,764
现金及现金等价物	21,879	24,048	20,784	49,028	68,797
其他流动资产	8,105	12,650	15,180	16,698	17,583
总流动资产	164,795	171,530	177,354	212,581	240,947
固定资产	112,668	134,345	165,496	192,216	204,742
无形资产	31,044	29,198	29,198	29,198	29,198
其他长期资产	56,757	67,319	69,154	71,478	73,844
总长期资产	200,469	230,862	263,848	292,891	307,784
总资产	365,264	402,392	441,202	505,472	548,731
应付账款	65,392	77,677	74,503	94,417	93,085
短期借款	0.00	0.00	0.00	0.00	0.00
其他负债	3,908	4,137	4,137	4,137	4,137
总流动负债	69,300	81,814	78,640	98,554	97,222
长期债务	27,202	25,713	25,742	26,003	26,213
其他长期债务	12,618	11,486	14,347	13,918	16,799
总长期负债	39,820	37,199	40,089	39,921	43,013
股本	68,184	76,534	76,534	76,534	76,534
储备/其他项目	187,960	206,845	245,939	290,464	331,963
股东权益	256,144	283,379	322,473	366,998	408,497
少数股东权益	0.00	0.00	0.00	0.00	0.00
总权益	256,144	283,379	322,473	366,998	408,497

估值指标

会计年度 (倍)	2022	2023	2024E	2025E	2026E
PE	41.27	32.42	24.49	20.96	18.02
PB	9.66	8.44	7.19	6.32	5.68
EV EBITDA	28.21	24.82	19.28	16.44	14.25
股息率 (%)	0.00	0.00	0.26	0.31	0.36
自由现金流收益率 (%)	3.03	2.70	2.25	4.11	4.39

现金流量表

会计年度 (美元百万)	2022	2023	2024E	2025E	2026E
EBITDA	82,986	94,106	121,495	140,730	161,184
融资成本	(1,817)	(3,557)	(2,200)	(3,585)	(6,248)
营运资本变动	29,759	12,264	(9,732)	14,450	(9,046)
税费	(11,356)	(11,922)	(16,630)	(19,407)	(22,565)
其他	(8,077)	10,855	4,071	9,238	17,859
经营活动现金流	91,495	101,746	97,003	141,425	141,184
CAPEX	(31,485)	(32,251)	(43,200)	(41,040)	(28,728)
其他投资活动	11,187	5,188	702.17	(2,696)	501.65
投资活动现金流	(20,298)	(27,063)	(42,498)	(43,736)	(28,226)
债务增加量	996.00	(1,489)	28.50	261.40	210.40
权益增加量	6,410	8,350	0.00	0.00	0.00
派发股息	0.00	0.00	0.00	(6,121)	(7,152)
其他融资活动现金流	(77,163)	(78,954)	(57,798)	(63,585)	(86,248)
融资活动现金流	(69,757)	(72,093)	(57,770)	(69,444)	(93,189)
现金变动	1,440	2,590	(3,264)	28,245	19,768
年初现金	20,945	21,879	24,048	20,784	49,028
汇率波动影响	(506.00)	(421.00)	0.00	0.00	0.00
年末现金	21,879	24,048	20,784	49,028	68,797

业绩指标

会计年度 (倍)	2022	2023	2024E	2025E	2026E
增长率 (%)					
营业收入	9.78	8.68	13.98	11.76	11.53
毛利润	6.77	11.13	17.03	12.31	11.91
营业利润	(4.92)	12.63	31.99	15.59	14.39
净利润	(21.12)	23.05	28.32	16.85	16.27
EPS	(19.33)	27.29	32.40	16.85	16.27
盈利能力比率 (%)					
毛利润率	55.38	56.63	58.14	58.43	58.63
EBITDA	29.34	30.61	34.68	35.94	36.91
净利润率	21.20	24.01	27.03	28.26	29.46
ROE	23.62	27.36	31.26	32.10	33.18
ROA	16.55	19.23	22.45	23.38	24.41
偿债能力 (倍)					
净负债比率 (%)	2.08	0.59	1.54	(6.27)	(10.42)
流动比率	2.38	2.10	2.26	2.16	2.48
速动比率	2.34	2.10	2.20	2.15	2.43
营运能力 (天)					
总资产周转率 (次)	0.78	0.80	0.83	0.83	0.83
应收账款周转天数	184.65	156.33	139.71	130.32	122.06
应付账款周转天数	180.22	193.15	186.79	186.79	186.79
存货周转天数	5.48	3.60	5.26	5.26	5.26
现金转换周期	9.91	(33.21)	(41.82)	(51.22)	(59.47)
每股指标 (美元)					
EPS	4.59	5.84	7.74	9.04	10.51
每股净资产	19.60	22.44	26.34	29.98	33.37

资料来源：公司公告、华泰研究预测

正文目录

图表目录

投资要点

我们预计 FY24/25/26 谷歌总营收分别为 3504/3916/4367 亿美元，同比为 14.0%/11.8%/11.5%。我们预计谷歌 Services 业务 FY24/25/26 营收为 3046/3332/3632 亿美元，对应同比为 11.8%/9.4%/9.0%，主要由行业渗透率和 YouTube 增长驱动。谷歌 Cloud 业务 FY24/25/26 营收为 434/553/694 亿美元，对应同比为 31.2%/27.4%/25.5%，主要由自研芯片 TPU、AI 大模型 Gemini 与全链条云端布局驱动。

AI 全栈技术壁垒逐步凸显，有望在 AI 应用下半场回到主导地位

谷歌具备自研 AI 芯片能力，硬件迭代升级支持模型端研发，生成式 AI 浪潮有望驱动谷歌 AI 全栈技术壁垒逐步凸显。我们认为虽然 AI 领域竞争持续激烈，但谷歌通过“硬件+模型+应用”三线布局，叠加云业务快速追赶，下游 B 端及 C 端客户分布广泛，未来 AI 有望带动营收及盈利增长。硬件层面，谷歌于 2016 年自研基于 ASIC 的 TPU 系列 AI 加速器打造成本优势，其 TPU v5p 能实现媲美英伟达 H100 的训练性能，最新 Trillium TPU（v6）在时钟速度、HBM、芯片间互连带宽和能效上对比 v5e 亦有较大提升，赋能大模型训练与云计算差异化服务。对比 AWS 和微软分别在 2018 年和 2023 年才发布第一款自研 AI 芯片。模型层面，谷歌通过自研 PaLM 系列大语言模型、Gemini 系列多模态模型等追赶 OpenAI，且已对标 GPT-4，对比 AWS 主要针对 B 端用户提供 Amazon Bedrock 模型库，自研 Titan 大模型也主要应用于广告场景，谷歌模型布局更为全面。此外，近期谷歌宣布推出量子芯片 Willow，可以在大约 5 分钟内，完成世界上最快的超级计算机估计需要 10^{25}年才能完成的计算任务。Willow 拥有 105 个量子比特，这使得它在量子纠错和随机电路采样（RCS）方面达到同类性能领先。Willow 也是首个在增加量子比特数量的同时能降低错误率的量子系统，**我们认为这表明谷歌技术研发实力持续为第一梯队，随着 AI 应用落地推广，技术壁垒有望逐步被市场认知并兑现业绩。**

广告业务高搜索渗透率或能维持，主要关注 AI 协同进展

搜索广告作为谷歌的基本盘，将继续带来稳定且持续增长的收入。我们预计谷歌 Services 业务 FY24/25/26 营收为 3046/3332/3632 亿美元，对应同比为 11.8%/9.4%/9.0%。谷歌搜索引擎在全球市场上占有绝大部分的搜索流量。我们认为，谷歌虽将继续享受高渗透率所带来的收益，**不过，我们同样认为需密切观察公司在 AI 与广告业务结合层面的改善进度。**

整体来说，从投放渠道看，全球广告线上渗透率仍有较大上升空间。从客户需求看，中国跨境电商和游戏等行业出海，将更加重视线上获客，催化全球在线广告增长。从市场竞争格局看，零售电商和长短视频的市占率虽在较快扩张，但谷歌和 Meta 仍占主导地位。近年来 TikTok 成为广告增速最快的应用，但或同时面临监管风险。另外，**YouTube 作为全球最大的视频平台之一，正在推动谷歌广告业务的转型。**YouTube 广告库存充裕，活跃用户群体众多和视频内容丰富，加上短视频变现率改善，Shorts 商业价值有较大增长空间。YouTube 视频浏览依赖频道订阅，平台积累大量用户兴趣数据，垂直类目广告主可借此高效锁定高潜用户，发挥广告与原生内容协同效应，我们认为将成为谷歌广告业务的增长极。

另外，一些针对垂直领域的 AI 应用小盘股，如 AppLovin 正在以“破局式科技”的方式崛起，但我们认为这恰恰证明 AI 大模型可以适配垂直广告领域，同样为谷歌的机会所在。虽然 AppLovin 目前的市场份额相对谷歌仍然较小，但他们通过在垂直广告领域如游戏和电商方面的专业数据积累，以生成式 AI 赋能实现了转化率的大幅提升，从而吸引价格敏感的中小企业广告商，或展现出削弱谷歌中小企业广告市场的趋势。其所代表的程序化广告，不断搜集并分析用户行为数据，包括用户在应用中的停留时间和每用户平均收入(ARPU)，以此来提升广告投放的精准度和效果。程序化广告作为一种高效连接长尾媒体与目标受众的方法，正在成为广告行业发展的趋势。

目前，谷歌在全球搜索引擎的市场份额已从 23 年 11 月的 91.5%小幅下降至 24 年 11 月的 90.0%。新兴的 Perplexity 等搜索引擎已将大语言模型（聊天机器人）与搜索功能结合。如何将 AI 有效融入到核心产品中以防止份额继续被蚕食，已成为其最大挑战之一。**谷歌拥有强大的 AI 研发实力和海量数据优势。我们认为，谷歌的滞后并非源于技术能力不足，而是商业化节奏和产品创新速度仍有待提升。**

近期谷歌在通过 AI 技术提升搜索体验方面已有明显进展，推出 Gemini Deep Search，目前只面向 Gemini 订阅用户，是谷歌结合大模型对搜索所做的更颠覆商业模式的尝试。但我们认为基于目前的订阅模式，再打磨基于 Gemini Deep Search 的新广告投放算法，或能成为 AI 搜索时代的谷歌新商业模式。**考虑到谷歌 CEO 曾表示 Gemini 将是谷歌下一个现象级应用入口，我们看好谷歌在打磨新搜索商业模式之后，凭借其 AI 技术积淀和庞大的应用生态，在 AI 时代继续保持搜索王者地位。**

公司于 24 年 10 月提拔前搜索广告主管 Prabhakar Raghavan 出任 CTO，其背景主要为信息检索，目前 CEO 也为产品背景，若后续观察到 DeepMind 管理团队融入搜索等核心生态，我们认为或可视为谷歌释放更积极的整合信号。

云业务份额持续提升，AI 全链条布局与头部差距缩小

全球云计算市场规模增长势头稳定，谷歌云业务具备 IaaS-PaaS-SaaS 全链条布局，并将 AI 服务引入云中，持续缩小与头部云厂商的差距。我们预计谷歌云业务 FY24/25/26 营收为 434/553/694 亿美元，对应同比为 31.2%/27.4%/25.5%。谷歌云市占率从 19Q4 的 6%攀升到 24Q3 的 12%。我们认为，目前谷歌云已将市场策略从构建技术转为产品和解决方案，叠加 AIaaS，持续赢得中大型企业青睐，包括最近获得 Airtel、Eiffage Partners、Humana、摩托罗拉等客户。从行业规模上看，目前云计算仍处于发展潮头。据 Gartner 数据，2023 年全球云计算市场规模为 5636 亿美元，预计 2024 年将达 6788 亿美元，同比增长 20.4%。其中 IaaS 同比增长达 26.6%，而谷歌云的 IaaS 市占率更是前五大提供商中增速最高。从竞争地位来看，谷歌云服务现已涵盖了 IaaS、PaaS 和 SaaS 三大层面，24Q3 谷歌云营收增长率为 35%，对比 AWS 和微软的 19%和 34%，持续巩固其全球第三大云地位。谷歌将 Gemini 模型集成为用户端和企业端统一的 AI 品牌，提升工作效率和用户品牌感知，有望进一步追赶微软 Co-pilot，继续提升云业务市场份额。公司表示通过 Gemini for Google Workspace，生产力得到了大幅提升，在客户中平均每个员工每周可节省 105 分钟。

与市场观点不同之处：我们认为谷歌被市场低估，抢回 AI 主导权正当时

市场此前认为谷歌相对其他科技巨头在 AI 领域有所落后，导致谷歌估值相对较低。近期谷歌股价因量子芯片技术突破和 Gemini 2.0 大模型的发布而出现显著上涨，我们认为这波行情背后或反映了市场对谷歌科技能力认知的重大转变，而导致的“价值发现”配置。与 Meta、微软、亚马逊、苹果对比来看，过去五年中谷歌和 Meta 的估值相对较低，而从 2022 年底开始，我们认为谷歌股价在这波 OpenAI 的 chatGPT 横空出现以来的 AI 行情中并没有跟随，2024 年之后落后于 Meta 成为科技巨头中估值最低，主要鉴于市场普遍认为谷歌的 AI 发展已掉队，尤其在与微软、OpenAI 的竞争中显得反应迟缓。然而，量子计算领域的突破性进展，以及 Gemini 的持续迭代，有力证明了谷歌在前沿科技领域的持续创新能力，我们认为应重新审视谷歌的技术实力与投资价值。

我们强调，谷歌在 AI 研究领域的技术实力根深蒂固。谷歌早在 2016 年惊艳发布基于深度学习的 AlphaGo，并以超出人类常见的步法战胜世界围棋冠军李世石。而 AlphaGo 中也使用了自研 AI 芯片 TPU。此芯片的发明也代表着谷歌当年已洞悉降低 AI 计算 TCO 以及软硬件匹配的重要性。TPU 经历多次迭代，对比其他科技巨头在 AI 芯片研发上具备先发优势。谷歌也在 2017 年发布大模型奠基算法 Transformer，随后在 2018 年发布蛋白结构预测系统 AlphaFold，发明者在 2024 年荣获诺贝尔化学奖。另外，有“AI 教父”之称的图灵奖得主、Google Brain 前员工 Geoff Hinton 也于同年获得诺贝尔物理学奖。**我们认为，凭借 TPU 和 Gemini 2 新大模型，以及庞大的搜索生态数据，叠加全链条云布局，谷歌抢回 AI 主导权正当时。**

谷歌：搜索王者站在 AI 的十字路口

谷歌是全球领先的科技公司，三大业务板块包括：1）谷歌服务，包括 Google Search、Android、YouTube、Google Drive、Gmail、Google Play 和硬件等，主要从广告解决方案、应用程序及硬件销售和 YouTube 订阅中获得收入。2023 年谷歌广告总收入为 2379 亿美元，占全球数字广告市场份额的 28%；2）谷歌云：主要从 IaaS/SaaS/PaaS 等服务中产生收入；3）Other Bets：主要包括公司对 Waymo、Verily 和 DeepMind 等其他公司的投资。

图表1：谷歌产品矩阵图

资料来源：谷歌官网、华泰研究

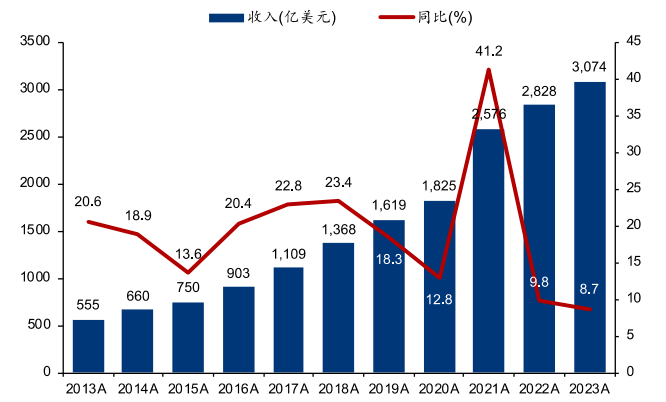
搜索保持谷歌营收的主力地位，云业务成为近年的增长引擎。从 2013 的 555 亿美元到 2023 年的 3,074 亿美元，谷歌总营收的十年 CAGR 为 18.7%。从收入结构来看，广告（尤其是搜索部分）一直以来都为谷歌贡献了大部分的收入，23 年谷歌广告占总收入的 77.4%，其中搜索部分占 56.9%。云业务作为近年谷歌收入增长的主要驱动力，营收从 2018 年的 58 亿美元到 2023 年的 331 亿美元，CAGR 为 41.5%，23 年占总收入的 10.8%。

图表2：谷歌业务分类及收入结构

业务	主要产品	商业模式	2023 全年收入(亿美元)	2023 收入占比
Google Services	Google search	Performance 广告：用户通过点击直接与广告主链接	1750.3	56.9%
	YouTube	Brand 广告：用于品牌建设和营销	315.1	10.3%
	Google Network		313.1	10.2%
	YouTube (非广告部分)	订阅服务，如 YouTube TV 和音乐的会员	346.9	11.3%
	Google Play	应用销售和应用内购买		
	Hardware	Pixel 系列设备的销售		
Google Cloud	Google Cloud Platform	基础设施或平台的消费制和订阅制收入	330.9	10.8%
	Google Workspace	协作工具的消费制和订阅制收入		
Other Bets	Waymo	开发和提供自动驾驶汽车服务	15.3	0.5%
	Verily	数据驱动的医疗技术公司		
	DeepMind	专注于人工智能研究和应用		

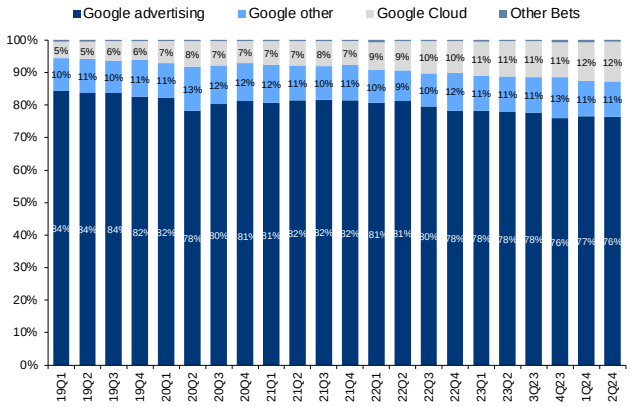
资料来源：谷歌官网、华泰研究

图表3：谷歌总营收及同比增长



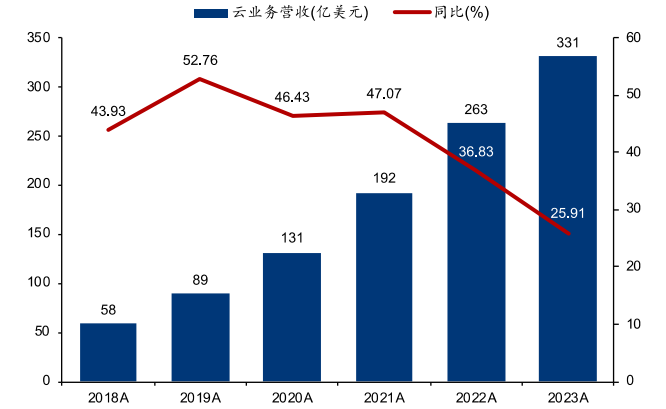
资料来源：谷歌官网、华泰研究

图表4：谷歌收入结构及变化



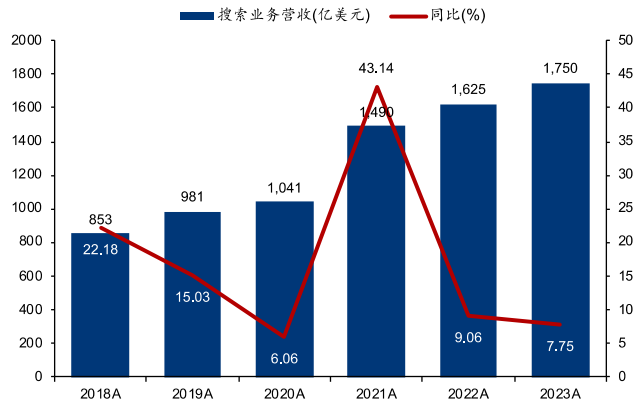
资料来源：谷歌官网、华泰研究

图表5：谷歌云业务营收及同比增长



资料来源：谷歌官网、华泰研究

图表6：谷歌搜索业务营收及同比增长



资料来源：谷歌官网、华泰研究

股价复盘：穿越三大科技时代

我们认为谷歌是互联网时代最具代表性的企业之一，通过复盘谷歌的股价表现，我们可深入探讨公司在 PC 时代、移动互联网时代、云计算时代及人工智能时代的战略决策和市场表现。

1） PC 和移动互联网时代：搜索广告的崛起

谷歌成立于 1998 年，正值 PC 互联网时代的快速发展期。凭借着 PageRank 算法，谷歌在搜索引擎市场中脱颖而出，并通过 AdWords 广告系统实现了搜索广告的商业化。从 2004 年谷歌上市到 2008 年全球金融危机爆发前，公司的股价一路走高。这一时期，谷歌的搜索广告业务几乎垄断了市场，收入和利润的稳健增长是股价持续上扬的主要驱动力。然而，2008 年金融危机的爆发导致全球广告支出大幅下滑，谷歌的股价也随之受到冲击。这一阶段的股价波动清晰地显示出，搜索广告作为公司的基本盘，一旦宏观经济层面出现广告支出下滑，谷歌的股价就会受到明显影响。

谷歌在 PC 时代的成功让其成为全球最具影响力的科技公司之一，但公司管理层察觉到，移动互联网时代的来临将颠覆现有的技术格局。因此，谷歌早在 2005 年就收购了 Android 操作系统，这一前瞻性的决策为公司在移动互联网时代的成功奠定了基础。

随着智能手机的普及和 Android 系统的广泛应用，谷歌成功将 PC 互联网时代的搜索广告模式移植到移动端，进一步扩大了市场份额和收入来源。在这一时期，谷歌的股价再度迎来快速增长。特别是 2011 年后，随着 Android 设备销量的激增和移动广告收入的持续增长，谷歌的股价表现出强劲的上涨势头。然而，谷歌在移动互联网时代的战略布局尽管前瞻，但在硬件产品层面相对谨慎。这种谨慎使得谷歌在与苹果等对手的竞争中保持了相对稳定的市场地位，但也错失了一些更激进的增长机会。比如早期手机无自有品牌主要与摩托罗拉合作，后第一代 Pixel 又仅由 Verizon 独家销售，错过了智能手机品牌的心智建立时期，Pixel 始终难以追赶苹果或三星的地位。

图表7： 谷歌 2004-2014 股价复盘（美元）

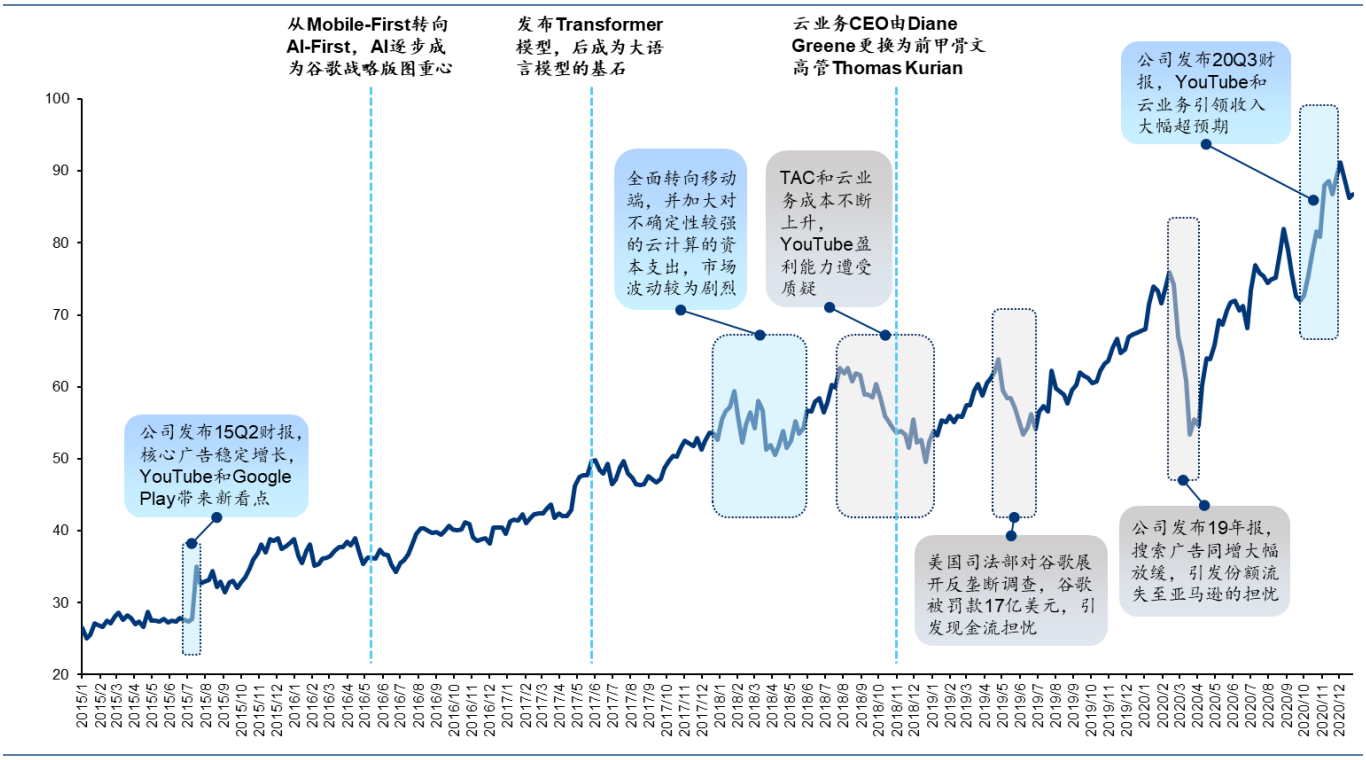
资料来源：谷歌官网、华泰研究

2） 云计算时代：持续扩张与股价波动

2015 年之后，云计算逐渐成为科技行业的主要趋势之一。谷歌管理层再次展现出对行业发展的前瞻性，通过不断加大对 Google Cloud 的投资，谷歌在云计算领域逐渐建立起自己的竞争优势。但由于市场竞争激烈且公司在这一领域的市场份额较低，谷歌的股价在这一阶段表现出较大的波动性。尤其是在 2018 年左右，谷歌云业务换帅，由结构化重组并将云建立为单独业务的前 VMWare CEO Diane Greene 更换为前甲骨文高管 Thomas Kurian，云战略方针经历转型阵痛，让市场对谷歌云业务增长速度有所质疑，导致公司股价在短期内出现了一定的回调。

然而，谷歌云计算业务的收入增长速度仍然可观，尤其是在 2020 年全球疫情的推动下，远程办公以及无接触 AI 需求激增，使得云计算业务成为谷歌新的增长点。2021 年，公司云计算业务的强劲表现再次推动股价走高。

图表8：谷歌 2015-2020 股价复盘（美元）



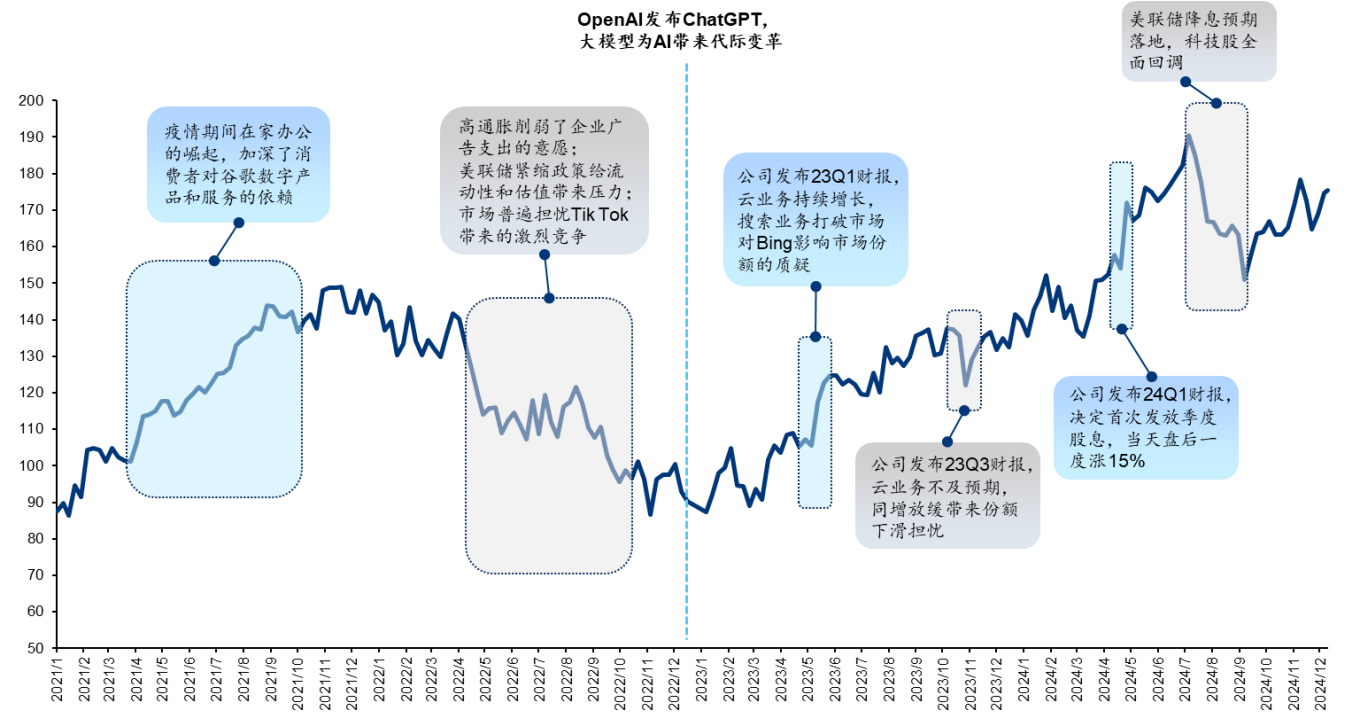
资料来源：谷歌官网、华泰研究

3） 人工智能时代：AI-First 战略的提前布局

进入 2020 年代，人工智能成为科技行业的核心驱动力之一。谷歌早在 2014 年就收购了 DeepMind，并在 2017 年发布大模型的基石算法 Transformer 和正式提出了“AI-First”的战略定位。这一战略使得谷歌在人工智能领域的布局比行业发展早了数年。

2022 年，随着 ChatGPT 等生成式 AI 产品的问世，该类人工智能开始对各行各业产生深远影响。虽然市场普遍认为谷歌在生成式 AI 的商业化落地上持谨慎态度，但我们认为，谷歌乃是科技大厂中少有的清楚认知到现阶段大模型能力边界的公司。市场也普遍认为，ChatGPT 和类似产品乃“有求必应”，事无大小都能解决，但实际上模型的幻觉与搜索结果的精确性要求存在冲突，且在短期内难以有完善的解决方案，这也是目前以大模型为基础的搜索无法迅速占领市场的原因之一。然而，Perplexity 等后起之秀已把聊天机器人（大模型）和搜索有机结合，我们认为这也是谷歌需追赶的重要一环。不过，在某些垂直领域，比如说在医疗行业的创新药研发，我们则认为大模型及生成式 AI 具备充足落地条件，鉴于该类问题可通过物理和化学定律有效控制和监管幻觉的出现，从而受其影响较少，因此能商业性落地，如谷歌 AlphaFold 通过生成式 AI 模拟及预测蛋白结构、DNA、RNA 和其他分子如何相互作用，能缩短药物研发时间，从而起到关键作用。谷歌在这一领域的领导地位和前瞻性认知或将为公司股价提供长期支撑。

图表9：谷歌 2021-至今股价复盘（美元）



资料来源：谷歌官网、华泰研究

总的来说，谷歌的股价复盘说明了以下两点：

1） **管理层始终展现出对行业发展趋势的高度敏锐性，且多次在技术变革的早期阶段做出了战略性抉择，但在执行层面相较竞争者更谨慎。**早在 2005 年，谷歌通过收购 Android，提前布局了移动互联网。当时，智能手机尚未成为市场的主流产品，而苹果 iPhone 也未正式发布，但谷歌管理层已预见到移动互联网将是未来的发展方向。2008 年，Android 系统正式推出，随着智能手机的普及，Android 迅速成为全球最广泛使用的移动操作系统之一。2006 年，谷歌以 16.5 亿美元的价格收购了当时尚处于早期发展阶段的 YouTube。尽管当时的估值看似高昂，但这一决策展现了谷歌对未来互联网视频消费趋势的敏锐洞察。随着宽带普及和用户对视频内容需求的增长，YouTube 逐渐发展成为全球最大的在线视频平台，成为谷歌广告生态系统中的重要组成部分。2014 年，谷歌收购了以 Demis Hassabis 挂帅的英国人工智能公司 DeepMind。当时，人工智能技术尚未广泛被应用，但谷歌管理层通过这一收购，显示了其对 AI 时代的前瞻性布局。后来的事实证明，AI 不仅是谷歌核心业务的关键推动力，也是全球技术创新的焦点。通过自研 AI 芯片 TPU、击败世界围棋冠军的 AlphaGo，以及大语言模型（如 PaLM、Gemini 系列），加上几位重要员工，Geoff Hinton、Demis Hassabis 等，均为本届诺贝尔奖得主，足以证明谷歌在 AI 时代的竞争中具备强劲优势、以及在资本市场上的具备长期竞争力。

2） **搜索广告始终是谷歌的基本盘，一旦宏观层面广告支出有下滑趋势，对公司股价有较大影响。**Google Ads 作为全球最大的在线广告平台，长期占据着全球互联网广告市场的领导地位。因此，宏观经济环境对广告支出的变化，往往对谷歌的股价有直接影响。在 2008 年次贷危机中，全球经济陷入衰退，企业广告支出急剧减少。作为依赖广告收入的公司，谷歌股价在这一时期受到明显冲击。股价从 2007 年底的高点急速下滑，直至 2009 年初才开始缓慢回升。2020 年新冠疫情爆发，全球经济陷入停滞，广告支出在短期内大幅下降。然而，随着疫情带动的数字化加速转型，尤其是电商、在线教育和视频会议等领域的需求激增，谷歌广告业务迅速复苏并实现了增长。2021 年，谷歌的股价也随之创下历史新高，展示了其核心业务的强大韧性。2022 年，全球通胀高企，企业广告支出增速放缓，再次对谷歌的广告收入构成挑战。尽管如此，谷歌依靠多元化的收入来源，特别是云计算，逐渐抵消了广告市场的下滑影响。股价虽然短期内出现波动，但在整体大盘中依然表现优异。

AI 生态站位：下一个科技周期谷歌该如何转身？

硬件层 TPU 持续发力，赋能云端和大模型发展

AI 芯片领域已多方入局，自研芯片乃科技厂商必经之路。目前科技巨头出于削减 TCO、提升研发可控性及满足软硬件生态等考量，均针对内存容量与延时两大 AI 计算瓶颈自研 AI 芯片，并通过与 AI 初创公司合作率先绑定用户，试图逐步摆脱对外部硬件的依赖。我们认为或将成为英伟达的最大竞争对手。

谷歌作为云大厂中自研芯片的先行者和 AI 领域的奠基者之一，其自研的 TPU 是专为神经网络设计的 ASIC 芯片，具有矩阵乘法单元（MXU）和专有的互连拓扑等专用功能。TPU 针对矩阵乘法优化，且具有跨芯片并行功能，适合训练大模型，且具备成本优势。2023 年 4 月，谷歌还把开发 TPU 的工团队转移到谷歌云，以提高其向云客户出租 TPU 服务器的能力。未来 Trillium TPU 有望在 AI 芯片激烈的竞争格局中脱颖而出。

谷歌 TPU 对比其他科技厂商拥有先发优势，主要在以下两个方面：

1） **专为神经网络的张量运算而设计，高度匹配自身生态。**TPU 专为神经网络设计，具有矩阵乘法单元（MXU）、专有互连拓扑、多重切片等功能来优化模型训练端。其不仅适合训练大模型，且对比通用 GPU 具备成本（TCO）优势。此外，谷歌不仅已通过 TPU 实现包括 Gemini 在内的自身模型训练，其专门为谷歌开源深度学习框架而定制，与 TensorFlow、PyTorch、JAX 等机器学习框架集成在一起，且已商业化落地，能为云端客户提供自研模型的硬件支持。对比 Meta 目前的 MTIA 系列芯片仅用于广告推荐、信息流等业务，在模型训练领域性能仍较为不足；AWS 在硬件支持自身生态系统方面稍逊一筹，其自研芯片主要通过实例来供客户使用；微软旗下 Maia AI 芯片则推出较晚，目前仅优先为自身云服务提供支持。

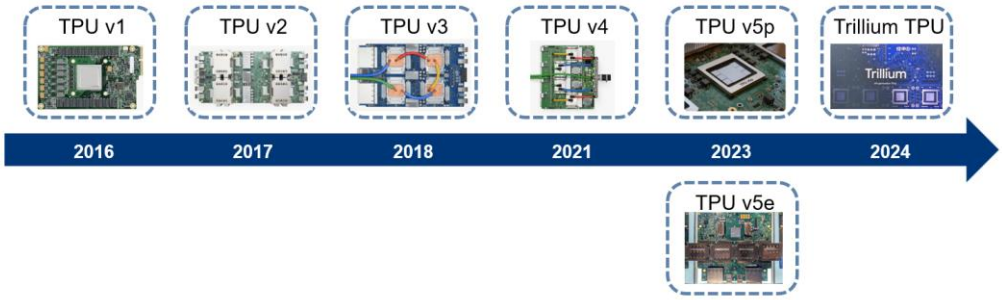
2） **研发经验丰富，已经历多次迭代。**TPU 自 2016 年开始推出，八年间已迭代至第六代，对比 AWS 及 Meta 分别在 18/23 年才陆续布局，而微软早年虽有布局但因选用 FPGA 而导致裹足不前。谷歌芯片研发经验优于同业，且具备先发优势。此外，AWS、微软及 Meta 在模型训练上也仍需英伟达等外部厂商的支持，其中微软计划将在 2024 年直接购买超过 40 万个 GPU，用于训练端和 Copilot /API 推理端；而 Meta 也声称在生成式 AI 与模型端，公司仍以采购英伟达芯片为主。对比下谷歌已成功基于 TPU 完成对 Gemini、Gemma 等模型的训练。

图表10： 海外云大厂和互联网巨头的自研芯片

	云厂商	芯片	（预计）发布时间	代际	制程	设计	种类	功能	特点
AI 芯片	amazon	Trainium	2023	2	5 nm	Marvell	ASIC	Training	适用于大模型训练
	amazon	Trainium	2025	3	3 nm	TBD	ASIC	Training	适用于大模型训练
	amazon	Inferentia	2019	2	5 nm	TBD	ASIC	Inference	主要为高性能深度学习推理应用程序而设计
	Google	TPU	2024	6	3 nm	Broadcom	ASIC	Training and Inference	少数能与英伟达高算力 GPU 匹敌的 AI 芯片
	Meta	MTIA	2024	2	5 nm	Broadcom	ASIC	Training and Inference	主要用于广告推荐、信息流等业务
	Microsoft	Maia 100	2023	1	5 nm	In-house	ASIC	Training and Inference	专为处理大型 AI 和生成式 AI 工作负载设计
CPU	amazon	Graviton	2023	4	5 nm	In-house	ArmCPU	-	支持广泛的云上工作负载
	Google	Axion	2024	1	5 nm	Marvell	ArmCPU	-	用于执行云端通用运算负载
	Microsoft	Cobalt 100	2023	1	5 nm	In-house	ArmCPU	-	用于在微软 Cloud 上运行通用计算工作负载

注：根据 The next platform 报道，Axion 或为基于 Crypress 的设计
资料来源：The Information 官网、各公司官网、华泰研究

图表11：谷歌发布 TPU 历程图



资料来源：谷歌官网、华泰研究

谷歌在硬件方面深耕已久，TPU 在架构与性能参数上不断迭代。第一代 TPU 于 2016 年谷歌 I/O 大会上正式发布，主要为谷歌云计算数据中心的机器学习应用提供，彼时仅面向推理端，但从 2017 年推出第二代开始，TPU 已同时拥有训练和推理能力，不仅能支持浮点数运算，且具有更高的片上内存。2018 年发布的 TPU v3 旨在提高性能和能效以满足不断增长的机器学习任务需求，但其应用范围仍然受限于谷歌的生态系统和软件包。第四代 TPU 于 2021 年发布，主要突破在于部署可重构光电路交换机（Optical Circuit Switch，OCS）来快速动态重新配置芯片之间的连接，有助于在出现故障时实时调整。专为中大规模训练和推理而构建的 TPU v5e 于 2023 年发布。与 TPU v4 相比，TPU v5e 可为大语言模型提供高达 2 倍的训练性能和 2.5 倍的推理性能，并能节约一半以上的成本。Gemini 1.0 是基于 TPU v4 和 TPU v5e 在人工智能优化基础上进行的大规模训练，在 TPU v5p 也会应用于加速 Gemini 开发。

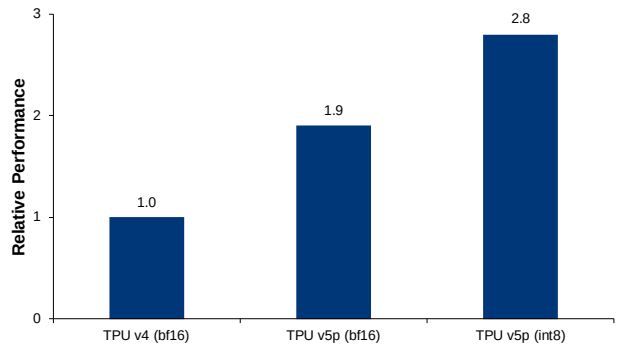
2024 年 5 月 15 日，谷歌于 I/O 大会宣布第六代 Trillium TPU，并于 10 月底正式推出预览版。Trillium TPU 通过改良芯片设计，包括扩大矩阵乘法单元（MXU）并提高时钟速度，以及提升 HBM 和芯片间互连（ICI）带宽至 v5e 的 2 倍，使单芯片峰值算力对比 TPU v5e 提高 4.7 倍，能效也比 v5e 高 67%以上。芯片扩展方面，Trillium 不仅通过 Multislice 技术能在单个 Pod 中扩展至 256 个 TPU，且能通过多切片技术实现集群，从而每秒处理 PB 级数据。目前谷歌 Trillium TPU 已于 24 年 10 月底推出预览版。此外，目前谷歌仅通过谷歌云服务平台向外部客户提供 TPU 的算力租赁服务，而未有将其作为硬件产品出售。

图表12：历代 TPU 性能对比

	TPU v1	TPU v2	TPU v3	TPU v4	TPU v5e	TPU v5p	TPU v6e
发布时间	2016	2017	2018	2021	2023	2023	2024
工艺制程（nm）	28	16	16	7	5	5	3
时钟频率（MHz）	700	700	940	1050	-	1750	-
每颗芯片的峰值计算能力（TFLOPS）	92 (int8)	46 (bf16)	123 (bf16)	275 (bf16 or int8)	197 (bf16)	459 (bf16)	918 (bf16)
HBM 容量与带宽	28 GB, 34 GB/s	16 GB, 700 GB/s	32 GB, 900 GB/s	32 GB, 1200 GB/s	16 GB, 819 GB/s	95 GB, 2765 GB/s	32 GB, 1640 GB/s
最小/平均/最大测量功耗（W）	40	-	123/220/262	90/170/192	-	-	-
TPU Pod 规模（芯片数量）	-	256	1024	4096	256	8960	256
互连拓扑结构	-	2D torus	2D torus	3D torus	2D torus	3D torus	2D torus
每个 Pod 峰值计算能力（PFLOPS）	-	12 (bf16)	126 (bf16)	1100 (bf16 or int8)	51 (bf16)	-	235 (bf16)
每个 Pod 的 All-reduce 带宽（TB/s）	-	120	340	1100	51.2	-	102.4
每个 Pod 的切分带宽（TB/s）	-	2	6.4	24	1.6	-	3.2
目标应用场景	仅推理端	训练&推理端	训练&推理端	训练&推理端	训练&推理端	训练&推理端	训练&推理端

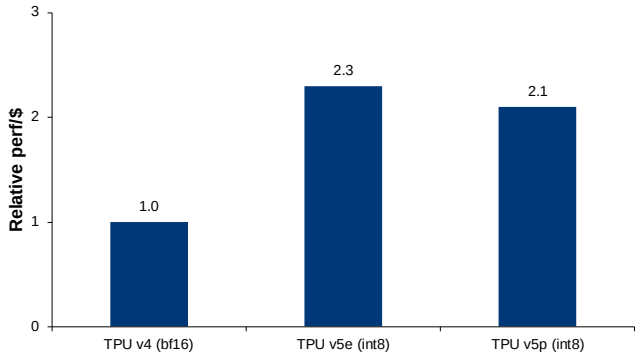
资料来源：谷歌官网、Next Platform 官网、华泰研究

图表13：谷歌历代 TPU 训练大模型速度表现比较



资料来源：谷歌官网、华泰研究

图表14：谷歌历代 TPU 训练大模型每美元相对性能比较



资料来源：谷歌官网、华泰研究

相较于 GPU，TPU 在 AI 领域具有以下优势：1）性能：TPU 专为张量运算而设计，能针对特定 AI 工作负载（训练、微调和推理）进行经济高效的扩缩，因此在特定情况下，神经网络的训练和推理效率或更高。**2）集成性：**TPU 专门为谷歌开源深度学习框架而定制，与 TensorFlow、PyTorch、JAX 等机器学习框架集成在一起，可加速其工作负载，在一同使用下效率或更高。3）**成本：**谷歌云上 TPU 相比 GPU 价格而言，配置 1 个 H100 芯片，内存为 234GB 的 A3 虚拟机价格为 11.06 美元/小时，H100 现货价格为 9.04 美元/芯片/小时，TPU v4/v5e/v5p/v6 的价格分别为 3.22/1.2/4.2/2.7 美元/芯片/小时，TPU 收费存在优势。

不过，TPU 作为 ASIC 存在通用性较弱等问题。此外，TPU 的应用也在一定程度上受到英伟达 CUDA 生态圈一家独大的影响。谷歌云作为 AI 云服务商，需满足有 AI 训练和推理需求的客户，而英伟达 GPU 拥有生态圈成熟和开发者众多的 CUDA，是目前大部分 AI 训练所必需的工具。因此，我们认为 TPU 或其他云大厂自研芯片不会完全取代英伟达的 GPU，二者与英伟达 GPU 应能形成良性互补，并非零和博弈，但若未来科技厂商算法相对成熟，设计 ASIC 去取代部分英伟达的算力也较为合适。此外，谷歌在自研 AI 芯片同时，也大量采购英伟达 GPU，包括 H100 以及 Blackwell 平台，也引入公司 AI 云基础设施和超级计算机架构中，目前已公布基于 Blackwell GB200 NVL 机架的预览照片。对比微软、AWS 与 Meta 亦同时采取自研+外购 AI 芯片齐头并进策略，实现优势互补与降本增效。

图表15：最新几代 TPU 价格对比

TPU 版本	地区	按需定价(USD)	1 年承诺定价(USD)	3 年承诺定价(USD)	*现货定价
Trillium TPU	美国东部-5	$2.7000	$1.8900	$1.2200	$1.8900
TPU v5p	美国东部-5	$4.2000	$2.9400	$1.8900	$2.4150
TPU v5e	美国东部-5	$1.2000	$0.8400	$0.5400	$0.6000
TPU v4 pod	美国中央-2	$3.2200	$2.0286	$1.4490	$3.6860

*现货定价将在一定时期动态调整

资料来源：谷歌云官网、华泰研究

总的来说，我们认为：1）谷歌 TPU 或其他云厂商的自研芯片不会在一夜之间取代所有英伟达的 GPU；2）若算法已相对成熟，可使用 TensorFlow 框架编程并在 TPU 上运行，可有效利用其优化和加速，节省成本，或是性价比较高的选择；3）面对英伟达 CUDA 的成熟生态圈，云厂商自研芯片无需以完全取代作为目标，而仅需为客户提供更多算力选择即可有效打开市场。

微软：自研芯片崭露头角，布局硬件端强化 AI 竞争力

微软研发芯片早已有迹可循，但其在云业务三巨头中布局自研 AI 芯片进度较慢。目前，微软在芯片上采取“自研+合作”双管齐下计划，以支持各种 AI 训练与推理端工作负载。微软不仅推出自研 AI 芯片，同时积极与英伟达、AMD 等头部 AI 芯片厂商保持紧密合作，为自身部署生成式 AI 提供不同选项。

定制 AI 芯片如期而至，以应对大模型训练成本挑战，并降低提供 AI 服务的成本。2023 年 11 月，微软在 2023 Microsoft Ignite 开发者大会上宣布将推出两款定制芯片：1）对人工智能任务和生成式 AI 进行优化的 AI 芯片（ASIC）Maia 100；2）首款基于 Arm 构建的 Cobalt 100 服务器 CPU。微软在 2024 年 8 月进一步公布了 Maia 100 的设计细节，并于 10 月正式推出基于 Cobalt 100 的虚拟机。不过 Maia 将优先为自身 Azure 云服务提供支持，欲降低对英伟达昂贵芯片的依赖。

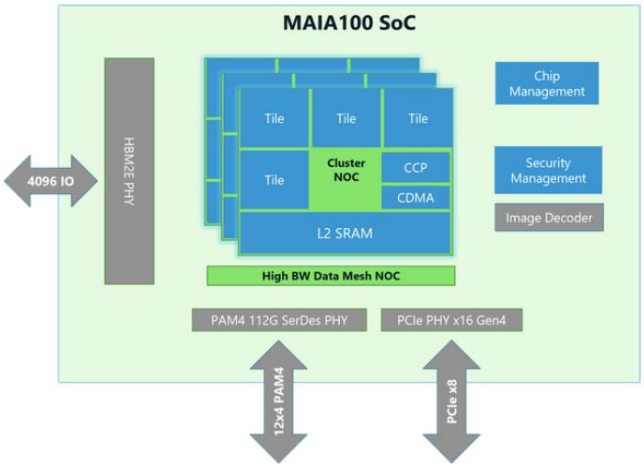
Maia100 和 Cobalt 100 CPU 的共同推出体现了微软多元化的芯片策略和对 AI 计算中不同工作负载的重视。此外，微软的 Cobalt 100 基于 Arm 的 CPU 也应证了我们此前的观点，即 ARM 架构相比 x86 能耗较低，加上目前 CPU 在 CPU+GPU 架构中承担的任务将会减少，因此不管在数据中心或在 AI 应用中，Arm 架构将越来越受到青睐。**Maia 100 芯片专为运行和优化云 AI 工作负载而设计，旨在运行大语言模型、帮助 AI 系统更快地处理大量数据。Cobalt 100 CPU 是微软在云中部署的第二款基于 Arm v9 指令集的 64 位处理器，旨在为微软 Azure 上的通用云服务提供支持，并针对通用工作负载的性能、功率和成本效益进行了优化。**

图表16： 微软 Maia 100 芯片参数

	参数及对比
制程及规格	采用台积电 5nm 制程工艺，拥有 1050 亿个晶体管 -对比英伟达 H100 的 800 亿晶体管多 31% -对比 AMD MI300X 的 1530 亿晶体管少约 30%
算力	首次支持低于 8 位数据类型（Int8），以便共同设计软硬件，且有助于支持更快的模型训练和推理 -在 MXInt8 下的性能为 1600 TFLOPS -在 MXFP4 下运算速度达 3200 TFLOPS
内存和带宽	虽然 Maia 片上放置了大量 SRAM，但其在片外只采用了 64 GB 的 4 层 HBM2E 堆栈，低于英伟达 H100 的 6 层 HBM3 80GB 与 AMD 的 8 层 HBM3 192GB 内存带宽 1.8 TB/s，略低于英伟达 H100NVL 和 AMD MI300X 的 3900GB/s 以及 5300GB/s
网络设计	Maia 采用定制的基于以太网的网络协议设计，让每个芯片都有内置 RDMA 以太网 I/O，以实现更好的扩展和端到端工作负载性能 单个 Maia 的 I/O 总单向带宽为 600GB/s，远超 AMD MI300X 的 896GB/s 和英伟达 H100 的 900 GB/s

资料来源：微软官网、华泰研究

图表17： 微软 Maia 100 芯片架构



资料来源：微软官网、华泰研究

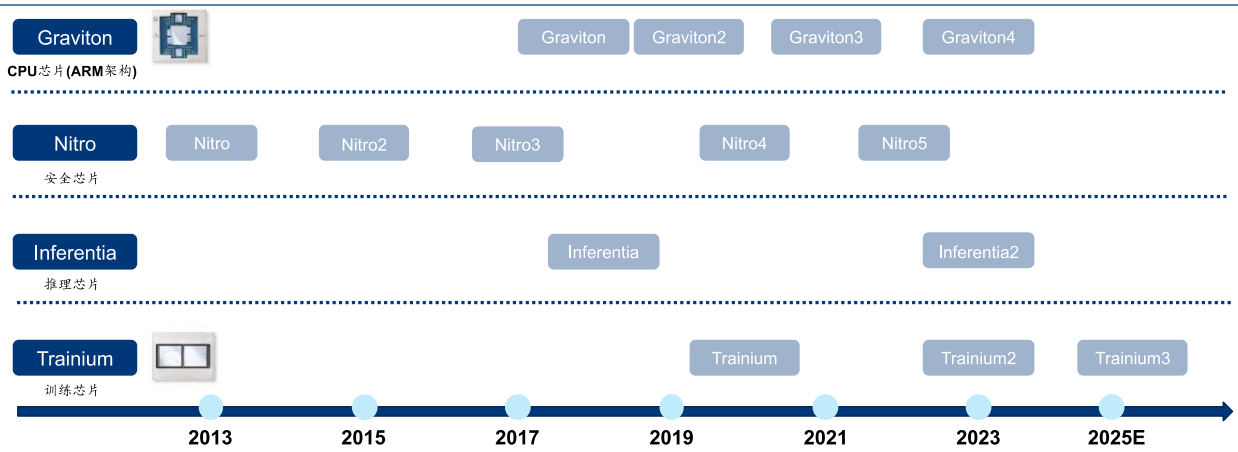
亚马逊：AI 芯片条线清晰，推理+训练端双轮驱动助力 AI 新征程

亚马逊目前拥有 3 条主要的自研芯片产品线，分别为基于 ARM 架构的 GravitonCPU 系列、安全芯片 Nitro 系列、人工智能自研芯片系列（Inferentia 推理芯片和 Trainium 训练芯片）。目前，也有部分客户在使用亚马逊自研 AI 芯片，例如 Anthropic、Airbnb、Snap 和 Hugging Face 等。

亚马逊未来将加大投入，主要取代场景为公司的自有实例。自研芯片能改善硬件供应端的限制，以及成本竞争力等关键问题。我们认为，未来 AWS 不会使用其专有芯片完全取代其实例中提供英伟达、AMD 等的外购芯片解决方案，但由于硬件创新对于成本竞争力至关重要，AWS 预计未来将增加对自研芯片开发的投入。

回溯亚马逊产品发展历史，2013 年亚马逊推出首颗针对底层系统的 Nitro 芯片；2015 年，亚马逊收购了以色列芯片初创公司 Annapurna Labs 作为专注于亚马逊的芯片定制团队，并于 2018 年推出基于 Arm 架构的服务器 CPU 芯片 Graviton。针对人工智能与机器学习关键环节，亚马逊分别在 2018 年与 2020 年首次推出自研 AI 推理芯片 Inferentia 和训练芯片 Trainium。目前亚马逊已决定停止开发 Inferentia，将资源集中转向 Trainium。公司在 2024 年 12 月的 re:Invent 大会上宣布 Trainium 2 芯片已全面投入使用，可用于训练和部署大型语言模型，并计划在 2025 年下半年发布采用 3nm 制程的 Trainium 3。

图表18：AWS 自研芯片推出时间线



资料来源：AWS 官网，华泰研究

图表19：AWS 加速型计算基础设施概述

	Amazon EC2 P5	Amazon EC2 G6e	Amazon EC2 Inf2	Amazon EC2 Trn1	Amazon EC2 DL2q
实例规模	p5.48xlarge	g6e.48xlarge	inf2.48xlarge	trn1n.32xlarge	dl2q.24xlarge
GPU	8 NVIDIA H100	8 NVIDIA L40S GPU	12 AWS Inferentia 2	16 AWS Trainium	8 Qualcomm AI 100
vCPU	192 AMD EPYC 7R13	192 AMD EPYC 7R13	192 AMD EPYC 7R13	128 Ice Lake SP	96 Cascade Lake P-8259CL
GPU 内存	640 GB HBM3	384 GB	384 GB	512 GB	128 GB
实例内存	2 TB	1536 GB	768 GB	512 GB	768 GB
网络带宽	3200 Gbps EFAv2	400 Gbps	100 Gbps	1600 Gbps	100 Gbps
EBS 带宽	80 Gbps	60 Gbps	60 Gbps	80 Gbps	19 Gbps
按需价格/小时	USD 98.32	USD 30.13	USD 12.98127	USD 24.78	-

资料来源：AWS 官网、华泰研究

Inferentia 系列芯片旨在提高推理任务性能的同时降低成本。2023 年 4 月，AWS 发布新一代 Inferentia 2 与 Inf2 实例算力方面，Inferentia 2 针对 INT8 数据类型可提供算力达 380 TFLOPS，针对 FP16/BF16 数据类型算力达 190 TFLOPS，并额外添加了对 FP32、TF32 和可配置的 FP8（cFP8）数据类型的支持。内存方面，Inferentia 2 具备 32GB HBM，带宽为 820 GB/s，对比 Inferentia 总内存增加了 4 倍，内存带宽增加了 10 倍。**针对训练机器学习模型，Trainium 系列训练芯片能优化成本。**在 2023 年 11 月的 re:Invent 大会上，AWS 宣布推出新一代 Trainium 2。Trainium 2 芯片具有 2 个计算核心与 4 个 HBM 内存堆栈。与第一代 Trainium 相比，Trainium 2 的训练性能提高了 4 倍，内存容量提高了 3 倍，同时能效提高了 2 倍。在 2024 年 12 月的 re:Invent 大会上，AWS 宣布 Trainium 2 正式应用于 Amazon EC2 Trn2 实例。

图表20：独角兽公司使用 Trainium 应用情况

AWS 芯片客户	公司简介	使用情况
Anthropic	AWS 投资的生成式 AI 独角兽	自 2021 年开始使用 AWS，正在构建基于 Trainium 2 的 UltraServers EC2 UltraCluster
Databricks	软件公司	利用 Trainium 2 促进 Mosaic AI 的训练，将 TCO 降低高达 30%，并服务于全球超过 10000 家组织机构
Hugging Face	AI 开放平台	Optimum Neuron 开源库集成在 Hugging Face 推理端点中，Trainium2 的推出能使 Hugging Face 用户获得更高的性能，从而更快地开发和部署模型
Helixon	医疗保健研究组织	利用 Trainium1 分析大量基因组数据集，利用其深度学习模型获得准确遗传见解
Money Forward	金融科技公司	更精确地预测市场趋势和客户行为

资料来源：AWS 官网，华泰研究

模型侧多模态和轻量化成为趋势，谷歌布局仍处第一梯队

如今模型应用发展有两大趋势：1）单一文本模型向多模态大模型转变；2）大参数模型向轻量化与端侧发展。谷歌在两方面部署上均处于第一梯队，对比同业模型竞争优势明显。

我们认为，在大模型竞争中，OpenAI 虽凭借早期先发优势与商业化落地率先抢占市场，但目前谷歌凭借自身硬件、数据与生态优势，叠加以 Gemini 为主的“AI 全家桶”重振旗鼓，逐渐抢回 AI 赛道上的主动权。如今 OpenAI 不仅模型迭代的速度开始变慢，且管理层接连出现变动，23 年 11 月 CEO Sam Altman 出现领导力危机，24 年 5 月联合创始人 Ilya Sutskever 离职，9 月 CTO Mira Murati、首席研究官 Bob McGrew 和研究副总裁 Barret Zoph 均宣布离职，显现出公司管理层内部或存在一定问题。对比下谷歌有望在此次 AI 竞赛中逐渐缩小与 OpenAI 的技术与份额差距。

各科技公司针对自身软硬件优势各有侧重：

1） **谷歌凭借自身 TPU 与数据集赋能大模型发展，持续拓宽模型应用范围。**先后推出大语言模型 PaLM、PaLM2，并加强垂直应用布局，依据特定领域的数据进行了模型微调，以执行企业客户的特定任务。多模态大模型 Gemini 在多项基准测试中的优秀表现已成功挑战到 OpenAI 一家独大的地位。此外，谷歌在轻量化大模型、可交互生成式世界模型上亦有部署，旗下 Gemma 系列及 Genie 先后推出以强化自身模型端竞争力。另外，作为 Transformer 的发明者，谷歌也非常明白大模型的幻觉问题，因此，他们并没有将此技术直接应用到搜索里。不过，面对像 Perplexity 等“先搜索、后整理”的正确运用大模型去整理爬虫等搜索结果的后起之秀，我们认为谷歌也有必要急起直追，以巩固其在搜索的霸主地位，防止份额被进一步蚕食。

2） **微软联盟 OpenAI 在大模型领域占据先发优势，多元布局欲打造 AI 模型帝国。**2022 年 11 月 30 日，基于 GPT-3.5 的 ChatGPT 正式发布，开创了人工智能新纪元。微软则通过与 OpenAI 相互赋能，成为一时无两的大模型领域领跑者。但近期 OpenAI 的各种问题，包括多位高管和创始人的离职，加上竞争格局愈发激烈，以及 Scaling Law 的发展开始众说纷纭，并涉及到通用性(Artificial Generative AI)和垂直领域(domain expertise)应用的争议等，也将影响微软在 AI 应用的落地。截至 24 年 10 月，微软已向 OpenAI 投资共计 137.5 亿美元，并拥有 GPT-4o 和包括 DALL·E、Embedding、Whisper 等在内的所有其他 OpenAI 人工智能模型的独家授权，让其 Azure OpenAI 服务旗下模型种类众多且能商业化率先落地。而微软自身也针对轻量化模型领域于 2024 年 4 月发布开源模型 Phi-3，包含单语言模型的 mini、small 和 medium 版本和多模态的 vision 版本，助力边缘智能终端部署，但先发优势能否维持则有待观察。

3） **亚马逊也具备较完整的硬件端产品布局，专注 B 端客户布局中间层服务。**亚马逊主要聚焦 B 端客户需求，其借助自研训练芯片 Trainium 与推理芯片 Inferentia 具备构筑大模型的成本优势，拥有自研大语言模型 Amazon Titan，亦为 B 端提供 Amazon Bedrock 服务平台，让用户能访问来自 Anthropic、Cohere、Meta with Llama2 和 Stability AI 等丰富的 AI 模型库。

4） **Meta 深耕开源大语言模型 LlaMa，多模态与轻量级模型齐发力。**Meta 深耕轻量化模型领域，旗下 LlaMa 能以较小的参数量媲美主流大模型性能，并通过开源免费提供给研究者和商业使用者。2024 年 4 月，Meta 发布 Llama 3，12 月更新到 Llama 3.3 版本，使性能与效率双升，并进一步节省算力。此外，Meta 不仅通过 ImageBind 实现跨多模态创建联合嵌入空间技术，还构建了针对翻译、语音、图像及视频的多款轻量化模型。

5） **字节跳动豆包大模型崭露头角，快速工厂式复制 AI 应用。**豆包大模型具备多模态处理能力，日均 tokens 使用量超过 4 万亿，相比发布时增长了 33 倍。至 2024 年 11 月，豆包 APP 的月活跃用户数（MAU）已达到约 5998 万。我们认为，字节跳动具备爆款应用的生产能力，有望在 AI 垂类领域复制，但后续盈利能力仍待观察。

6） 百度飞桨+文心大模型打造中间技术层，赋能模型应用蓬勃发展。文心大模型 ERNIE 涵盖 NLP、视觉、跨模态、生物计算和行业模型五大领域。目前，百度 AI 架构已具备垂直一体式布局，下游行业覆盖面广泛，已有多家合作伙伴接入文心生态。

图表21： 各科技巨头大模型版本与参数对比

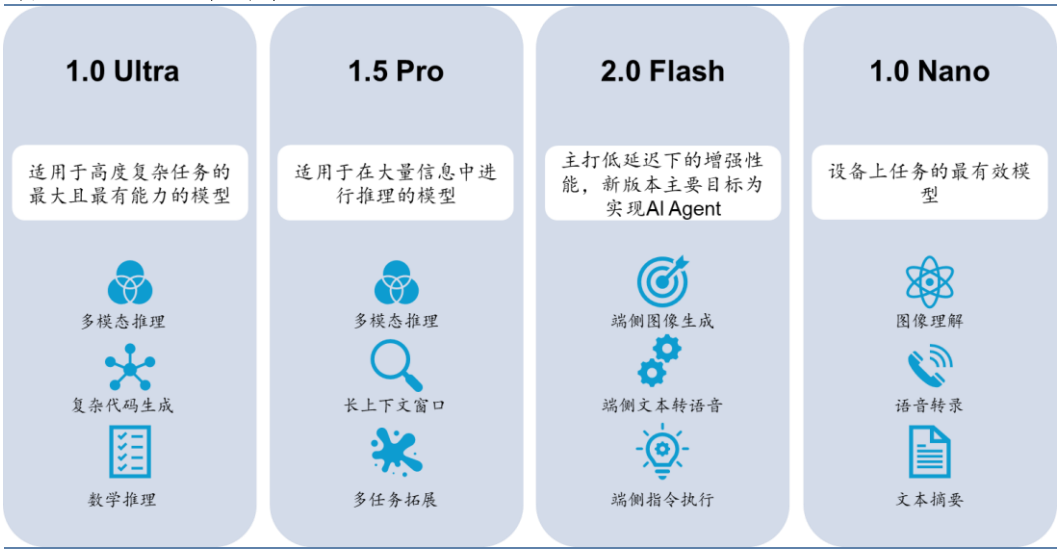
	大模型名称	发布时间	参数量（亿）	训练芯片	预训练令牌数	输入类型	应用功能
谷歌	PaLM	2022.4	8/620/5400	TPU v4	780 Billion	文字	大语言模型，能够执行常识推理、算术推理、文本解释、代码生成和翻译等任务
	PaLM2	2023.5	3400	TPU v4	3600 Billion	文字	具有改进的多语言、推理和编码功能
	Gemini	2023.12	18/32.5/16000	TPU v5e TPU v5p	-	文字、图像、音频、视频	原生多模态模型，能够处理文本、图像、音频、视频理解、代码编写和数学推理
	Gemini 1.5 Pro	2024.2	-	TPU v5p	-	文字、图像、音频、视频	中型多模态模型，性能与 1.0 Ultra 相当，能够实现 200 万个 tokens 的长上下文窗口
	Gemini 1.5 Flash	2024.5	-	TPU v5p	-	文字、图像、音频、视频	轻量化多模态模型，是 API 中速度最快的 Gemini 模型，旨在快速高效地大规模提供服务
	Gemma	2024.2	20/70	TPU v5e	2000/6000 Billion	文字	轻量级大语言模型，能够处理数学推理、编码、文本对话于推理任务等
	Gemma 2	2024.5	27	-	-	文字	采用全新架构，性能提升、降低部署成本并拥有多功能调优工具链
	Genie	2024.2	107	TPUv5p	9420	文字、图像、	通过互联网视频训练的基础世界模型，可以从合成图像、照片甚至草图生成可玩世界
微软	GPT-3	2020.6	1750	A100	-	文字	可以理解并生成自然语言（文章、诗歌、故事、新闻报道和对话）或编辑代码
	GPT-4.0 Turbo	2023.11	17600	A100/H100	-	文字、图像 音频	大型多模态模型，具备自然语言处理、语言翻译、代码生成、图像处理、推理等功能
	GPT-4o	2024.4	-	H100	-	文字、图像 音频、视频	OpenAI 最先进的多模态模型，具有与 GPT-4 Turbo 相同性能，但效率更高且成本更低
	DALL • E 3	2023.10	-	-	-	文字、图像	能够根据文本创建原创图像和艺术；将图像扩展到原始画布之外；根据文字对图像进行编辑
	Whisper v3	2023.11	15.5	-	-	音频	自动语音识别系统，可以进行多种语言的转录，以及语言翻译成英语
	Embeddings	2024.1	-	-	-	文字	文本向量化模型，可以计算文本字符串的特征向量，通过向量来衡量字符串之间的语义相关性，从而为真实文本做推荐、分类、搜索等任务
	Moderation	2022.8	-	-	-	文字、图像	审查模型，能够检测仇恨、威胁、自残、性内容、涉及未成年人的性内容、暴力等内容
	Sora	2024.2	30	-	-	文字、视频	文本转视频模型，可以生成长达一分钟的视频，同时保持视觉质量并遵循用户的提示
	Phi-3	2024.4	38/42/70/140	-	-	文字、视觉	小型语言模型，在各种语言、推理、编码和数学基准测试中表现优于同等规模和下一个规模的模型
亚马逊	Titan	2023.9	-	Trainium	-	文字、图像	大语言模型，支持文本处理、生成代码、优化搜索结果、创建编辑图像等功能
Meta	LlaMa	2023.2	70/130/330/650	A100	1000/1400 Billion	文字	大语言模型，能够构建聊天机器人、生成内容、多语言语音识别
	LlaMa2	2023.7	70/130/700	A100/H100	2000 Billion	文字	改进的大语言模型，主要用于指令型语言模型，基于指令进行操作和回应
	LlaMa3	2024.4	30/80/700/900/4050		15000 Billion	文字	引入了分组查询注意力技术以提升推理效率和速度，并通过开发新堆栈、改进硬件可靠性、优化存储系统等提高训练效率
	ImageBind	2023.5	-	A100/H100	-	图像、视频、音频、文本、热测量和惯性测量单元	ImageBind 支持六种模态中任何输入，从而实现基于音频的搜索、跨模态搜索、多模态算术和跨模态生成

资料来源：各公司官网、华泰研究

谷歌：大模型推陈出新追赶 GPT，强化 AI 领域竞争优势

2023 年 12 月 6 日，谷歌宣布多模态大模型 Gemini 1.0 正式上线。模型分为 Gemini Nano、Gemini Pro、Gemini Ultra 三个版本，被用于从数据中心到移动设备的所有设备上，并将渗透于整个 Google 生态中。谷歌在大模型领域深耕多时：公司于 2023 年 3 月发布聊天机器人 Bard；4 月份将 Google Brain 和 DeepMind 人工智能实验室合并；5 月份在 Google I/O 大会上，宣布新实验室 Google DeepMind 开始研发 Gemini；历经半年大规模开发，团队成果得以面世。2024 年 5 月，谷歌于 I/O 大会宣布推出 Gemini 1.5 Pro 和轻量化版本 Gemini 1.5 Flash。2024 年 12 月，谷歌宣布推出 Gemini 2.0，先行推出 Flash 实验版本可供用户和开发者使用，目前尚未明确定价方式。

图表22：Gemini 四种版本介绍



资料来源：谷歌官网、华泰研究

1） **Gemini 1.5 Pro：首创大型模型长上下文窗口。**不仅升级了翻译、编码、推理、音频和图像理解等功能，且其私人预览版上下文窗口已达 200 万 tokens，对比远超 GPT-4o 的 12.8 万。在文本、代码、图像、音频和视频评估测试中，1.5 Pro 87%的表现优于 1.0 Pro，与 1.0 Ultra 大致相似。在 NIAH 评估中，1.5 Pro 在 99%的时间内都能从长文本块中找到所需嵌入文本。此外，Gemini 1.5 Pro 还具备情境学习技能，可从长信息中学习新技能，而不需要额外微调。

2） **Gemini 1.0 Ultra：谷歌最大、性能最强的模型。**用户能访问 Gemini Advanced 使用 Gemini Ultra。Gemini Ultra 能有效执行编码、逻辑推理、遵循复杂指令以及协作创意项目等复杂任务。此外，移动端用户也能使用 Android 和 iOS 来体验 Gemini Ultra。

3） **Gemini 2.0 Flash：最受开发者欢迎的模型，主打低延迟下的增强性能。**2.0 Flash 在关键基准测试中速度是 1.5 Pro 的两倍。除了支持图像、视频和音频等多模式输入外，2.0 Flash 还支持多模式输出，例如与文本混合的原生生成图像和文本转多语言音频。此外还可以原生调用 Google 搜索、代码执行以及第三方用户定义函数等工具。

4） **Gemini 1.0 Nano：设备端任务模型。**可在支持 Android 的设备上运行，此外，需要使用 Android 版 Google AI Edge SDK。2023 年 12 月，Google 官宣 Gemini Nano 将在 Pixel 8 Pro 上正式运行，有助于防止敏感数据离开手机，并提供在没有网络连接的情况下获取录制的对话、采访、演示等内容的摘要。

Gemini 模型采用多模态模型技术，发展前景广阔。Gemini 1.0 模型建立在 Transformer 之上，并使用 TPUv5e 和 TPUv4 进行训练与优化，Gemini 2.0 则基于最新的 Trillium TPU 上进行训练和推理。预训练数据集选取于网络文档、书籍和代码。谷歌没有像 OpenAI 构建 DALL·E 和 Whisper 那样单独训练图像和语音模型，而是直接建立原生多模态模型，这有助于 Gemini 无缝理解和推理各种输入，不仅优于 GPT-3.5 纯文字大语言模型，也不需要像 GPT-4 那样依赖插件和集成来实现推理、编码、文本、图像、音频等功能。

多维度基准测试展现 Gemini 性能，挑战 OpenAI 一枝独秀的地位。得益于谷歌专有数据与多模态特性的支持，Gemini 1.0 Ultra 在基准测试中的 30 项上领先于 GPT-4，但幅度较小。多任务语言方面，Gemini 1.0 Ultra 在 MMLU（大规模多任务语言理解）中的得分率高达 90.0%，首次超越人类专家，对比 GPT-4 的准确率为 86.4%；图像基准方面，Gemini 1.0 Ultra 无需从图像中提取文本，能直接进行 OCR 处理，在 MMMU 基准测试中准确率为 59.4%；代码处理方面，Gemini 创建的代码生成系统 AlphaCode 2 表现优于 87%的竞赛参与者，能够理解 Python、Java、C++等高质量代码。随着谷歌 AI 产品逐步商业落地，未来将成为 OpenAI 有力的竞争对手，估值低于微软的局面或将扭转。

Gemini 2.0 推出思维模式，对标 OpenAI o1 模型，在“幻觉”处理方面更进一步。基于 Gemini 2.0 Flash，谷歌经过专门训练推出 Thinking 模式，可使用思维（thoughts）来增强其推理能力。与 o1 思维链（chain of thoughts）的关键区别是，Gemini 2.0 Flash Thinking 在进行过程中会明确展示其推理过程，而 o1 则会隐藏其步骤，对于需要确保在长思维链中不会出现幻觉的领域来说，我们认为这是谷歌相对于 OpenAI 产品的重大进步。截止 2024 年 12 月，Gemini 2.0 Flash Thinking 模型已经跃居 Lmarena Chatbot Arena 的第一位，且在编程、数学、创意写作等各项评测任务上均为第一名。Targum 创始人和 CEO Alex Volkov 通过 10 个难题对 o1-2024-12-17 和 Gemini-2.0-flash-thinking 进行了对比测试，结果发现这两个推理模型的表现相当，而后者的速度更快。

图表23：Chatbot Arena LLM 评测排名

Chatbot Arena Overview

Model	Overall	Overall w/ Style Control	Hard Prompts	Hard Prompts w/ Style Control	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
gemini-exp-1206	1	1	1	1	1	1	1	1	1	1
gemini-2.0-flash-thinking-exp-1219	1	1	1	1	1	1	1	1	1	1
chatgpt-4o-latest-20241120	1	1	4	5	1	6	1	2	1	1
gemini-2.0-flash-exp	4	4	1	3	2	2	2	3	1	2
o1-preview	5	3	1	1	1	1	5	3	5	3
o1-mini	6	8	3	5	1	1	19	6	6	6
gemini-1.5-pro-002	6	7	7	8	8	6	5	7	6	9
grok-2-2024-08-13	8	11	12	13	10	10	7	10	9	9
yi-lightning	8	12	7	10	8	6	7	9	7	6
gpt-4o-2024-05-13	8	7	10	10	8	10	7	9	8	8
claude-3-5-sonnet-20241022	8	6	7	1	6	6	7	6	6	6
deepseek-v2.5-1210	8	16	7	10	8	8	7	9	6	8
qwen2.5-plus-1127	8	20	7	9	7	5	11	7	7	6
athene-v2-chat	10	19	7	10	8	6	23	9	7	9
glm-4-plus	12	17	11	15	10	13	11	11	9	10
gpt-4o-mini-2024-07-18	12	19	14	23	10	19	9	14	9	10
gemini-1.5-flash-002	12	20	21	27	26	13	7	14	9	26
llama-3.1-nemotron-70b-instruct	12	32	12	17	10	11	7	14	27	11
claude-3-5-sonnet-20240620	15	8	11	8	8	6	23	9	9	8

资料来源：Chatbot Arena，华泰研究

Gemini 定价对标 GPT Plus，踏上大模型商业化落地万里长征的第一步。目前，用户能免费使用 Gemini 聊天机器人，谷歌 Gemini Ultra 1.0 通过 Google One AI Premium 提供，具有两个月免费试用期，此后订阅价格为 19.99 美元/月。对比 OpenAI 提供“GPT Plus”订阅价格为 20 美元/月。新贵 AI 搜索公司 Perplexity 同样将其 Perplexity Pro 版本价格定位在 20 美元/月。Google Gemini 开发者模式则分不同模型定价，其中 1.5 Flash 免费版本速率限制在 15RPM，付费版本无限制，价格与 GPT-4o-mini 对标；1.5 Pro 免费版本速率限制在 2RPM，付费版本无限制，价格与 GPT-4o 对标。

图表24：Gemini API 定价

模型	免费版限制	价格	
		提示词<128k tokens	提示词>128k tokens
Gemini 1.5 Flash	15 RPM 1 million TPM 1,500 RPD	Input Pricing: $0.075 / 1M tokens Output Pricing: $0.30 / 1M tokens Context Caching: $0.01875 / 1M tokens	Input Pricing: $0.15 / 1M tokens Output Pricing: $0.60 / 1M tokens Context Caching: $0.0375 / 1M tokens
Gemini 1.5 Pro	2 RPM 32,000 TPM 50 RPD	Input Pricing: $1.25 / 1M tokens Output Pricing: $5.00 / 1M tokens Context Caching: $0.3125 / 1M tokens	Input Pricing: $2.50 / 1M tokens Output Pricing: $10.00 / 1M tokens Context Caching: $0.625 / 1M tokens
Gemini 1.0 Pro	15 RPM 32,000 TPM 1,500 RPD	Input Pricing: $0.50 / 1M tokens Output Pricing: $1.50 / 1M tokens	
GPT-4o		Input Pricing: $2.50 / 1M tokens Output Pricing: $10.00 / 1M tokens Context Caching: $1.25 / 1M tokens	
GPT-4o-mini		Input Pricing: $0.15 / 1M tokens Output Pricing: $0.60 / 1M tokens Context Caching: $0.075 / 1M tokens	
o1-preview		Input Pricing: $15.0 / 1M tokens Output Pricing: $60.0 / 1M tokens Context Caching: $7.5 / 1M tokens	

注：RPM (requests per minute) 表示每分钟请求数，TPM (tokens per minute) 表示每分钟 tokens 使用数量，RPD (requests per day) 表示每天请求数；
资料来源：谷歌官网、华泰研究

全新轻量化大语言模型 Gemma 系列除在 TPU v5e 和 v5p 上优化以外，也针对英伟达芯片进行优化，跨设备兼容性提升。2024 年 2 月，谷歌推出首款轻量化开源大语言模型 Gemma，其采用与 Gemini 相同的技术构建，分为 20 亿参数和 70 亿参数两种规模，更类似 Gemini Nano 模型的 18 亿和 32.5 亿参数量。Gemma 配备多框架工具，能够在 Keras 3.0、本机 PyTorch、JAX 和 TensorFlow 进行推理和微调。此外，Gemma 还具备跨设备兼容性，可在笔记本电脑、台式机、物联网、移动设备、谷歌云中运行，并可部署在 Vertex AI 和 Google Kubernetes Engine（GKE）上。2024 年 5 月，谷歌宣布宣布推出 Gemma 2（27B 参数），其采用全新架构，通过算法改进实现轻量化，其性能可媲美 Meta 参数更大的模型 Llama 3（70B 参数）。

图表25：Gemma 2 与 Llama 3 和 Grok-1 基准测试对比

			Gemma 2		Llama 3		Grok-1
	BENCHMARK	METRIC	9B	27B	8B	70B	314B
General	MMLU	5-shot, top-1	71.3	75.2	66.6	79.5	73.0
Reasoning	BBH	3-shot, CoT	68.2	74.9	61.1	81.3	–
	HellaSwag	10-shot	81.9	86.4	82	–	–
Math	GSM8K	5-shot, maj@1	68.6	74.0	45.7	–	62.9 (8-shot)
	MATH	4-shot	36.6	42.3	–	–	23.9
Code	HumanEval	pass@1	40.2	51.8	–	–	63.2 (0-shot)

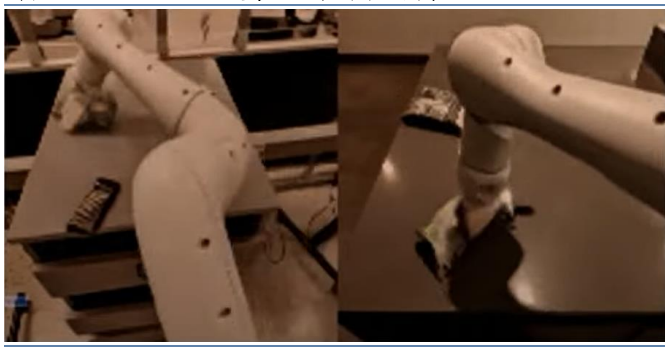
资料来源：谷歌官网、华泰研究

谷歌 Genie 打造可交互生成式世界。2024 年 2 月，谷歌发布 Genie，作为谷歌继推出大模型 Gemini、开源大模型 Gemma 之后的新模型。Genie 能接受文本和图像提示，并生成类似视频游戏的交互式环境。Genie 参数量达 110 亿，在 2D 平台游戏的超过 200000 小时的视频上进行训练。Genie 的核心组件基于 Vision Transformer 构建，可用于处理视频等具有时间和空间维度的数据，底层数据库则基于大量游戏视频建立。不同于 Sora、Runway 等模型，Genie 生成的内容具备可交互属性，用户可通过文本对所生成虚拟环境中的角色动作进行操控。此外，Genie 由潜在动作模型、视频分词器、动态预测模型三大核心组件组成，其不仅能理解并推理每帧之间的潜在动作，还能逐帧预测视频，并生成符合运动规律的序列帧。Genie 也具备可模拟性，其能通过短视频模拟物体的动态变化来训练机械臂等多功能智能体，有助于机器人的发展。**2024 年 12 月，谷歌发布 Genie 2，在上一代的基础上实现了通用性的飞跃，不同于 Genie1 只能生成 2D 元素，Genie 2 能够生成丰富的 3D 世界，并具备例如对象交互、复杂的角色动作、物理建模以及预测其他 NPC 行为的能力。**

图表26：Genie 2 能够生成不同场景轨迹

资料来源：谷歌官网、华泰研究

图表27：Genie 模拟可变形物体来训练机械臂



资料来源：谷歌官网、华泰研究

2023 年初至今谷歌加码投资多家 AI 独角兽，丰富 AI 版图。伴随 2022 年末以 ChatGPT 为首的生成式 AI 走入大众视野，谷歌在 2023 年对各领域 AI 初创企业的投资持续推进。根据 Pitchbook 数据，自 2023 年起，谷歌已经投资包括 Anthropic、Hugging Face、Runway 在内的多家 AI 独角兽企业。我们认为此举或出于谷歌对拓宽业务渠道与巩固生态壁垒的需求，如推出 Claude 3 的初创公司 Anthropic。

2023 年初，谷歌已对 Anthropic 投资 3 亿美元，并获得该公司 10%的股权，2023 年 10 月谷歌再次向 Anthropic 追加投资 20 亿美元，AWS 也于 2024 年 3 月向 Anthropic 追加 27.5 亿美元投资，总投资额达 40 亿美元，使得 AWS 成为其主要云提供商。24 年 11 月 Anthropic 宣布将使用亚马逊的 Trainium 和 Inferentia 芯片来训练和部署其未来的基础模型。**我们认为，谷歌二度注资 Anthropic 能增强谷歌云在 AI 模型部署上的生态丰富程度，同时也能在与 OpenAI 的模型竞争中减少新树敌，未来科技巨头争夺 AI 霸主的军备竞赛或更为激烈。**

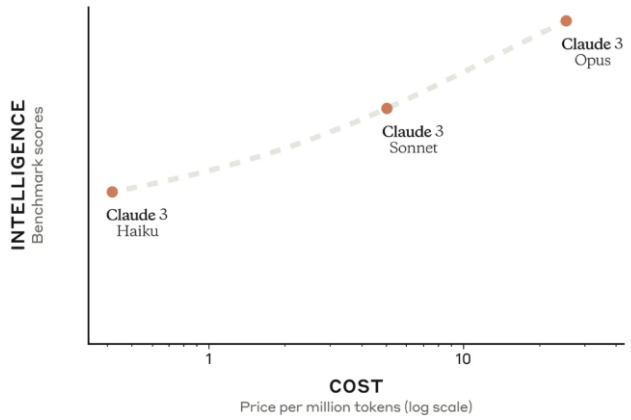
图表28：谷歌 2023 年以来所投资的部分公司

行业	公司名称	交易时间	融资总金额（百万美元）	被投资公司主营业务	合作与投资方面其他情况
AI+能源	Intersect Power	2024.12	-	可再生能源和存储解决方案提供商	此次合作将生产足够的清洁电力来驱动数个千兆瓦级数据中心，第一阶段将于 2026 年投入运营，并于 2027 年全面完成
软件	Cameyo	2024.6	-	开发虚拟化工具以在 ChromeOS 设备上运行 Windows 应用程序	此次收购将使谷歌 ChromeOS 受益，因为用户能够直接使用 Windows 应用程序，而无需复杂的安装或更新
电子商务	Flipkart	2024.5	350	印度电子商务市场的领军企业，为数亿消费者提供服务	谷歌还将向 Flipkart 提供云服务，沃尔玛和微软也参与了投资
AI+软件	Charm	2023.11	5.87	开发开源编码平台，为命令行界面（CLI）提供强化视觉外观的工具。	由谷歌的风险投资基金 Gradient Ventures 带头
AI	Essential AI	2023.10	51.5	提供基于大模型的全栈型智能产品	英伟达和 AMD 也参与此轮投资
AI	Anthropic	2023.10	2000	专注于开发通用 AI 系统和语言模型，2024 年 3 月发布最新 Claude 3 系列大模型	亚马逊也参与此轮投资，谷歌本轮投资为上轮追加
AI+软件	Hugging Face	2023.8	235	自然语言处理（NLP）模型库和社区，提供大量预训练模型、工具和资源	亚马逊、英伟达、英特尔、 AMD 等公司也参与此次 D 轮融资
AI+教育	Elsa	2023.7	23	AI 驱动的语言平台，旨在帮助非英语母语学习者提高语言和发音	由谷歌的风险投资基金 Gradient Ventures 投资，资金将用于支持 Elsa 平台扩大其全球产品
AI+多媒体	Runway	2023.6	191	人工智能视频软件公司，致力于提供文生视频的 AI 服务	英伟达和 Salesforce 也参与此轮投资
AI+货币	WorldCoin	2023.5	115	加密货币项目，通过虹膜扫描实现身份认证	由区块链资本(Blockchain Capital)牵头
AI+房地产	NoBroker	2023.3	198.89	房地产中介公司，印度第一家房地产科技独角兽公司，23 年 10 月宣布推出 CallZen.AI	资金将用于简化整个房地产流程的技术研发
AI	Anthropic	2023.2	300	专注于开发通用 AI 系统和语言模型，2024 年 3 月发布最新 Claude 3 系列大模型	谷歌获得 Anthropic 约 10%的股份，并与 Anthroic 签订了一项大型云计算合同
云计算	Chronosphere	2023.2	115	云原生可观测性平台，23 年 4 月与谷歌云建立市场合作伙伴关系。	资金将用于本地云观测平台的创新和上市准备
AI+无人驾驶	Oxa	2023.2	140.92	为自动驾驶汽车开发软件，主要产品包括组件 Oxa Driver 和用于模拟无人驾驶汽车测试的生成性 AI 工具 Oxa MetaDriver	该项投资预计会促使 Oxa 与谷歌在自动驾驶项目上展开合作

资料来源：Pitchbook 官网、CNBC 官网、华泰研究

Anthropic 由 OpenAI 前研究副总裁 Dario Amodei 等人于 2021 年创立，旗下 Claude 系列模型基于 Transformer 架构，且具备推理、视觉分析、代码生成与多语言处理等功能，全面对标 GPT 系列。2023 年 7 月 11 日发布的 Claude 2 单次可处理高达 10 万 Tokens 的上下文窗口，对比 GPT-4.0 Turbo 为 12.8 万 Tokens。而在 MMLU 准则中，Claude 2 评分也仅次于 GPT-4，且已被 Slack、Notion 和 Quora 等众多公司使用。2024 年 3 月 4 日，Anthropic 宣布推出 Claude 3 系列，包括 Haiku、Sonnet 和 Opus 三种型号，能支持文字与图像输入，并在速度与准确性优化的前提下支持 20 万 Tokens 的上下文窗口。其中，Claude 3 Opus 首次在 MMLU、GPQA 与 GSM8K 等 10 项评分准则中超越了 GPT-4 与 Gemini 1.0 Ultra，在 3 月发布之后曾经成功登顶 Chatbot Arena 排行榜第一，目前仍保持前十地位，成为 OpenAI 在 LLM 领域的有力竞争者。

图表29：Claude 3 系列模型性能与成本对比



资料来源：Anthropic 官网、华泰研究

图表30：Chatbot Arena 排行榜

排名	模型	Arena Elo 得分
1	Gemini-Exp-1206	1372
1	Gemini-2.0-Flash-Thinking-Exp-1219	1368
1	ChatGPT-4o-latest (2024-11-20)	1364
4	Gemini-2.0-Flash-Exp	1354
5	o1-preview	1335
6	o1-mini	1306
6	Gemini-1.5-Pro-002	1302
8	Grok-2-08-13	1288
8	Yi-Lightning	1287
8	GPT-4o-2024-05-13	1285
8	Claude 3.5 Sonnet (20241022)	1283
8	Deepseek-v2.5-1210	1278
8	Qwen2.5-plus-1127	1278
10	Athene-v2-Chat-72B	1277

注：数据截至 2024 年 12 月
资料来源：LMSYS Chatbot Arena 官网、华泰研究

微软：多元布局稳固市场，携手 OpenAI 欲打造 AI 大模型帝国

微软多次向 OpenAI 投资，巩固自身在 AI 大模型领域地位。OpenAI 成立于 2015 年，2019 年 3 月，OpenAI 设立 OpenAI LP 并转型为有限盈利公司，同年 7 月微软向 OpenAI 注资 10 亿美元。双方开始在 Azure 云计算服务上合作开发 AI 超级计算技术。同时，OpenAI 逐渐将云计算服务从谷歌云迁移到 Azure。2023 年 1 月 23 日，微软以 290 亿美元的估值投资 OpenAI 约 100 亿美元，并获得 OpenAI 49%的股权，扩大与 OpenAI 的合作伙伴关系。**目前，OpenAI 提供语言类大模型 GPT 系列、图像多模态大模型 DALL·E、通用语音识别模型 Whisper、文本向量化模型 Embeddings、文本审查模型 Moderation 等丰富模型。**微软拥有 GPT-4 和所有其他 OpenAI 人工智能模型的独家授权，并通过 Azure OpenAI 服务提供对 OpenAI 模型的 REST API 访问。

GPT 系列：基于 Transformer 架构的大型语言模型，能理解和生成自然语言和代码。OpenAI 在 2018 年 6 月推出 1.17 亿参数的初代生成式预训练模型 GPT-1 后，陆续推出了 GPT-2（15 亿参数）、GPT-3（1750 亿参数）、GPT-3.5（2000 亿参数），并于 2022 年 11 月推出基于 GPT-3.5 的 AI 对话聊天机器人 ChatGPT。2023 年 3 月，OpenAI 发布 GPT-4，用户能通过 ChatGPT Plus 获得 GPT-4 有限访问权限。

2024 年 5 月，OpenAI 推出 GPT-4o，朝着更自然的人机交互前进。GPT-4o 作为首个采用输入+输出由同一个神经网络处理的多模态模型，接受文本、音频、图像和视频的任意组合作为输入/输出，且可在最短 232 毫秒内响应音频输入，平均输出延迟为 320 毫秒，与人类对话反应时间相似，对比 GPT-3.5 和 GPT-4 分别为 2.8 秒/5.4 秒。此外，GPT-4o 在文本、推理和编码方面能与 GPT-4 Turbo 相当，在多语言、音频和视觉能力上的表现则有显著改善。目前，GPT-4o 已能在 Azure OpenAI 服务 API 和 Azure AI Studio 中提供，其输入/输出成本对比 GPT-4.0 Turbo 降低了 50%。

图表31：GPT-3.5 Turbo、GPT-4、GPT-4.0 Turbo 与 GPT-4o 参数对比

	GPT-3.5 Turbo	GPT-4.0	GPT-4.0 Turbo	GPT-4o
训练日期	2021.9	2021.9	2023.12	2024. 5
参数数量	1750 亿	大于 10000 亿	-	-
输入类型	文本	文本、图像	文本、图像、文本转语音	文本、音频、图像和视频
上下文窗口	16385 (GPT-3.5 Turbo-0125)	8192 (GPT-4) 32768 (GPT-4-32K)	128000	128000
最大输出令牌	4096	4096	4096	-
模型定价	输入：$0.50 /1 M tokens 输出：$1.50 /1 M tokens	输入：$30 / 1 M tokens 输出：$60 / 1 M tokens	输入：$10 / 1 M tokens 输出：$30 / 1 M tokens	输入：$5 / 1 M tokens 输出：$15 / 1 M tokens

资料来源：OpenAI 官网、华泰研究

2024 年 9 月，OpenAI 推出 o1 模型，通过强化学习和“思维链”增强推理能力，分为 preview 和 mini 版本，并于 12 月推出正式版。o1 能解决比以前更复杂的问题，尤其是在科学、数学和编程领域，o1-preview 在国际数学奥林匹克竞赛(IMO)中，GPT-4o 仅正确解决了 13%的问题，而 o1 得分为 83%。但作为早期模型，它还不具备 ChatGPT 的许多实用功能，例如浏览网页信息以及上传文件和图片。

亚马逊：长期软硬件积淀托底，多维度布局生成式 AI 领域

亚马逊以芯片-模型-应用构筑 AI 产业大厦，多层布局优势显著。与 ChatGPT 专注于顶部应用程序层不同，亚马逊首先通过使用 Trainium 和 Inferentia 芯片提供基础层服务。此外，由全托管基础模型服务平台 Amazon Bedrock 提供中间层服务，并提供 AI 模型 Amazon Titan，以多维度布局生成式 AI 领域。

Amazon Bedrock 提供业界多款基础模型的访问权限，资源优势较为显著。2023 年 9 月 25 日，亚马逊向 Anthropic 投资 40 亿美元，成为其主要云提供商，Anthropic 将利用 Trainium 和 Inferentia 来训练和部署模型。同时，Anthropic 承诺为 AWS 客户提供通过 Amazon Bedrock 访问其 Claude 的权限。2023 年 11 月 29 日，AWS 宣布 Meta 的大型语言模型 Llama 2 13B 与 70B 将托管在 Amazon Bedrock 中。此外，Amazon Bedrock 还提供了来自 Cohere、Stability AI、AI21 Labs 等丰富的模型库。

Amazon Titan 自研大语言模型，通过完全托管的 API 提供三类模型：用于内容生成的文本模型、可创建矢量嵌入的多模态嵌入以及图像生成模型，主要应用场景为广告领域的精准推荐和图像生成。

图表32：Amazon Titan 模型版本与参数

模型版本	最大令牌数	语言	支持微调	应用场景
Titan Text Express	8K	英语，100+种语言（预览版）	是	检索增强生成、开放式文本生成、头脑风暴、汇总、代码生成、表格创建、数据格式化、释义、思维链、重写、提取、问答聊天
Titan Text Lite	4K	英语	是	摘要和文案写作
Titan Text Embeddings	8K	25 种以上语言	否	文本检索、语义相似度和集群化
Titan Multimodal Embeddings	128	英语	是	搜索、推荐、个性化
Titan Image Generator (preview)	77	英语	是	文本到图像生成、图像编辑、图像变体

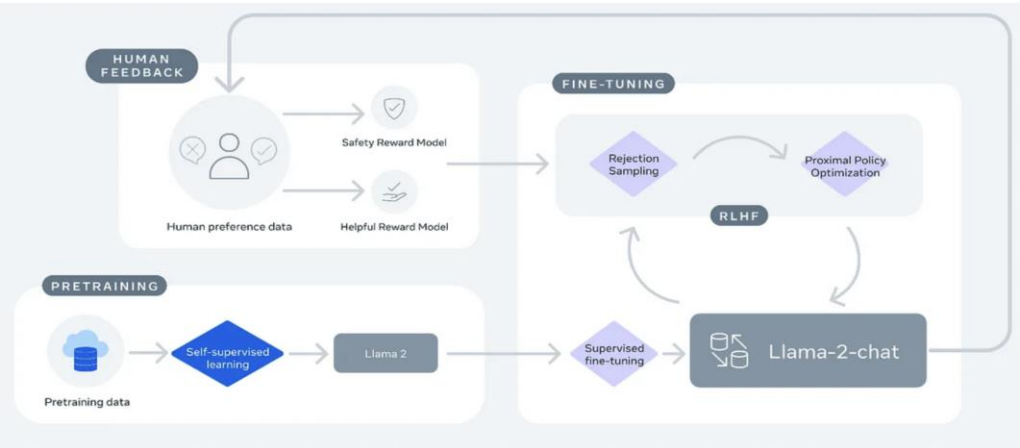
资料来源：AWS 官网、华泰研究

Meta：开源模式另辟蹊径，发展数据融合技术切入模型市场

Meta 在 OpenAI、谷歌主导的 AI 闭源模型之外开辟开源 AI 模型赛道，推出 LlaMa 系列大模型，并在 2023 年推出开源文本生成音乐模型 Audiocraft、公开多模态视频数据集 Ego-Exo4D、开源视觉模型 DINOv2。此外，2024 年 3 月 Facebook 负责人 Tom Alison 表示，Meta 到 2026 年的技术路线图涉及开发人工智能推荐模型，以支持类似 TikTok 的 Reels 短视频服务和传统的较长视频，为整个视频生态系统助力。

LlaMa 仅用约 1/10 的参数规模，实现了媲美 OpenAI GPT-3、DeepMind Chinchilla、谷歌 PaLM 等超大模型的性能。我们认为，Llama 竞争力的关键是实现改进算法以减小模型参数，其不仅能降低算力需求，还可加速 AI 与边缘端融合。2023 年 2 月，Meta 发布大语言模型 LlaMa，LLaMa 采用 Transformer 架构，分为 70 亿、130 亿、330 亿和 650 亿四种参数规模。LlaMa 2 于 2023 年 7 月 18 日发布，架构基本保持不变，但预训练数据增加到 2 万亿个 tokens，上下文长度增加了一倍。2024 年 4 月，Meta 发布 Llama 3，首发 80 亿和 700 亿参数版本。对比同参数规模的模型（谷歌 Gemma、Llama 2），Llama 3 表现较佳。

图表33：LLaMA 2 架构图



资料来源：Meta 官网、华泰研究

字节跳动：以豆包大模型为基准，快速工厂式复制 AI 应用

2023 年 8 月，字节跳动推出了底层大模型“云雀”，并随后推出了人工智能对话产品“豆包”。2023 年 11 月，字节跳动成立专注于 AI 应用的 Flow 部门，将其提升至与抖音、火山、飞书等同等重要的业务部门地位。为了统一字节跳动旗下的 AI 产品品牌，增强市场认知度和用户接受度，2024 年 5 月云雀大模型正式更名为豆包大模型。在 2024 年 5 月至 12 月间，豆包大模型的日均 tokens 数据呈现快速增长的趋势，从 1200 亿增长到 40000 亿，增幅最大，最新数据已达所有模型之首。定价方面，豆包通用大模型为每千输入 tokens 0.0008 元，约为 GPT-4o 价格的 1/23，GPT-4o-mini 价格的 3/4，豆包视觉理解大模型为每千输入 tokens 0.003 元，约为 GPT-4o 价格的 1/6，具备相当的价格优势。

在豆包大模型的技术基础上，我们认为字节跳动的主要竞争力为基于各个场景推出不同 AI App，积极布局 AI 市场。字节跳动在 AI 领域采取了“火力覆盖”的战略，意在不遗漏任何一个潜在的市场机会。通过推出多款产品，字节能够与市场上的优质竞品进行直接竞争，迅速占领市场份额。例如，豆包对标 ChatGPT 和百度文心一言，而猫箱则与 MiniMax 旗下的社交 AI 产品星野竞争。“豆包”是字节跳动整合封装豆包大模型能力推出的同名 AI 助手产品，囊括搜索、写作、数据分析、图像生成、音乐生成等多种能力，有 PC 端、移动端、网页端、浏览器插件 4 种产品形态。在 2024 年 11 月的 AI 产品全球月活跃排行榜中，豆包 App 的 MAU 达到了 5998 万，仅次于 OpenAI 的 ChatGPT，位列全球第二，全国第一。

图表34：字节跳动旗下 AI 产品汇总

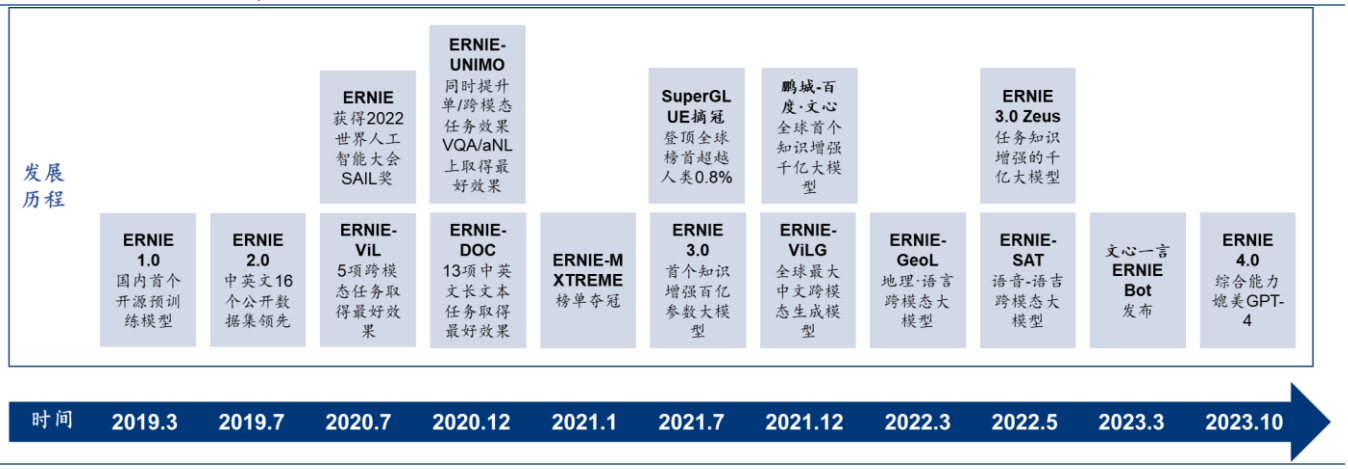
字节 AI 应用			
国内版	海外版	对标产品	功能
豆包	Cici	ChatGPT	聊天机器人、办公助手
猫箱	Anydoor	Character AI，星野	AI 虚拟角色对话
扣子	Coze		AI 应用开发平台
星绘	PicPic	妙鸭相机	AI 写真集
豆包爱学	Gauth	Question.AI	AI 拍照搜题
即梦 AI	Dreamina	Sora	AI 视频、图像生成
小悟空	Chitchop	HIX AI	AI 工具集平台，包括智能对话、翻译、创意生成等
即创		OptiMonk, Jasper, Synthesia	AI 电商广告制作
海绵音乐		Suno	AI 音乐生成
Marscode		GitHub Copilot	AI 代码生成

资料来源：字节跳动官网、华泰研究

百度：文心大模型成为国内 AI 模型的早期实践者

文心大模型 ERNIE 涵盖 NLP、视觉、跨模态、生物计算和行业模型五大领域，为国内 AI 大模型的早期实践者。2019 年 3 月，百度基于早期的神经网络语义匹配技术，发布预训练模型文心 ERNIE1.0，其在语言推断、语义相似度、命名实体识别、情感分析和问答匹配等 NLP 中文任务上超越 BERT；同年 7 月，百度发布文心 ERNIE2.0，在 16 项中英文任务上超越 BERT 和 XLNet。目前，文心大模型包含 NLP 大模型、CV 大模型、跨模态大模型、生物计算大模型、行业大模型，截至 2024 年 2 月，文心大模型已形成“模型层+工具与平台层+产品与社区层”的整体布局，整体日调用量超过 5000 万次。

图表35：百度文心大模型发展阶段



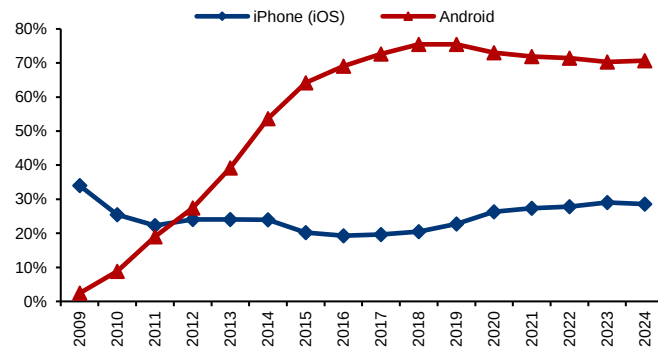
资料来源：百度智能云官网、华泰研究

百度自身优势赋能+企业合作双轮驱动模型端商业化进程。我们认为，百度早期搜索业务为 AI 布局提供基底，其 AI 架构具备一定的垂直一体式布局，并有多家合作伙伴接入文心生态，对比国内其他大模型虽具备先发优势，但面对来势汹汹的后起之秀豆包，能否脱颖而出仍需观察。

安卓生态或有端侧 AI 爆发潜力

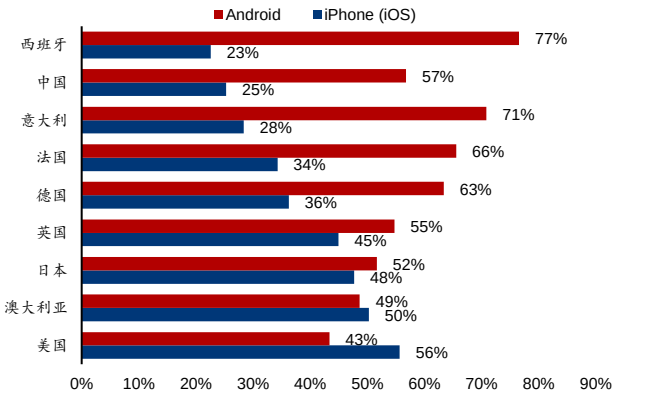
安卓系统已经在全球范围内拥有超过 30 亿活跃设备，这一庞大的用户基础为谷歌带来了无与伦比的数据资源和市场影响力。安卓生态的开放性和灵活性，使得谷歌能广泛覆盖从入门级到高端旗舰的各种设备。谷歌将其“GMS”(Google Mobile Services Stack，谷歌移动服务堆栈)捆绑到 Android OEM 协议中，这使得 Play store、Chrome、谷歌地图、谷歌搜索和其他服务在 Android 设备上占据近 100%的份额，为 AI 技术的普及和商业化提供了天然的优势。

图表36：安卓和 iOS 系统全球市场份额趋势



资料来源：StatCounter、华泰研究

图表37：安卓和 iOS 系统各国市场份额（23Q4）



资料来源：Kantar、华泰研究

谷歌正积极将 AI 技术嵌入安卓设备中，以提升用户体验并开辟新的收入来源。谷歌搜索以 Circle to Search 的方式融入安卓硬件生态中，用户能够通过，圈出、突出显示、涂鸦或点按图片、视频或文字的方式，即时搜索在手机上看到的任何内容，并获得由 AI Overviews 所总结的结论。另外，通过对海量数据的处理，谷歌可精确地了解用户需求，从而在手机端推出定制化广告、推荐系统等商业化应用。

谷歌将 AI 助手集成于 Pixel 及部分安卓手机，移动端 AI 已初见成效。在 made by Google 2024 发布会上，谷歌展示了如何通过 AI 提升 Pixel 手机的摄像头功能，如实时优化照片效果、自动生成短视频等。随着 Gemini 与 Android 深度集成，谷歌正在以 AI 为核心重建操作系统，并重新定义手机的功能。例如，用户可在应用顶部调出 Gemini 的叠加层，询问屏幕上内容的问题，包括找到正在观看的 YouTube 视频的具体信息，还可直接从叠加层生成图像，并将其拖放到 Gmail 和 Google Messages 等应用中。

图表38： Pixel 9 通过 AI 优化合照

资料来源：谷歌官网、华泰研究

图表39： Gemini 与安卓手机深度集成

资料来源：谷歌官网、华泰研究

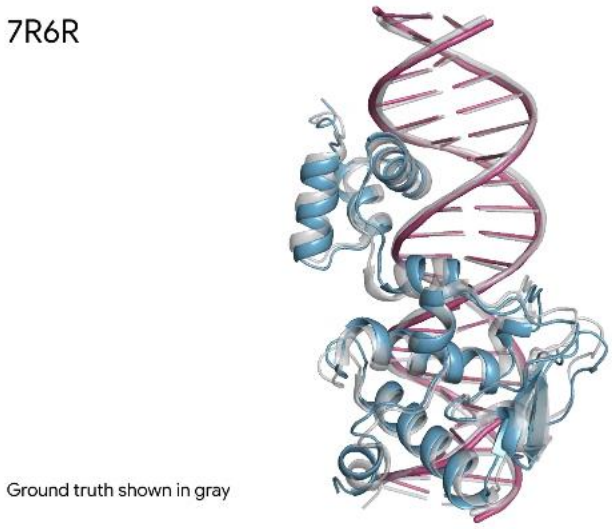
AlphaFold 专攻 AI 医疗，走在 AI 落地 B 端商业化前沿

谷歌的 AlphaFold 是旗下 DeepMind 开发的一项医疗领域人工智能模型，主要应用于药物开发、生物学研究和新蛋白质设计等场景。AlphaFold 能够：1）通过准确预测靶点蛋白质的三维结构，加速新药研发过程；2）理解蛋白质功能、相互作用及其在细胞中的角色，促进基础科学研究的进展；3）激发新算法的开发，例如设计自然界中不存在的新型蛋白质，为合成生物学提供了新的工具。

AlphaFold 的研究始于 2017 年，目前已发展至第三代。2018 年，DeepMind 推出 AlphaFold 1，并在 CASP13 竞赛中获得冠军，展示了其在蛋白质结构预测方面的潜力。该模型结合了 ResNet 网络架构和共进化信息。2020 年，AlphaFold 2 发布，该模型在 CASP14 竞赛中取得了 92.4 的 GDT-score。AlphaFold 2 利用 Transformer 和大量数据库数据进行训练，使得结构预测的准确率超过 90%，全球数百万研究人员已使用 AlphaFold 2 在疟疾疫苗、癌症治疗和酶设计等领域获得新发现。

2024 年 5 月，DeepMind 推出了 AlphaFold 3，它不仅能预测单体蛋白质的结构，还能模拟蛋白质与 DNA 或 RNA 等其他分子的相互作用，进一步推动了生物预测模型的发展。AlphaFold 3 采用新一代架构和覆盖所有生命分子的训练体系，核心是在基于 AlphaFold 2 改进的 Evoformer 模块。在处理输入的分子列表后，AlphaFold 3 使用类似于 AI 生成图像的扩散网络组装预测结果，生成相互配合的三维结构。AlphaFold 3 能够对包括蛋白质、DNA 和 RNA 在内的大型分子进行建模，也可以对小分子（即配体，包括许多药物）建模。此外，AlphaFold 3 还能够建模分子的化学修饰，化学修饰调控着细胞的正常功能，当被破坏时可能导致疾病。

图表40：AlphaFold 3 预测的分子复合物（蓝色、粉色）与发现的真实分子结构（灰色）几乎匹配



资料来源：谷歌官网、华泰研究

我们认为，大模型在医疗行业如创新药研发，具备充足落地条件，鉴于可通过物理和化学定律有效控制和监管幻觉的出现，从而受其影响较少。生物学结构具备一定的规则，因此在大模型优化方向上更具备目的性，AlphaFold 能够通过技术优化、数据增强和置信度设置使大模型在进行蛋白质预测时提高准确性并减少“幻觉”问题：

- **Cross-Distillation（交叉蒸馏）：**AlphaFold 3 使用了交叉蒸馏技术，集成了来自其前身 AlphaFold 2 和 AlphaFold-multitimer 的预测，能够用预测结构的例子丰富训练数据集。这种方法有助于模型从以往的经验中学习，从而提高其处理紊乱区域的准确性，显著减少幻觉行为。
- **Diffusion（扩散）技术：**AlphaFold 3 使用基于扩散的生成技术，从噪声输入开始并逐步完善，以产生准确的预测。尽管扩散技术可能会带来幻觉，但 AlphaFold 3 可以直接在三维原子坐标上操作，而不是依赖传统方法，有助于在管理复杂性的同时保持预测准确性。
- **专业增强的培训数据：**该模型已经在更广泛和多样化的数据集上接受了训练，其中包括容易产生幻觉的结构。通过在训练期间将 AlphaFold 3 暴露在更广泛的蛋白质配置中，它学会了生成更多生物学相关的结构，同时最大限度地减少生成不可信结构的可能性。
- **损失函数优化：**AlphaFold 3 包含复杂的损失函数，不仅能够提高结构的准确性，而且还教会模型预测其生成结构的置信度。这种自我评估能力使得 AlphaFold 3 能够识别生成预测的质量高低。
- **置信度设置：**AlphaFold 3 还包括评估原子级别和最终输出中的成对错误的置信度措施。这种机制使我们能够从不同的样本中选择更高质量的结构，进一步减轻幻觉的影响。

广告业务：谷歌份额领先，但 AI 搜索进度仍需观察

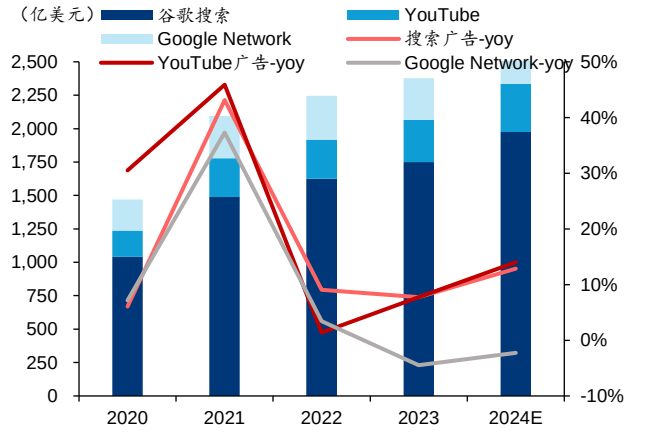
展望全球广告市场，受益于经济前景改善，我们预计 2025 全球广告行业恢复超预期，最大广告市场美国增长势头良好。我们预期全球广告行业 23-26 年 CAGR 为 8.8%，同时线上渗透率保持扩张，支出占比将从 23 年 68.0%抬升至 26 年 74.0%。

我们预计谷歌 Services 业务 FY24/25/26 营收为 3046/3332/3632 亿美元，对应同比为 11.8%/9.4%/9.0%。

广告是谷歌的主要营收来源。FY23 广告收入 2378.6 亿美元(+6.0% yoy)，占总收入 77.4%。**广告业务由三大板块构成：**

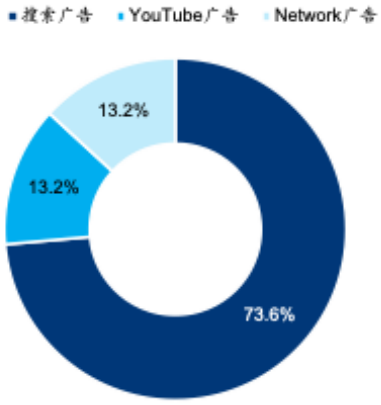
1） **谷歌搜索和其他广告，**包括 Google 搜索引擎广告，及 Google 其他自有产品广告收入，如 Gmail、Google Maps 及 Google Play。搜索广告是谷歌广告基本盘；该板块 FY23 收入 1750.3 亿美元（占广告收入 74%），+7.7% yoy；
2） **YouTube 广告，**包括 YouTube 移动、PC 及电视大屏端广告收入。乘线上视频东风，YouTube 广告近年增长动能较强，18-23 CAGR 23.1%，但 22 年以来增速略有放缓。FY23 创收 315.1 亿美元，+7.8% yoy；
3） **Google Network 广告，**覆盖第三方合作媒体资源，FY23 营收 313.1 亿美元，-4.5% yoy。

图表41：谷歌各类型广告收入规模及增速



资料来源：公司公告，华泰研究

图表42：谷歌各类型广告收入占比（FY2023）



资料来源：公司公告、华泰研究

谷歌于 21 年推出 AI 驱动的自动化投放工具 Performance Max（简称 Pmax）。该工具集成了谷歌各类自有和第三方广告资源，覆盖搜索、视频、购物、发现和展示广告网络（Google Display Network）。该工具优势在于：**1) 全自动：**用户可根据投放目标和预算，一键生成广告计划，优化难度低；**2) 数据互通：**过往广告主需针对不同目标，分别设置广告计划，跨计划数据追踪、比对和归因难度较高；此外，随着 Cookies 使用率下降，此前部分广告资源数据可用性受限。

对比其他广告位，Pmax 广告凭借出色投放效果获得高单价。据 Gupta Media，24 年以来 Pmax 全球平均 CPM 约为 9.90 美元，+13.6% yoy，显著高于 Google Display（4.65 美元）与 YouTube 广告（1.67 美元），并高于 Meta 旗下 Facebook 和 Instagram CPM 均值（5.83 美元）。

图表43： 谷歌 Performance Max 与 Meta Advantage+对比

	Meta Advantage+	Performance Max
推出时间	2022	2021
广告类型	购物广告、展示广告、视频广告、文本广告	购物广告、展示广告、视频广告、文本广告
产品	多种套件，针对不同场景	单款产品
优化目标	每款产品专注一个优化目标，如 Placement Campaign 在给定预算基础上最大化曝光量，拉动销售增长，更加专业细分	根据广告主需求选择不同优化目标
广告目标	推动销售、吸引客户、提升网站流量	推动销售、吸引客户、提升网站流量
目标客群	电商客户、应用开发者、内容创作者等	电商客户、应用开发者、内容创作者等
广告位投放	可选合作网站投放广告	某些类型广告强制在所有合作网站上投放
出价方式	智能出价 + 手动出价	智能出价 + 手动出价
社交互动	支持	不支持
投放渠道	Facebook，Instagram，WhatsApp	3P Display, Gmail, YouTube, Discover, Maps.等

资料来源：Meta 官网，Google 官网，华泰研究

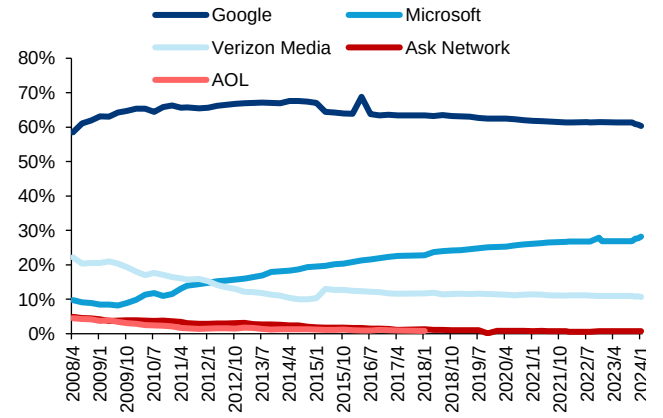
谷歌搜索广告：份额稳定居于全球头部，广告份额受电商和 AI 应用挑战

谷歌搜索广告主要依托 Google Search 搜索引擎，汇集全球大部分搜索流量。据 Similar Web，Google.com 访问量居全球榜首，22 年 12 月-23 年 11 月，月均访问量超 856 亿次、月均访问人数高达 24.4 亿。以美国为例，据 comScore，08-23 年间，谷歌旗下站点处理了全美 60%的搜索请求，处于绝对头部。

谷歌搜索广告收入居于全美第一，但在电商挑战下，份额逐年下降。分类型来看，传统综合搜索广告份额下降，如谷歌搜索收入占比由 18 年的 60%降至 23 年的 52%，或于 25 年跌破 50%。与此相比，随着零售线上化深入，并逐渐向头部平台集中，以亚马逊为代表的电商平台内搜索广告发展迅速，如亚马逊 23 年市占率 21%，较 18 年高增 11pct。

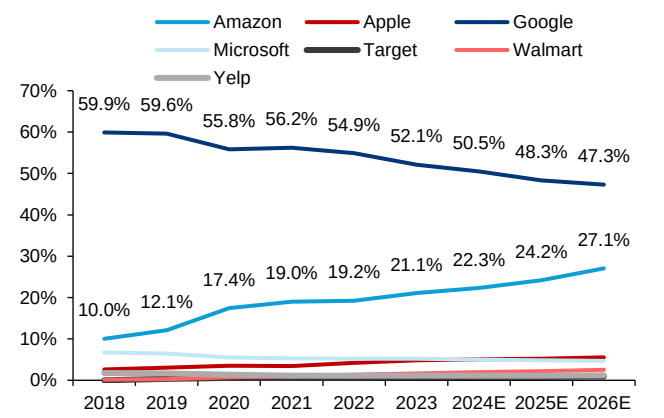
但我们认为，搜索广告仍具备明显优势，可较好捕捉高转化潜力用户，是效果广告的主流形式，吸引零售、金融等高价值广告客户。公司在 3Q 业绩会表示，2Q23 以来，搜索板块广告增长主要由零售板块驱动，尤其是 APAC 地区的零售商。亚太地区电商投放增长迅猛：例如，23 年以来，Temu、Shein 等跨境电商投放激增，二者美国广告预算中，分别有 12%和 5%花费于 PC 桌面；此外，垂直领域的广告增长更聚焦政治与金融服务业广告投放，主要得益于竞选活动和保险行业的复苏，贡献更多高价值广告客户。

图表44： 美国在线搜索请求份额：谷歌约占 60%



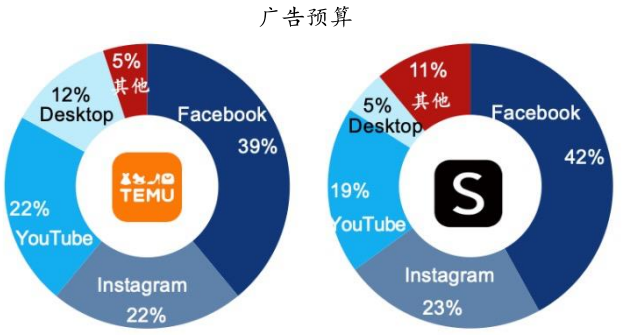
资料来源：comScore，华泰研究

图表45： 谷歌搜索广告收入份额居于全美第一，但呈逐年下降趋势



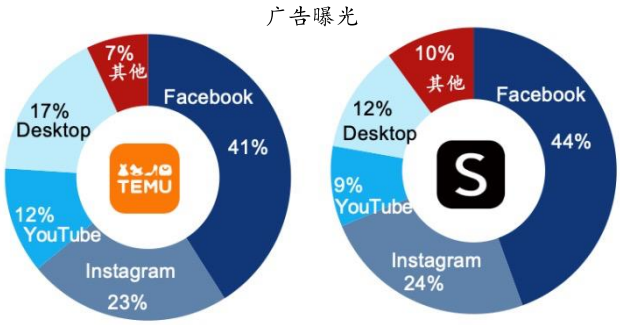
资料来源：eMarketer，华泰研究

图表46：社媒平台外，Temu 和 Shein 各有 12%和 5%在美投放预算花费于 PC 桌面



资料来源：Sensor Tower，华泰研究

图表47：PC 桌面曝光占比高于预算分配，曝光效果较好

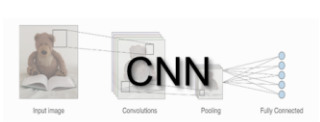


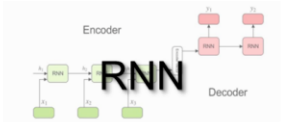
资料来源：Sensor Tower，华泰研究

谷歌作为 AI 大模型 Transformer 奠基者，具备先发技术优势

谷歌具备 AI 大模型领域的深厚技术积淀，大模型基础 Transformer 架构由谷歌 Deep Mind 团队 2017 年提出，首次亮相于论文《Attention Is All You Need》。该架构利用自注意力机制，能高效处理序列数据，学习上下文信息并生成输出。与传统的 RNN 模型相比，Transformer 能并行处理序列中的所有元素，显著提升计算效率。在 Transformer 出现之前，深度学习模型主要依赖监督学习，需要大量标注数据。而基于 Transformer 的 GPT 模型则采用自我监督学习（Self-supervised Learning）进行预训练，辅以少量监督学习进行微调。Transformer 架构在处理关联性强的任务时表现出色，适合创造性的生成任务，但仍需增强逻辑判断能力。

图表48：Transformer 架构与 CNN 和 RNN 对比情况

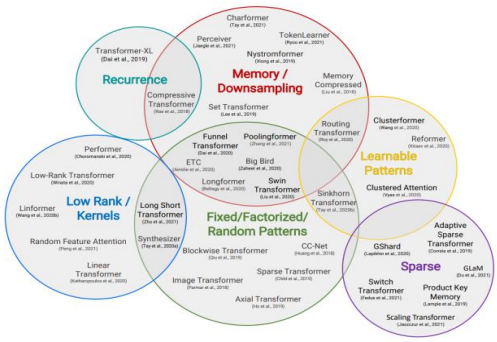




	CNN架构	RNN架构	Transformer架构
	• 局部连接、权值共享的卷积结构 • 通常包括：输入层（input layer）、卷积层（convolutional layer）、池化层（pooling layer）和输出层（全连接层+softmax layer）	• 每个时间片t的计算依赖t-1时刻的计算结果： • Encoder:$h_t = \mathrm{RNN}(x_{t-1}, h_{t-1})$ • Decoder:$y_t = \mathrm{RNN}(y_{t-1}, h'_{t-1})$ • 具有循环连接结构，如LSTM、GRU等	• 采用attention机制建模序列中位置关系 • 将序列中的任意两个位置之间的距离缩小为一个常量 • 使用位置编码代表位置信息
特征表示能力	对局部特征提取能力强，适用于图像识别、视觉任务中	适用于处理序列数据，能够捕捉序列信息的演化过程，适用于语音识别、自然语言处理任务	具有较强的建模能力，适用于处理声音序列、图像等多种数据形式
训练效率	每个卷积操作都是独立的，可并行化处理	每个时间步的计算都依赖于前一个时间步的输出，故与其他二者相比难以并行化	使用attention机制绘制输入和输出之间的全局依赖关系，可并行化处理
模型复杂度	模型结构相对简单，参数较少，不适用于处理序列数据	模型结构相对复杂，参数较多	模型结构较为复杂，参数较多

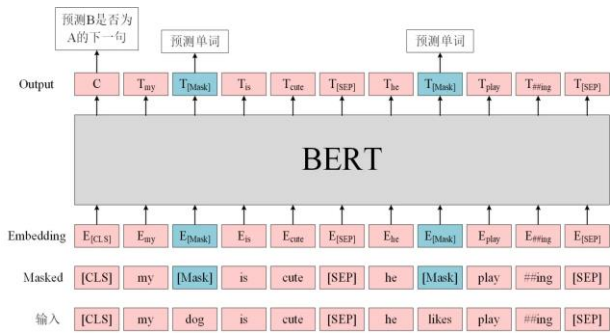
资料来源：Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).、OpenAI、KDnuggets、斯坦福大学官网、华泰研究

图表49：基于 Transformer 开发的高效模型（按技术和应用分类）



资料来源：谷歌《Efficient Transformers: A Survey》、华泰研究

图表50：BERT 模型预训练过程



资料来源：谷歌《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》、华泰研究

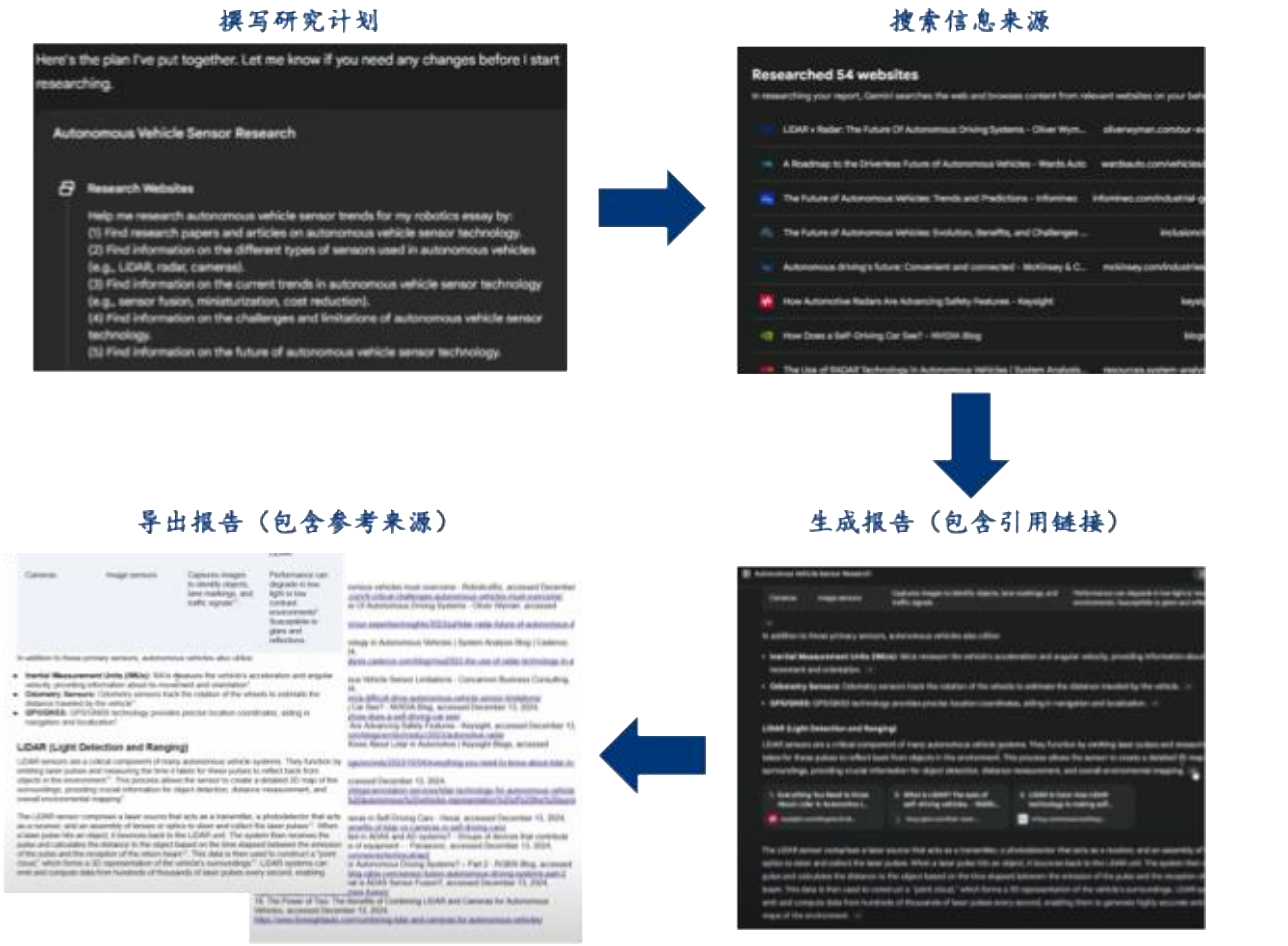
AI Overviews 基于搜索小步尝试 AI，Gemini Deep Search 试图以 AI 颠覆搜索范式

谷歌将人工智能技术整合至其搜索引擎核心业务，推出了多项创新搜索功能。谷歌已引入 AI Overviews 在原有搜索生态的基础上提高搜索体验并保持广告模式，并叠加基于 Gemini 入口的 Deep Search 提供实验性质的搜索服务。谷歌搜索已经过数次大更新，其中不乏 AI 技术的引入，以提高搜索结果的准确性和相关性：从 2013 年用于优化搜索引擎语义理解的蜂鸟算法（Hummingbird），到 2015 年用于搜索结果排序的机器学习算法 RankBrain 助力搜索引擎处理复杂查询、同义词以及提供更加个性化的结果，再到 2019 年基于 Transformer 架构的 BERT 进一步提升搜索引擎联系上下文进行语义理解的能力。2023 年 5 月，谷歌在 I/O 开发者大会宣布将生成式 AI 引入搜索引擎，并推出名为搜索生成体验(SGE)功能。此外，谷歌于同年 11 月升级谷歌商店，开辟了“AI 赋能的扩展程序”栏目，利用各种以生成式 AI 为后盾的扩展程序，从而提升浏览器性能。

AI Overviews 在谷歌搜索页面顶部提供多源信息摘要，并附带链接，并继续采用在 AI 搜索结果中嵌入广告的模式。2024 年 5 月，谷歌在 I/O 开发者大会推出同时采用深度学习+Transformer 技术的 AI Overviews 搜索引擎，其能针对复杂问题进行算法价值判断，并决定是否提供 AI 生成答案与传统链接摘要，最终展示在搜索结果顶部。截止 2024 年 10 月，AI Overviews 功能已拓展至超过 100 个国家或地区，已触达 10 亿用户。

2024 年 12 月，谷歌与 Gemini 2.0 同期发布基于 Gemini 的 Deep Search，在谷歌 Gemini Advanced 订阅服务中可用，截止 2024 年 12 月仅支持英文版。Deep Search 基于 AI Agent 原理，类似 Perplexity 的模式，首先提供一个多步骤的研究计划，用户可以进行修改和批准，然后利用 Google 搜索引擎在网络上查找相关信息，从而结合 Gemini 的高级推理和长上下文处理能力，帮助用户探索复杂的主题并撰写报告，且每条信息均有确定的来源。在生成报告后，用户可以导出到 Google Docs 继续编辑。

图表51： Gemini Deep Search 撰写报告流程

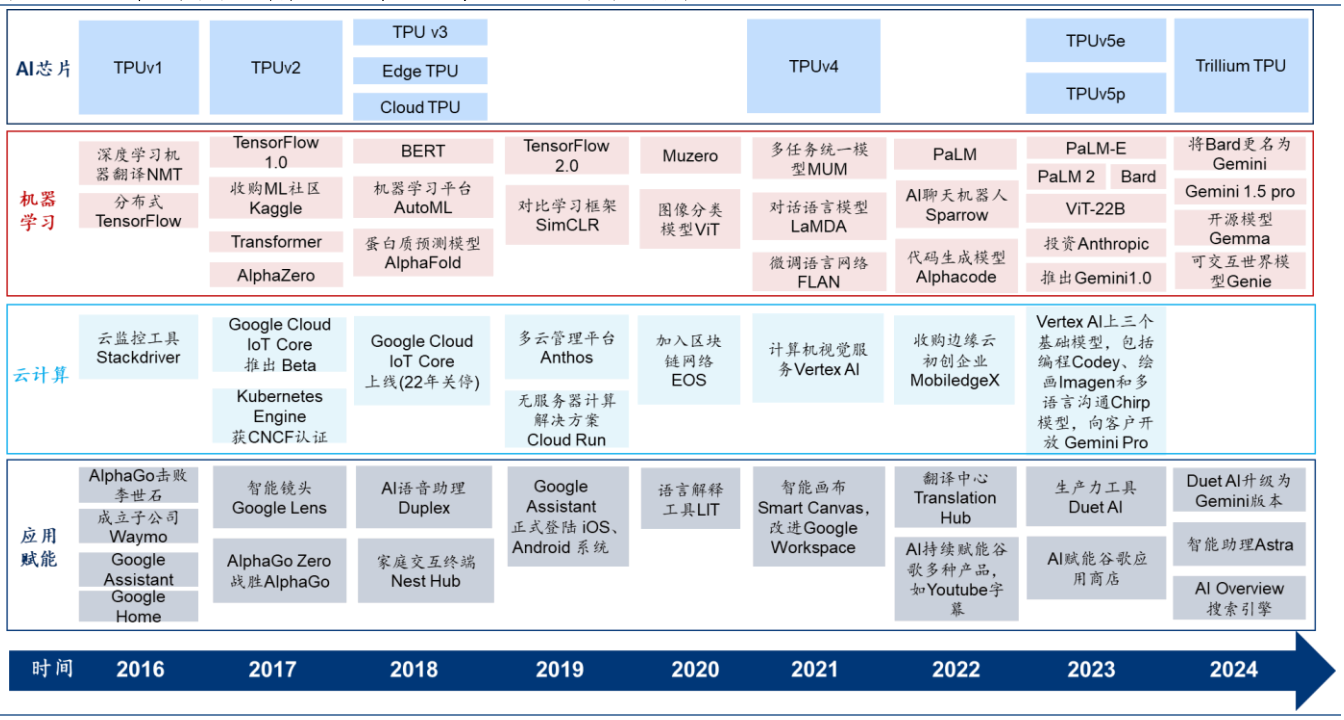


资料来源：谷歌官网、华泰研究

我们认为，AI Overviews 是在谷歌最主要的搜索入口上方提供 AI 索引，并未改变其搜索广告的盈利方式，广告主依然可以按此前谷歌的成熟算法模式投放广告，是一种面向大范围用户但小步探索 AI 的方式。与其不同的是，Gemini Advanced 所包含的 Deep Search 目前只面向 Gemini 订阅用户，是结合大模型对搜索所做的更颠覆商业模式的尝试。但我们认为基于目前的订阅模式，再打磨基于 Gemini Deep Search 的新广告投放算法，或能成为 AI 搜索时代的谷歌新商业模式。考虑到谷歌 CEO 曾表示 Gemini 将是谷歌下一个现象级应用入口，我们看好谷歌在打磨新搜索商业模式之后，凭借其 AI 技术积淀和庞大的应用生态，在 AI 时代继续保持搜索王者地位。

其他 C 端应用包括，Circle to Search 功能允许手机用户通过圈选屏幕上的对象来触发搜索，支持对图片、文本和视频内容的识别。视觉搜索，如 Google Lens，可利用模型识别手机相机所拍摄的物品并提供与之相关的内容搜索，3Q 月搜索量达 200 亿次，占购物搜索量的 20%。

图表52：2016 年至今谷歌云计算、AI 芯片、机器学习及 AI 应用赋能进程梳理



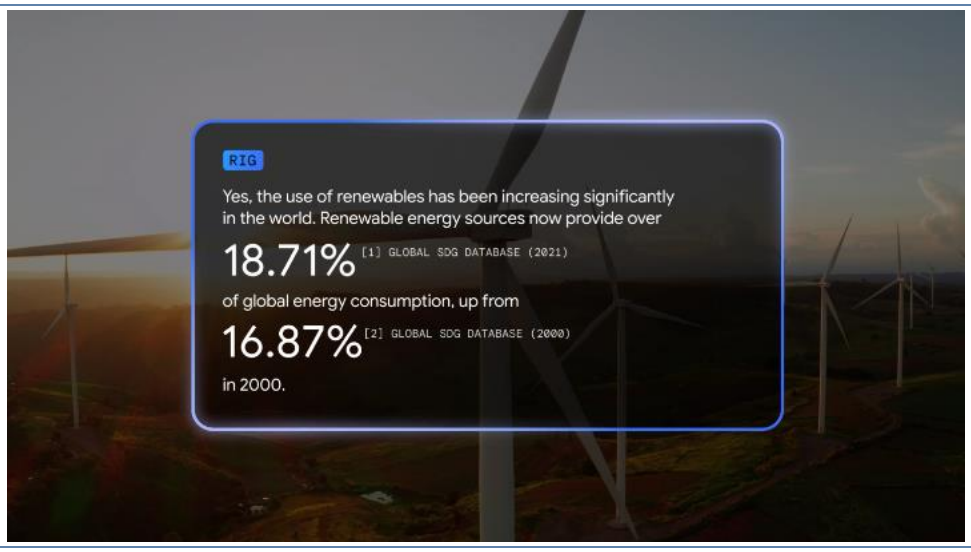
资料来源：谷歌官网、华泰研究

安全风险和幻觉问题仍为 AI 搜索落地关键

生成式 AI 风潮迭起，安全底线与大模型幻觉仍需注意。经过大规模预训练的模型、云计算与开源的融合正一齐推动生成式 AI 的全民化，但之中仍有不少道德与法律风险，包括版权保护、生成暴力、毒性、严重刻板印象、以及侵权的图像等。这些事件凸显了 AI 大模型全民化进程中 AI 信任风险和安全管理需求的迫切性和重要性。

我们认为，大模型幻觉（Hallucinations）与搜索结果的精确性要求存在冲突，这是 AI 搜索应用存在已久的矛盾。谷歌已走在行业前列清晰地认识到解决这一问题的必要性，并已有相关产品实践：1）谷歌推出的 Gemini Deep Search 通过保证每一条生成内容都具备可追溯的引用来源，保证检索的精确性，并通过强推理能力和长上下文窗口提升文字交流、内容生成与总结方面性能；2）24 年 9 月谷歌新发布解决 AI 幻觉的工具 DataGemma，能使用检索交错生成（RIG）和检索增强生成（RAG），使用 Data Commons 真实世界公开数据库帮助大型语言模型根据可靠数据核实其响应，并向用户更透明地引用事实来源，增强 LLM 事实性和推理能力。

图表53：DataGemma 使用 RIG 进行信源核查

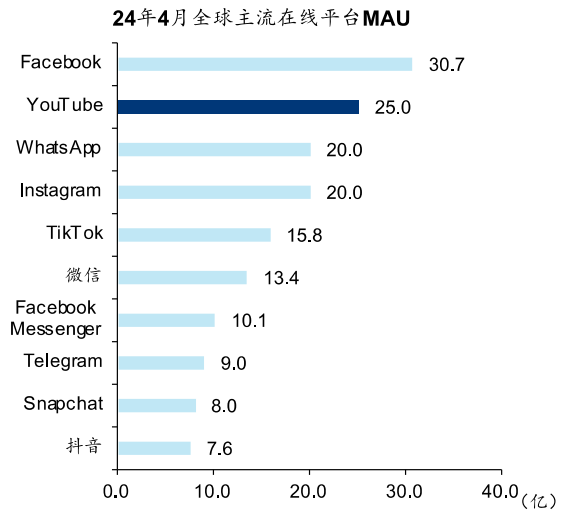


资料来源：谷歌官网、华泰研究

谷歌 YouTube 广告：在线视频推动业务转型

YouTube 是全球最大的在线视频平台。24 年 4 月 YouTube MAU 约为 25 亿，仅次于 Facebook。YouTube 拥有网页、PC 及移动端 App、电视 App 等多重入口，且视频窗口可嵌入第三方网页，无需登陆账户即自动播放，流量覆盖范围广阔。仅考虑网页端，据 SimilarWeb 统计，22 年 12 月至 23 年 11 月，YouTube 月均访问人数（14.5 亿）、访问次数排名全球第二，仅次于 Google 主站。

图表54：YouTube 月活用户数仅次于 Facebook



资料来源：公司公告，DataReportal，Kepios，华泰研究

图表55：YouTube 网页端流量充裕（22 年 12 月-23 年 11 月）

	月均访问量（亿次）	月均访客数（亿）	平均单次访问时长（分钟）
Google	856	24.4	11
YouTube	330	14.5	20
Facebook	171	11.5	10
Instagram	65	9.11	8
Twitter	64	7.64	11
百度	50	2.06	5
Wikipedia	45	7.41	4
Yahoo	34	2.94	9
Yandex	34	1.5	9
WhatsApp	29	3.6	16

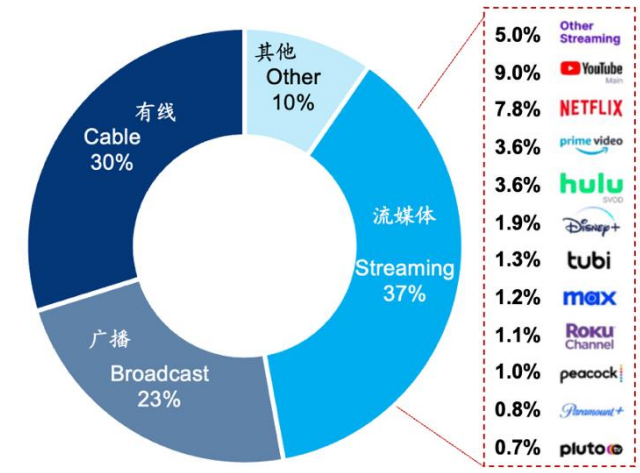
资料来源：SimilarWeb，华泰研究

驱动因素#1：把握电视媒体转型趋势，CTV 用户覆盖和观看时长增长，广告潜力扩大

美国联网电视（Connected TV，即 CTV）普及率迅速增长，广告支出不断上升。eMarketer 预计，全美 CTV 普及率将在 24 年内增至 68%，较 18 年上升 11pct；其中 25-34、35-44 岁人群普及率高达 73%、76%。CTV 广告总支出 23 年同比增长 18%，达 242 亿美元，占数字广告大盘 9%。

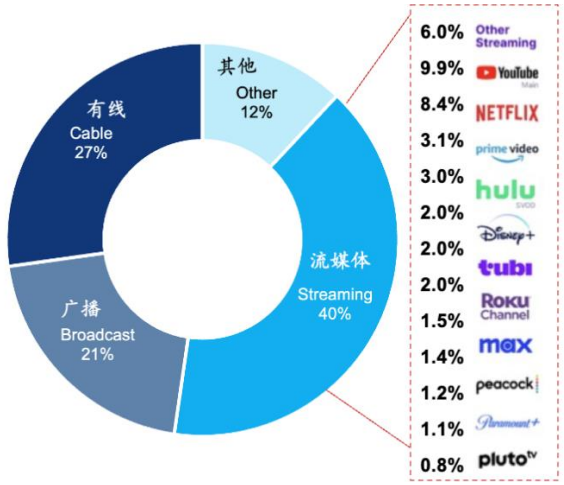
据 Nielsen 数据，23 年 2 月以来，YouTube 连续 17 个月位居全美流媒体电视观看时长之首；而若考虑到传统电视媒体内容，YouTube 市占率位居第二位，仅次于迪士尼。有线和广播等传统内容时长占比持续下滑，而流媒体占比自 37%（3Q23）提升至 40%（2Q24）。其中 YouTube 是市占率最大的流媒体平台，时长占比上升 0.9pct 至 9.9%。

图表56： 美国电视使用时长占比（3Q23）



资料来源：公司公告，Nielsen，华泰研究

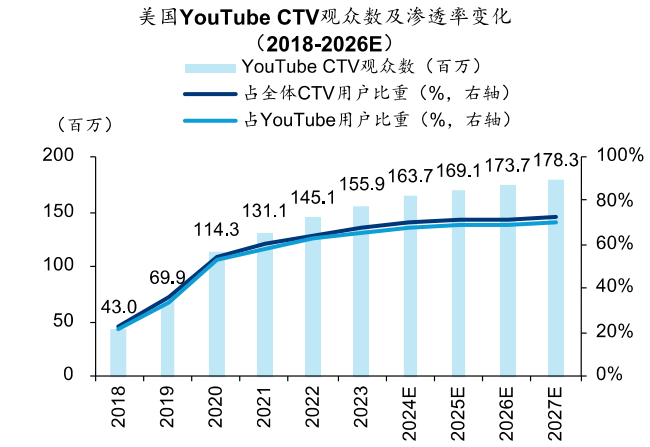
图表57： 美国电视使用时长占比（2Q24）



资料来源：公司公告，Nielsen，华泰研究

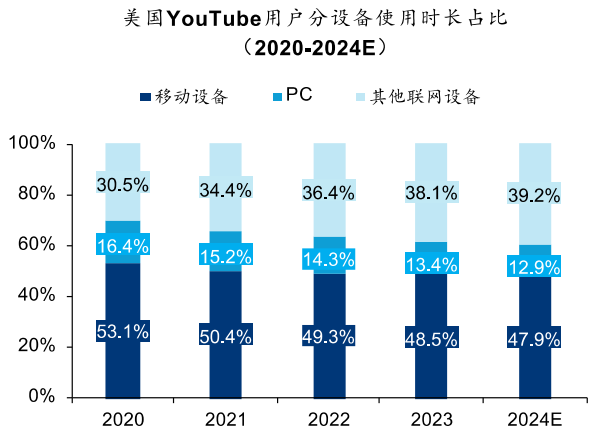
YouTube CTV 覆盖用户数、观看时长占比逐年上升。从观众规模看，23 年 CTV 观众数突破 1.5 亿，18-23 CAGR 29.4%；覆盖 66%的 YouTube 用户，较 18 年上升 44pct。预计 24 年 YouTube CTV 整体市占率超 70%。从观看时长看，CTV 观看时长占比持续增长，预计 2024 年将占 YouTube 总观看时长的 39.2%。从广告收入看，YouTube CTV 广告收入逐年稳步抬升，预计 2024 年可达 33 亿美元，同比抬升 14%。

图表58： YouTube CTV 观众数及渗透率稳步提升



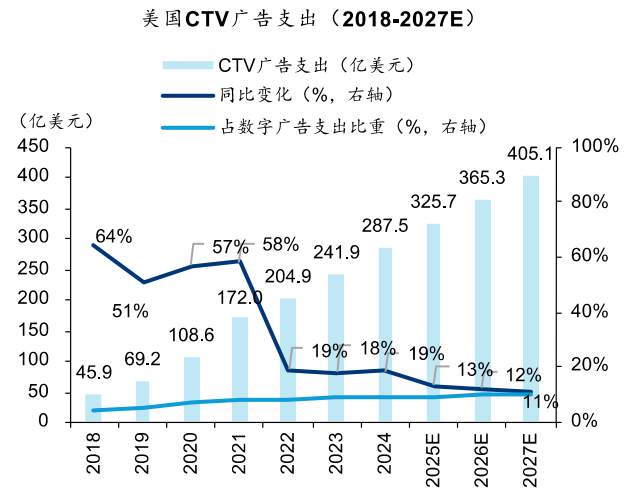
资料来源：eMarketer，华泰研究

图表59： 联网设备（以 CTV 为主）占用户总时长比重逐年扩大



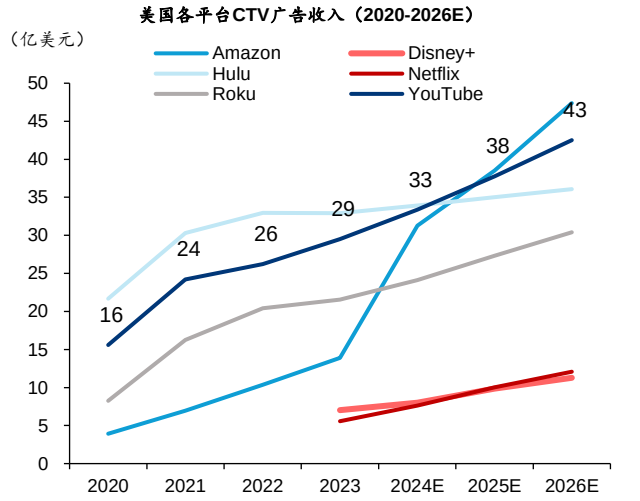
资料来源：eMarketer，华泰研究

图表60：美国 CTV 广告支出高速增长，23 年占数字广告大盘 9%（较 18 年上涨 5pct）



资料来源：eMarketer，华泰研究

图表61：YouTube CTV 广告收入不断增长，2023 年位居全美第二



资料来源：eMarketer，华泰研究

大屏内容质量更高、广告体验更好，成为 YouTube 广告增长新引擎。YouTube 在美约四成收入来自 CTV 端（eMarketer）。我们认为，与移动端和 PC 端相比，YouTube CTV 广告有以下优势：

1） **流量规格更高：**大屏观看视频通常制作较为精良，有利于提升品牌形象，吸引头部广告主预算。目前 75% YouTube Select 精选广告曝光来自 CTV 端。2024 年 5 月，YouTube 在其年度广告推介会 Brandcast 上正式推出精选买断计划，允许广告主买断前 1%的头部流量；我们预计该部分流量将大多在大屏上兑现。

2） **广告体验更好：**2023 年，YouTube CTV 端推出了暂停广告和 30 秒不可跳过广告。暂停广告仅在用户暂停播放时出现，干扰小，单价高。30 秒不可跳过广告虽然时长较长，但频率低，更符合具有明确受众广告主。

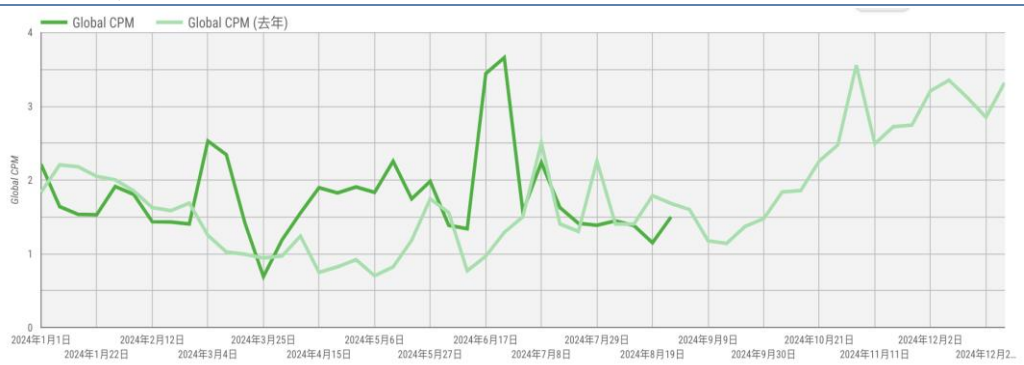
此外，YouTube CTV 广告相对传统电视广告的优势，也有助于承接广告主预算转移：1）效果导向：CTV 与移动、PC 端账号互通，能够综合用户多端行为信息，投放效果更精准、可追踪。**2）用户友好：**时长短于电视插播广告，体验更好；**3）互动性强，转化链路短：**如通过附加二维码，将电视推送无缝传输至手机端浏览，方便后续转化。

驱动因素#2：YouTube 广告营收增速较快，单价仍有提升空间

基于活跃的用户群体和丰富的视频内容，YouTube 广告库存充裕，4Q23-3Q24 广告和订阅收入超过 500 亿；短视频变现率改善，Shorts 商业价值有较大增长空间。YouTube 广告分为品牌（Brand）和效果（Direct Response）两类，前者侧重于通过高规格流量曝光，扩大品牌声量；后者更强调效果的可追踪性，意在直接实现转化目标（如 App 下载、线索收集和商品购买）。

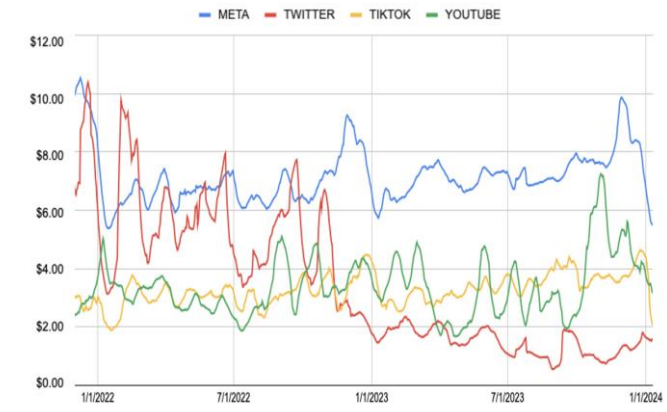
与其他主流社媒平台对比，YouTube 广告 CPM 较低。据 Gupta Media CPM Tracker，24 年以来 YouTube 全球平均 CPM 约为 1.67 美元，+24.7% yoy，低于 Meta 的 Facebook、Instagram 广告均价（5.83 美元）和 TikTok（3.70 美元）；美国平均 CPM 约为 2.98 美元，-2.1% yoy，与 Meta 差距较大（7.02 美元）。根据 eMarketer 预估，24 全年，美国成年人单日数字媒体使用时长中，YouTube 和 Meta 占比相似，但吸引的广告支出只有后者的四分之一，广告收入仍有较大提升空间。

图表62：24 年 2 月以来 YouTube 全球 CPM 同比略增



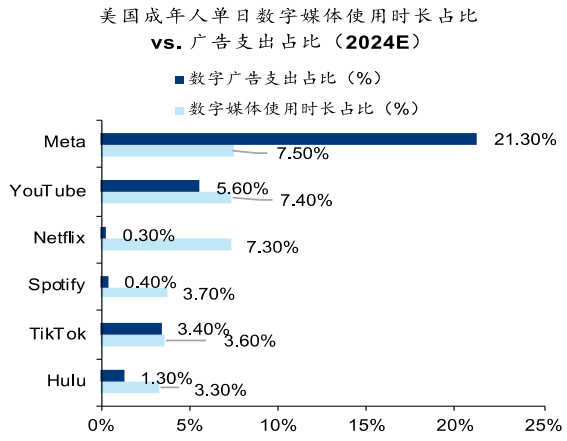
资料来源：Gupta Media，华泰研究

图表63：2022-2023 全球主要社媒平台 CPM 走势



资料来源：Gupta Media，华泰研究

图表64：YouTube 广告价格较低，变现能力不及 Meta



注：Meta 包括 Facebook 和 Instagram 平台；YouTube 使用时长包含电视端
资料来源：eMarketer，华泰研究

YouTube 的消费和货币化趋势呈现出积极的发展态势。公司 3Q 表示，Shorts 的观看时间在 YouTube 上不断攀升，每天有超 700 亿条短视频被观看，显示出用户对短视频内容的强烈需求。在货币化方面，短视频的货币化率与长视频相比呈现出健康增长态势，同时，短视频已被纳入 YouTube Select 视频广告活动中，为品牌提供了更精准的目标选择。此外，谷歌还推出 Shorts 的 SEO 优化工具，强调标题、描述和标签对视频可见性的重要性，以及分析工具，通过分析关键数据如平均观看时长和参与率，助力创作者优化内容策略，把握趋势。

驱动因素#3：YouTube 强化短视频战略以对抗 Netflix、TikTok 和 Reels 的竞争

面对 Netflix 等流媒体平台和 TikTok、Instagram Reels 等短视频媒体的冲击，YouTube 通过发力短视频、加大创作者支持力度和完善版权储备，巩固用户粘性：

1) **发力短视频：**YouTube 于 21 年推出总金额 1 亿美元的短视频创作基金，每个频道每月最高获得 1 万美元奖励 。**用户端，**23 年 7 月，公司表示，每月超 20 亿 YouTube 登录用户会观看 Shorts，较 22 年增长 5 亿。**创作者端，**1Q24 业绩会上管理层表示，23 全年发布 Shorts 的频道数量同比高增 50%。**而商业化方面，**与 TikTok 和 Instagram Reels 相比，目前 Shorts 仍处于商业化早期，管理层并未对其变现效果进行较多点评。

2) **赋能长尾创作者，提升平台粘性：**23 年 6 月以来，YouTube 陆续在美国、英国、加拿大等 99 个国家和地区扩展其伙伴计划（YPP）覆盖范围，降低准入门槛。新计划下，创作者只需满足 500 粉丝+过去 1 年视频播放时长超 3000 小时（或 90 天内短视频浏览量超 300 万）的基础要求，就可进行产品推广，并开启粉丝变现功能（如频道订阅）。截至 1Q24，参与 YPP 计划的频道数已达 300 万，3 年累计激励金额超 700 亿美元。

3) **丰富影视、体育、音乐版权储备，扩充专业内容库：**影音方面，10 年代以来，YouTube 相继与索尼、华纳、环球等主流厂牌签订内容合约，以规避版权风险。面对 AIGC 的发展，据金融时报，6 月以来，YouTube 积极和各大唱片公司开展谈判，意在取得更多艺术家的版权许可，训练其 AI 音乐生成技术。体育方面，22 年 12 月，YouTube 战胜苹果、亚马逊，取得了 7 年期 NFL（美国橄榄球联盟）周日赛转播权，打破了 DirecTV 近 30 年的垄断；转播服务以订阅制提供，每赛季费用为 299-439 美元。NFL 是美国国民级赛事，24 年冠军赛（即“超级碗”）收视人次超 1.23 亿（Nielsen），转播权的取得将进一步完善体育内容布局，并提升平台广告价值。

图表65：YouTube 伙伴计划（YPP）准入门槛 23 年以来大幅下降，拓展长尾频道变现空间

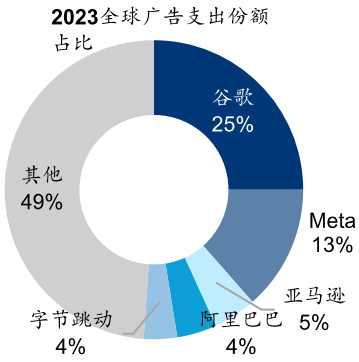
准入要求	新政策	原政策
粉丝数	**基础：**500 以上 **升级：**1000 以上	1000 以上
内容产量	过去 90 天内曾上传≥3 个视频	-
内容影响力	**基础：**过去一年视频观看总时长超 3000 小时，或过去 90 天内短视频浏览量超 300 万 **升级：**过去一年视频观看总时长超 4000 小时，或过去 90 天内短视频浏览量超 1000 万	过去一年视频观看总时长超 4000 小时，或过去或过去 90 天内短视频浏览量超 1000 万
激励形式	**基础：**可进行产品推广；可开启粉丝变现功能，如频道付费会员、超级留言和超级贴纸 **升级：**广告及 YouTube Premium 订阅收入共享	广告收入共享，粉丝变现，产品推广等

资料来源：公司官网，华泰研究

行业趋势#1：生成式问答或有幻觉，AI 搜索商业落地仍未明朗

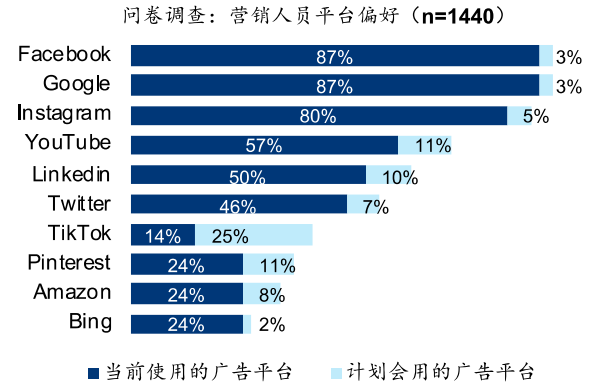
广告市场来看，虽然长短视频和零售电商广告增长强劲，谷歌和 Meta 仍旧占据主导地位。随着社交媒体和短视频广告增长，以及电商内搜索广告的发展，谷歌基本盘搜索广告份额下降，对其市占率产生不利影响；而 YouTube 视频不断丰富内容类型，仍有较大增长潜力。

图表66：全球线上+线下广告份额：五家科技公司占据 50%以上



资料来源：WARC，华泰研究

图表67：Meta 和 Google 旗下平台是广告主数字营销首选

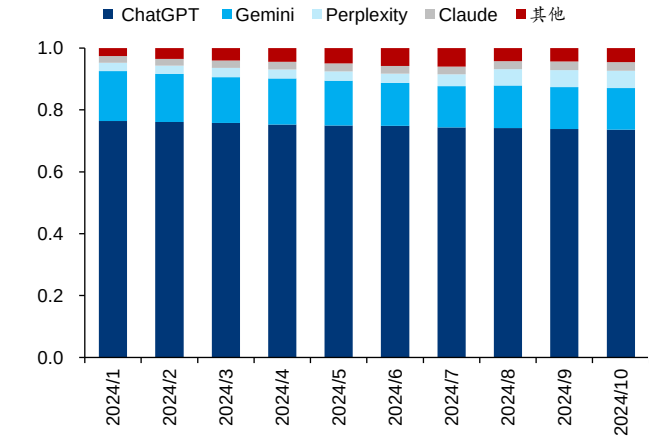


资料来源：Instapage，华泰研究

全球搜索市场份额仍为谷歌搜索主导，ChatGPT 占据生成式聊天机器人七成市场份额。Statcounter 数据显示，截至 24 年 9 月，全球互联网搜索引擎市场份额中，谷歌渗透率达 88%，同比下滑 0.5pct；雅虎和必应的市场份额分别为 2.5%和 7.0%；而集聊天机器人和搜索于一身的 Perplexity 市场份额抬升明显，10 月占比 5.6%，对比年初市占率 2.7%。聊天机器人方面，ChatGPT 仍占据主要市场份额，10 月占比达 74%，但市占率较年初有所下滑。

我们认为以 ChatGPT 为代表的对话机器人直接赋能的搜索将不会取代谷歌搜索，相反类似 Perplexity 这种先搜索后整理的搜索引擎，或将成为谷歌劲敌。2023 年 2 月 OpenAI 携微软 Bing 先下一城，谷歌虽于 2023 年 3 月发布 Bard，但其应对速度和回答表现欠佳，市场认为谷歌的搜索业务将受到冲击，同时对谷歌能否在未来维持人工智能的领头羊地位产生质疑。相比围绕 ChatGPT 重新设计界面的竞争对手 New Bing，谷歌搜索引擎的 AI 化升级似乎更为谨慎，公司始终将 Gemini 定位为生成式 AI 赋能搜索的补充而非替代品。我们认为，ChatGPT 较难改变当前全球搜索引擎市场格局，但新兴初创 AI 企业如 Perplexity，则有机会挑战谷歌搜索龙头的地位。

图表68：生成式 AI 聊天机器人市场分额占比（单位：%）



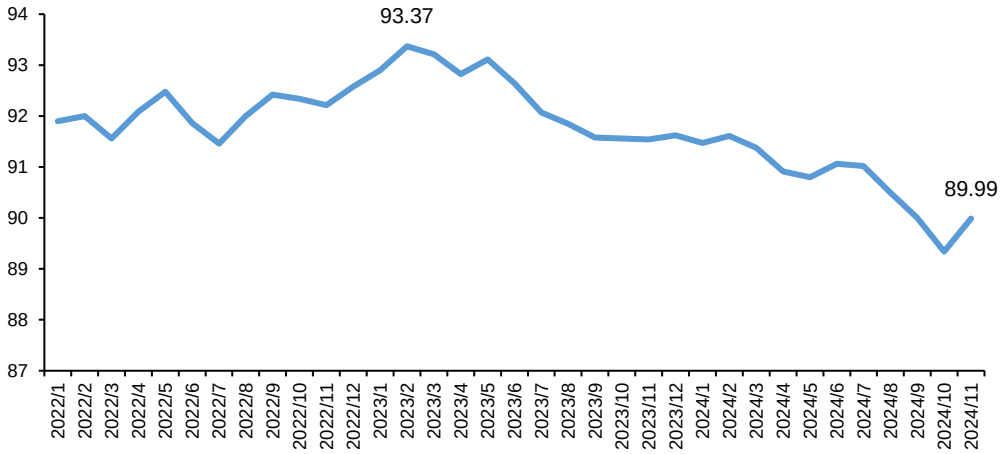
资料来源：FirstPageSage，华泰研究

图表69：主要生成式 AI 聊天机器人对比

	Gemini	Grok	Perplexity
平台	网页，安卓	网页	网页，iOS，安卓
目标	建设人工智能生态系统，并提供多种工具和套件	增强应用中对话能力	综合信息检索
重要能力	广泛分析功能	上下文保留能力	多源数据查询
	与多个数据库、分析平台无缝连接	与 Slack、Discord 等通信平台连接	高效处理用户查询歧义
	能够实时处理和分析数据流	用户反馈将纳入迭代循环	
应用环境	密集数据处理	自然互动	全面数据分析
集成	安卓系统应用	X 社交平台	

资料来源：Graph，华泰研究

图表70：2022 年至今谷歌搜索引擎市场份额变化（%）



资料来源：Statcounter 官网、华泰研究

我们认为，AI 商业化落地是胜负未定的万里长征，谷歌搜索业务第一梯队的地位难以一朝被颠覆：1）从竞争动态来看，随着新兴对手 Perplexity 的崛起和苹果与 OpenAI 的端侧合作，谷歌在短时间内或承受一定的竞争压力和份额的蚕食。2）从长期来看，谷歌具备广泛的生态基础，本身是大模型 Transformer 的奠基者，具备深厚的技术基础，或有希望解决大模型搜索的精确性要求与幻觉的冲突问题，从而在商业化角度获得长远收益。

但我们发现公司确实存在一些值得关注的短板。最典型的例子是，新兴的 Perplexity 等搜索引擎已能将大语言模型与搜索功能结合，而谷歌仍在探索这种模式，我们认为这种滞后并非源于技术能力的不足，而是商业化节奏和产品创新速度有待提升。

谷歌拥有强大的AI研发实力和海量的数据优势，但如何将这些优势快速转化为产品竞争力，是公司当前面临的最大挑战。目前谷歌 CEO 原为产品经理，CTO 经验更多在信息检索，我们认为，若观察到谷歌内部既为 AI 专家并同时熟悉产品的高管，在人事任命上得到重视并引导公司转型战略，或为公司整合 AI 与主营业务的积极信号。

Perplexity 致力于把聊天机器人和搜索完美结合，打造没有幻觉的“答案发现”引擎，会否成为谷歌的劲敌？

Perplexity 在 2022 年由前 OpenAI 研究科学家 Aravind Srinivas，联合前 Databricks 创始成员之一 Andy Konwinski、前 MetaAI 研究科学家 Denis Yarats 和前 Quora 工程师和量化交易员 Johnny Ho 共同创办，可见大部分始创人员均具备深厚 AI 研究和应用背景。目前，Perplexity 在 2024 年 11 月的最新一轮融资中预计筹资约 5 亿美元，估值达约 90 亿美元。

Perplexity 创立的目的在于通过将搜索和大语言模型（聊天机器人）结合起来，改变互联网上获取搜索问题答案的方式，通过确保答案的每个部分都带有引用来源，最大限度地减少了大语言模型中常见的“幻觉问题”。Perplexity 的搜索结果主要是基于真实网络搜索结果，而不是一般大模型（聊天机器人）所提供的自我思维后的最佳猜测答案。Perplexity 同时会提供引用的网站来源，因此，它是基于类似 RAG（检索增强生成）的方式，在给定查询的情况下，检索内容并提取相关段落，然后输出此信息以及来源。但有别于 RAG，它不是将搜索到的相关信息用作上下文内容发送给大语言模型以生成答案，而只是使用大模型调整答案而不是生成答案，即令大模型阅读链接，提取相关段落，并给出格式工整的答案，为每一个观点提供引用来源。此外，在前一个问题结束后，大模型也会提供其他相关问题供用户继续探索。

图表71：Perplexity 融资历程与管理团队

融资时间	融资轮次	主要投资方	融资金额（百万美元）	估值（百万美元）
2022.09	种子轮	Elad Gil	3.1	-
2023.03	A轮	New Enterprise Associates	25.6	121
2024.01	B轮	Institutional Venture Partners, New Enterprise Associates, NVIDIA, Kindred Ventures	73.6	540
2024.04	B轮	Daniel Gross, Institutional Venture Partners, New Enterprise Associates, NVIDIA	62.7	~1000
2024.11	-	Institutional Venture Partners	500.0	~9000

团队介绍

Aravind Srinivas
Co-founder & CEO

- 2021-2022在OpenAI担任研究科学家
- 2019-2021在Deepmind和谷歌实习
- 博士毕业于加州大学伯克利分校

Berkeley
Google OpenAI

Andy Konwinski
Co-founder & President

- 2013年开始作为Databricks的联合创始人
- 博士毕业于加州大学伯克利分校

Berkeley
databricks

Denis Yarats
Co-founder & CTO

- 2016-2022在Facebook担任AI研究科学家
- 2013-2016在Quora担任ML工程师
- 2011-2013在微软担任软件工程师
- 博士毕业于纽约大学

NYU
Quora Microsoft

Johnny Ho
Co-founder & CSO

- 2017-2022在Tower担任量化交易员
- 2013-2014在Quora工作
- 毕业于哈佛大学

Tower RESEARCH CAPITAL
Quora

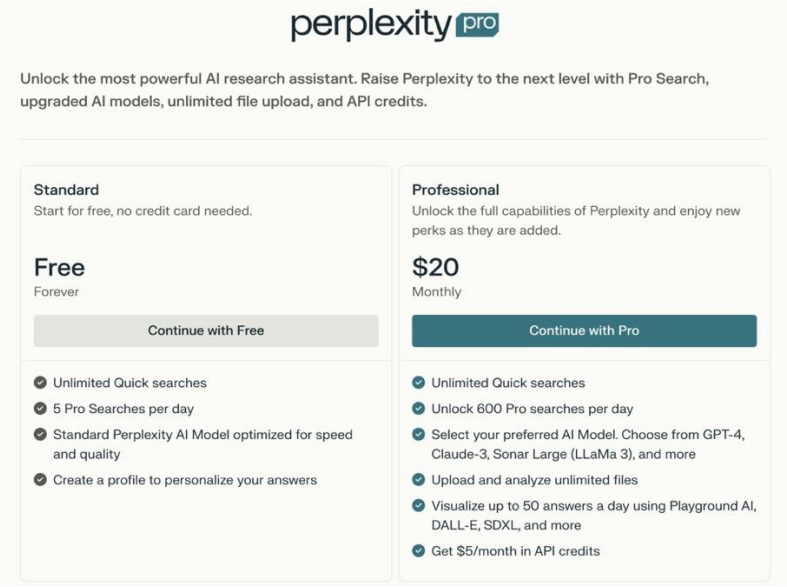
资料来源：Perplexity 官网，LinkedIn，Techcrunch，华泰研究

Perplexity 在商业模式方面基本采用免费增值和订阅结构，免费产品使用基于 GPT-3.5 的产品功能，而付费的 Perplexity Pro 可访问 GPT-4、Claude 3.5、Grok-2、Llama 3 和 Perplexity 自研的大模型。2024 年第一季度，其月活跃用户已达 1500 万。Perplexity 在 2024 年 5 月记录了约 7300 万次访问，每月的增长率约为 20%。ChatGPT 在 2024 年 10 月记录访问量约为 36 亿次，大约是 Perplexity 的 50 倍。

图表72： Perplexity 的搜索答案以及相关问题呈现

资料来源：Perplexity 官网、华泰研究

图表73： Perplexity 免费增值和订阅结合的商业模式



资料来源：Perplexity 官网、华泰研究

在商业模式上正在转向广告和电商使得 Perplexity 成为 Google 搜索的更直接竞争对手。Perplexity 表示将通过相关问题功能引入广告，允许广告商通过付费，在后续相关问题中展示其内容。Perplexity 于 24 年 11 月启用电子商务 AI 推荐，即在 Perplexity 的搜索结果中提供购物推荐，并且用户无需访问零售网站即可下订单。此举目的为与谷歌和亚马逊展开竞争，争夺购物搜索的份额。在推出 AI 购物搜索工具的同时，Perplexity 还推出了卖家计划。如果卖家加入，其产品将有更大的机会获得推荐。**另外，Perplexity 也有 API 和企业解决方案的盈利模式。**公司提供 API 访问其搜索功能，允许企业和开发人员将 Perplexity 集成到应用程序中，预计该服务将根据使用情况定价。Perplexity 也为企业客户开发定制的解决方案，可能涉及针对行业需求的定制人工智能环境和集成，进一步多样化其收入，但目前尚未从用户购买中抽取佣金。

不过，Perplexity 也面临新闻媒体内容方的指控，或进一步加大其商业模式的成本。Perplexity 于 24 年 6 月份被 Forbes 和 Wired 指控抄袭，这两家媒体批评 Perplexity 在没有明确注明来源的情况下使用其内容，并抓取了明确屏蔽爬虫的网站。Perplexity 首席商务官 Dmitry Shevelenko 承认了这些指控，并表示公司已经接受批评。Perplexity 随后对其用户界面进行了修改，使引用更加突出。2024 年 7 月，Perplexity 宣布启动一项新的出版商计划，与合作伙伴分享广告收入，每篇赞助文章 Perplexity 都会与媒体分享“两位数百分比”收入，目前已与 Time、Der Spiegel 和 Fortune 等多家媒体签署了协议。推出后的两周内，有 50 家媒体要求加入 Perplexity 的收入分成计划，该收入分成即主要来自于广告模式。

就目前而言，谷歌完全采用类似 Perplexity 的 AI 搜索模式的意向并不算强烈，原因也许是意味着推翻以往的搜索商业模式。考虑到公司的历史文化，我们认为谷歌并非做不了 Perplexity 的模式，而是需要打磨完整的商业模式才能发布产品，正如同云时代即使谷歌拥有当时最优秀的分布式工程师但仍然被亚马逊抢占先机，主要系云利润率低于广告利润率，所以谷歌没有动机。AI 时代，Perplexity 所定义的搜索逻辑里，信息排序的流程被省略，直接给出答案，公司 CEO 在采访中提到，Perplexity 早期打造产品时存在是否继续保留答案相关链接的争议，因为 AI 搜索得到的答案仍然可能产生幻觉。但团队决定以一种更激进的立场直接提供答案，而不是像谷歌等传统搜索一样继续保留链接排序。这种选择的背后是 Perplexity 押注于 AI 大模型会变得更智能、更便宜、更高效，而幻觉会在未来大幅下降。对谷歌来说，收入来自让大量用户点击和浏览链接，赚取广告费，同时打造数据飞轮，优化搜索排名。

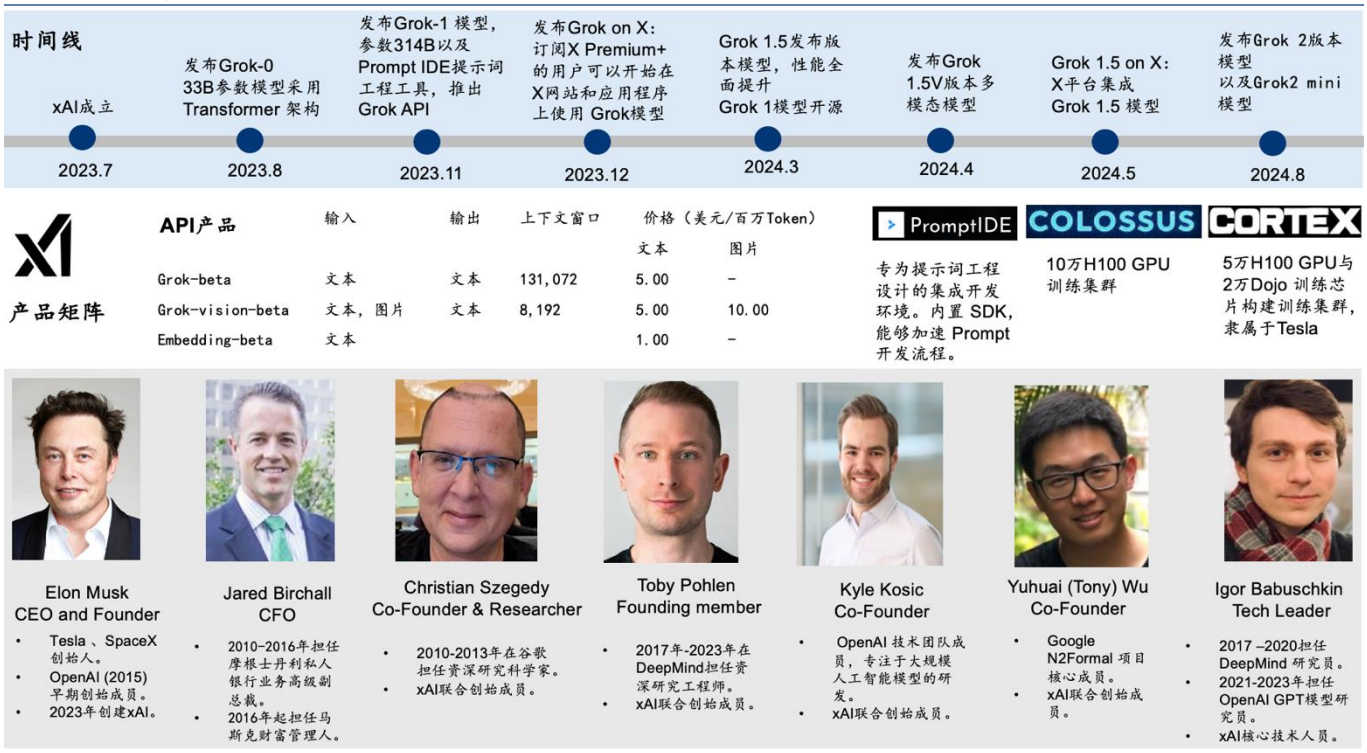
Grok 后发先至，即时更新知识库完善布局智能问答与搜索

2023 年 7 月 xAI 公司成立，并于同年 11 月发布首款 AI 大模型产品 Grok-1。公司随后陆续发布 Grok 系列模型：Grok-1.5、多模态大模型 Grok-1.5V、Grok-2 以及 Grok-2 mini，并拓展模型至 X（原 Twitter）平台供订阅用户使用。24 年 9 月，xAI 宣布建成 Colossus 计算集群，以 10 万块 H100 GPU、EB 级存储和超高速网络为支撑，旨在推动 Grok 后续迭代。

与市场上主流 AI 模型相比，如 OpenAI 的 GPT 系列或 Anthropic 的 Claude 系列，Grok 的关键技术突破包括：1）实时学习能力：与传统语言模型基于静态数据训练不同，Grok 引入基于 Twitter 实时数据的学习机制，能不断更新知识库。2）情感共鸣与个性化定制：xAi 团队注入 Grok 情感分析和动态调整模块，使其能根据用户需求调整语气和表达方式。

2024 年 3 月，Grok-1 大模型正式开源，该模型采用 MOE 架构，具有 314B 参数，对比同期开源产品 Llama 2（70B）和 ChatGPT 3.5 Turbo（70B），体积更大。我们认为 Grok 开源主系：1）大模型竞争激烈，开源策略有效促进技术的快速迭代，占领更多用户硬件平台，提高产品生命力；2）开源模型有望吸引更多用户参与，从而构建起应用生态系统，有效促进商业模式的自我增强和快速成长。

图表74： xAI 发展历程与管理团队



资料来源：xAI 官网，机器之心公众号、新智元公众号，华泰研究

图表75： 主要 ChatBot 订阅价格

产品	订阅价格（美元/月）	订阅价格（美元/年）
Grok Premium	16	168
ChatGPT Plus	20	
ChatGPT Team	30	300
Claude Pro	18	216
Claude Team	30	300
Perplexity Pro	20	
Gemini Advanced	20	

资料来源：各公司官网，华泰研究

图表76： 主要模型 API 调用价格

模型名称	输入（美元/百万 token）	输出（美元/百万 token）
Grok-beta	5	15
Grok-v-beta	10	
GPT-o1 Preview	15	60
GPT-4o mini	0.15	0.6
GPT-4o	2.5	10
Gemini 1.5 Pro -128k	2.5	10
Claude 3.5 Sonnet	3.75	15

资料来源：各公司官网，华泰研究

Grok-2 已实现对主流大模型的追赶，性能位居第一梯队。Grok-2 在多项基准测试中表现出色，包括推理、阅读理解、数学、科学和编码等领域，相较 Grok-1.5，Grok-2 和 Grok-2 mini 性能均有提升。在研究生水平的科学知识（GPQA）、常识（MMLU）和数学竞赛问题（MATH）等测试中，Grok-2 表现已达到行业前沿水平。在视觉数学推理（MathVista）和基于文档问答（DocVQA）方面，Grok-2 实现 SOTA。

图表77： 主流大模型对比，Grok-2 实现对主流大模型的追赶

	Grok 1.5	Grok 2 mini	Grok 2	GPT 4 Turbo	GPT 4o	Gemini Pro 1.5	Llama 3 405B	Claude 3.5 Sonnet
公司	xAI	xAI	xAI	OpenAI	OpenAI	Google	Meta	Anthropic
推出时间	2024 年 3 月	2024 年 8 月	2024 年 8 月	2023 年 11 月	2024 年 5 月	2024 年 4 月	2024 年 4 月	2024 年 3 月
GPQA	35.9	51.0	56.0	48.0	53.6	46.2	51.1	59.6
MMLU	81.3	86.2	87.5	86.5	88.7	85.9	88.6	88.3
MATH	50.6	73.0	76.1	72.6	76.6	67.7	73.3	76.1
HumanEval	74.1	85.7	88.4	87.1	90.2	71.9	89.0	92.0
MMMU	53.6	63.2	66.1	63.1	69.1	62.2	64.5	68.3
MathVista	52.8	68.1	69.0	58.1	63.8	63.9		67.7
DocVQA	85.6	93.2	93.6	87.2	92.8	93.1	92.2	92.2

资料来源：公司官网，机器之心公众号，华泰研究

我们认为，在本轮 AI 驱动搜索应用革新中，用户生态与智能搜索的结合或将对谷歌搜索的主导地位构成挑战。Grok 通过将模型能力集成到 X 社交平台（原 Twitter）中，有望为 6 亿平台月活用户推广多项应用内功能，包括文本搜索，评论摹写与优化等。X 平台也在其“探索”页面推出基于 Grok 的全新功能——“Stories on X”。该功能旨在通过 AI 生成技术为用户提供平台上热门新闻及话题的故事摘要，帮用户快速概览近日头条内容以及用户感兴趣的内容。此外，Grok 将追踪热点新闻报道和引发广泛讨论的公共议题，生成重点摘要，并在平台上进行推广。

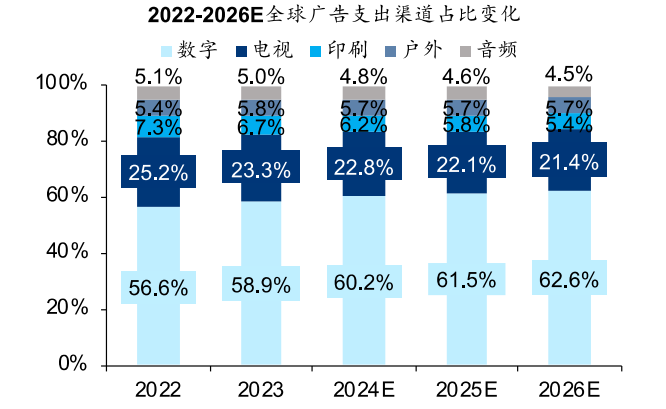
行业趋势#2：短视频电商驱动增长，AI 营销应用仍有深化空间

线上广告逐步取代传统媒体：从 23 年全球广告总支出看，电视占比仍偏高，为 23.3%。我们预测线上广告渗透率保持扩张，其支出占比有望从 23 年 68.0%抬升至 26 年 74.0%，主要由社交媒体、长短视频和零售电商等驱动。

电商和短视频广告渗透率仍有较大空间：目前海外线上零售渗透率仍较低（如 23 年美国仅为 15.4%，远低于中国的 27.6%），广告仍有较高增长空间。此外，19-23 年期间，随着商业化进程加速，美国视频广告份额攀升 5.8pct 至 23.2%，广告主逐渐由图文转向更具互动性的视频内容。

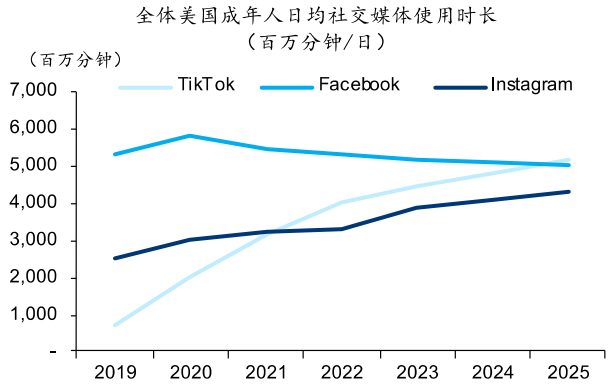
线上+线下媒体总时长步入存量竞争：美国成年人媒体总使用时长见顶，预期 23-25 年 CAGR 仅为 0.3%。总量稳定之下，**电视观看时长持续下降：**电视触达范围收缩，且投放效果较难监测，广告主进而向线上渠道转型。据 eMarketer，24 年美国成年人平均每天电视观看时长将降至 175 分钟，对应 21-24E CAGR 为-3.7%；线上视频观看时长则升至 230 分钟，对应 CAGR 达 4.8%，主要涉及 YouTube、Netflix 等平台。

图表78：全球线上广告逐步取代传统媒体



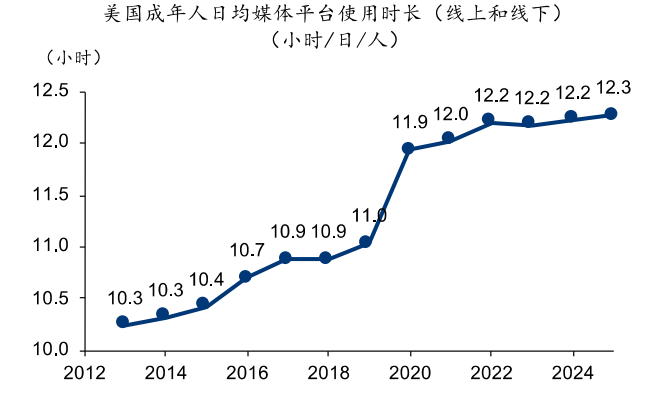
资料来源：Dentsu，华泰研究

图表79：短视频内容仍在创造时长增量



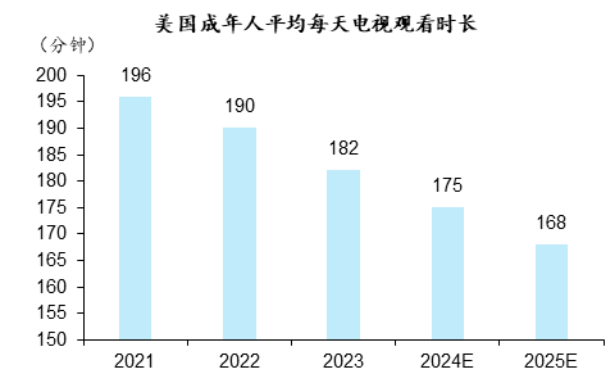
资料来源：eMarketer，华泰研究

图表80：媒体总时长步入存量阶段，网络逐步取代线下



资料来源：eMarketer，华泰研究

图表81：美国成年人平均每天电视观看时长逐年下降

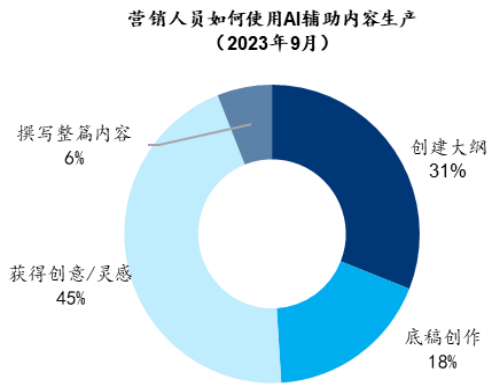


资料来源：eMarketer，华泰研究

我们认为海外数字营销行业面临以下两大趋势：

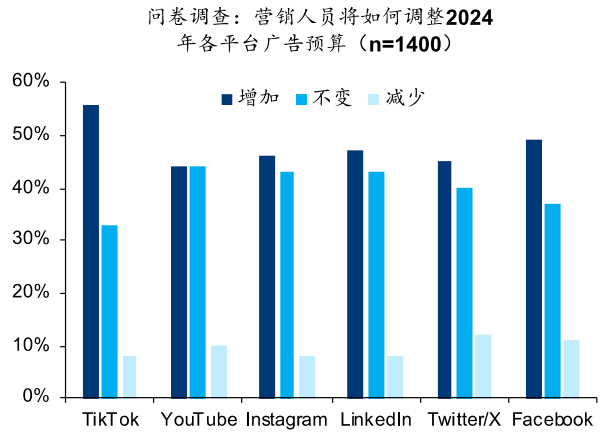
1）AI 在数字营销领域应用逐渐深化，内容生产与广告策略运营去中介化，激励更多长尾广告主入局。内容生产方面，AI+营销应用可提升素材生产效率，降低创意生产成本，提升长尾商户的创意质量。目前 AI 应用仍停留在较早期，如协助创作者获取灵感，在脚本撰写、多模态内容生成与编辑、甚至一键成片等方面，仍有深化空间。**投放方面，**各平台陆续推出 AI 驱动的自动化投放工具，如 Google Performance Max 和 Meta Advantage+，用户可根据投放目标，一键生成广告计划，无需手动优化。AI 可有效降低广告投放门槛，减轻对于广告投放人员的依赖，并有效提升转化效率。

图表82： AI 在营销领域的应用仍有深化空间



资料来源：HubSpot，华泰研究

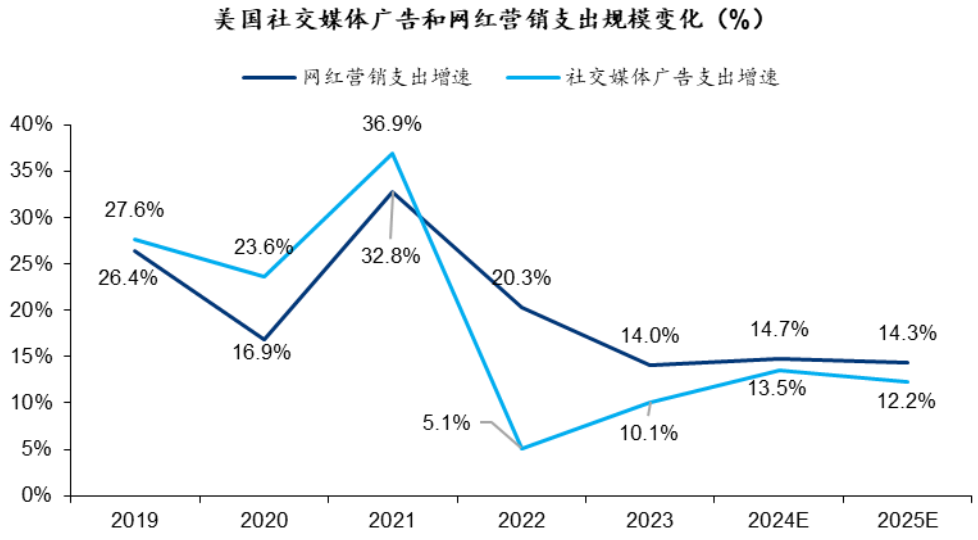
图表83： 广告主投放意愿：TikTok 营销预算增长态势最为显著



资料来源：HubSpot，华泰研究

2）广告投放与网红营销并重，KOL+UGC 内容宣传品牌。我们预期 24-25 年间，美国网红营销支出增速将保持 14%以上。与直接买量不同，网络 KOL 的 UGC 内容依托于名人粉丝群体，具备更高的互动与转化意愿。因此，在有限的品宣预算与流量成本攀升的背景下，网红营销有时更具性价比，其支出增长具有更强韧性。

图表84： 网红营销支出增长更有韧性



资料来源：eMarketer，华泰研究预测

AppLovin：AI 驱动业务增长，展现垂直领域精准广告趋势

AppLovin 是移动广告行业领先的用户获取和货币化平台，拥有 AppDiscovery、MAX、Adjust 等核心产品，产业链布局完善。公司推出 AI 广告引擎 AXON 2.0，采用自学习型技术，能够随着规模的扩大不断自我改进，构建公司的核心竞争力并推动广告业务持续增长。AppLovin 正积极拓展游戏领域之外的市场，特别是在电商领域取得显著进展。公司近期进行的电商广告测试表现优异，计划进一步增加投资并扩大业务规模。 此外，AppLovin 还通过 Wurl 探索了联网电视（CTV）渠道的广告机会，并计划将 AXON 2.0 的技术能力应用于 CTV 产品。公司业务主要分为两个部分：一方面，协助广告主进行客户获取，即利用广告主提供的广告预算在各类应用上投放广告，并根据广告效果（点击、展示、转化）进行收费；另一方面，通过 MAX 广告聚合平台为流量主提供服务，帮助他们通过竞价方式销售广告流量。

自成立以来，AppLovin 通过自主研发和战略并购，不断丰富其产品线，形成“软件平台+应用程序”双轮驱动：目前，公司的核心产品包括：1）App Discovery：广告投放工具，帮助触达优质用户进行广告投放；2）Max：广告聚合平台，通过集合竞价优化广告投放；3）Adjust：端到端分析工具，提供数据跟踪和归因分析；4）Sparklabs：利用人工智能技术，为客户提供创意素材。5）ALX：广告交易平台，连接广告主和发布商。6）Array：定期推荐精选的优质移动应用和游戏，为公告主投放区域提供建议。在应用程序方面，AppLovin 以超休闲游戏起家，先后收购 PeopleFun 和 Machine Zone 两家知名手游开发商，开发超 350 款流行的移动应用。

图表85： AppLovin 产品

产品	类型	简介
App Discovery	广告投放工具	帮助广告主触及优质客户，发掘增量受众，促进增长；优化广告投放，并提供自动化投放工具
SparkLabs	素材生成工具	利用人工智能，生成多种广告素材供客户选择
Max	广告聚合平台	聚合多家竞价商、广告主和广告平台，联同 Discovery 工具实现投放环节优化
ALX	广告交易平台	程序化广告交易，帮助广告主与理想客户建立链接，对重要广告信号覆盖，避免无效流量
Adjust	广告效果分析工具	为广告主提供自能化、精准的广告数据分析工具，包括归因分析和深度报告；提供数据保护和防作弊功能
Array	广告平台推荐工具	为广告主定期推荐 App 和游戏，不断鼓励用户发现新应用。
Wurl	投放平台	增加内容交互，提交订阅量，触达目标 CTV 受众

资料来源：AppLovin 官网，华泰研究

我们认为 AppLovin 定位为广告行业的“全链条玩家”。传统广告营销公司绝大多数定位为“中间商”，专注服务单一客户群体，如 IronSource 擅长帮开发者变现，或 The Trade Desk，致力于帮广告主实现精准投放。“中间商”商业模式痛点在于依赖客户的内容和需求，无法直接控制广告行业上下游。

AppLovin 的业务模式具有显著差异，它同时运营需求方平台（DSP）和广告聚合平台，从而全面掌握广告投放的整个链条，实现对广告主的预算分配、出价信息、流量主的广告位表现、竞价动态以及用户转化率等关键行为数据的深入洞察。依托这些全面的数据资源，AppLovin 能够显著提升其广告投放模型的效能，尤其是其广告引擎 Axon2。公司运用大型语言模型（LLM）来深入分析并解读数据中的复杂语义关系，包括用户意图和上下文需求，以构建更为精确的用户画像。在用户画像构建完成之后，AppLovin 采用强化学习技术，通过实时数据模拟不同的广告投放策略，从而显著加快了广告策略的验证过程，并有效缩短冷启动周期。

我们认为随着广告主业务的持续成长与完善，主流媒体在吸引新用户方面的能力正逐步降低，而中小型及长尾媒体正日益成为广告主获取新用户及重新激活现有用户的关键途径。程序化广告，作为一种高效连接长尾媒体与目标受众的方法，正在成为广告行业发展的不可逆转的趋势。利用先进的技术，程序化广告实现了数字广告采购和销售流程的自动化。它能够依据广告主设定的参数，在最佳时机以最优价格向目标受众展示定制化的广告内容。程序化广告系统不断搜集并分析用户行为数据，包括用户在应用中的停留时间和每用户平均收入（ARPU），以此来提升广告投放的精准度和效果。

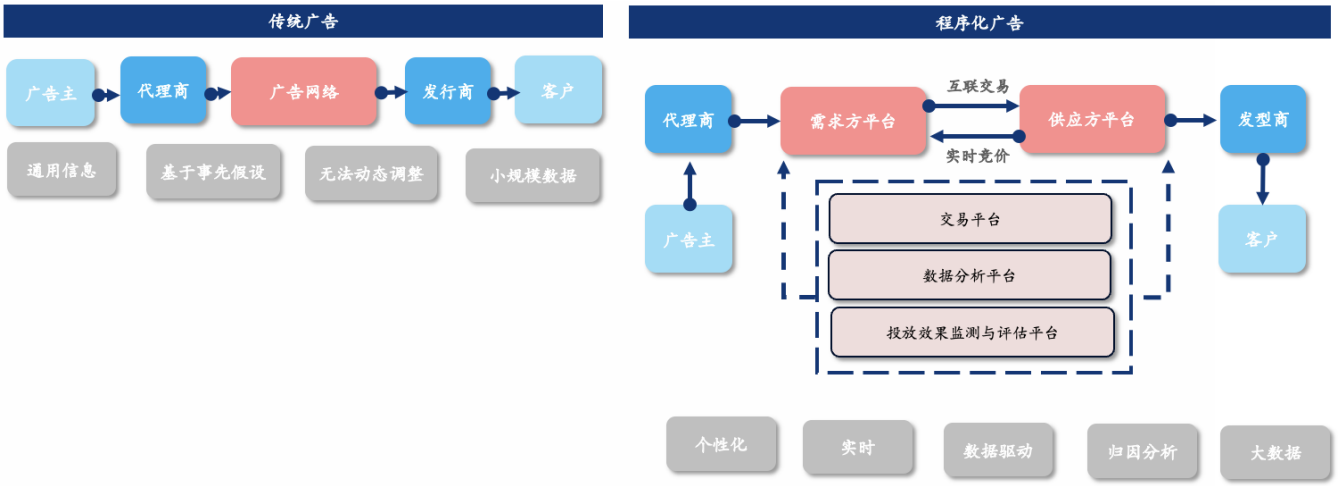
图表86：北美主要广告平台对比

公司	主要投放平台	广告投放类型	核心特点
		自有流量+广告平台	
Google	谷歌应用包括 Gamil、地图等，以及 YouTube 平台，甚至安卓平台	横幅广告、原生广告、视频广告、插屏广告、开屏广告等	1. 提供全面广告产品，数据优势明显，市场地位稳固。2. 竞争激烈，投放成本不可控。3. 多种服务模式可选，但标签精确度低。
Meta	FoA 应用包括 Facebook、Instagram 、Whatsapp 、Messenger 等	横幅广告、原生广告、视频广告、插屏广告	1. 广告投放成本相对较高。2. 广告内容需经过严格的审核流程。3. Meta 的广告标签系统更加精细，更适合广告主实现更精准的广告投放。
		第三方独立广告商	
Unity	游戏引擎+程序化广告平台	横幅广告、插屏广告、视频广告	1. 通过与 IronSource 的合并，公司在游戏发行领域，尤其是在超休闲游戏的推广上，实现能力和资源提升。2. 公司在获取中重度游戏用户方面的竞争力尚有待加强。
AppLovin	游戏发行+程序化广告平台+广告聚合平台	横幅广告、插屏广告、视频广告	1. 在超休闲游戏市场占据领先地位，同时在中重度游戏的用户获取上也取得了显著进展。2. 业务主要集中在北美和欧洲市场。3. 利用自有游戏生态系统收集的原始数据，增强了广告营销的预测和算法学习能力。

资料来源：谷歌、Meta、Unity、AppLovin 官网，华泰研究

我们认为，AppLovin 的发展或表示出大模型与具备丰富反馈数据的垂直领域广告的适配性。首先，我们须认识到营销行业与大模型之间存在着天然的协同效应。从训练数据看，作为互联网经济的基石，广告营销领域历经多年的发展，积累了丰富的反馈数据资源，包括但不限于实时点击率和转化率等关键指标，为大模型的实际应用奠定了坚实的基础。从营销的工作流程来看，大模型应用与之高度契合。在营销过程中，创意内容的制作和素材的快速迭代是常态，而这些正是大模型所擅长的领域之一。从营销结果看，营销的具体场景中，广告平台的目标明确且易于量化，这为大模型的训练提供了理想的条件。广告主普遍关注的关键绩效指标，如投资回报率（ROI）或广告支出回报率（ROAS），可以直接作为强化学习算法中的奖励函数。平台在设计这些奖励函数时，能够根据不同广告主的具体需求，定制出既量化又明确的目标，无论是侧重于短期转化还是长期用户留存。

图表87：传统广告和程序化广告对比



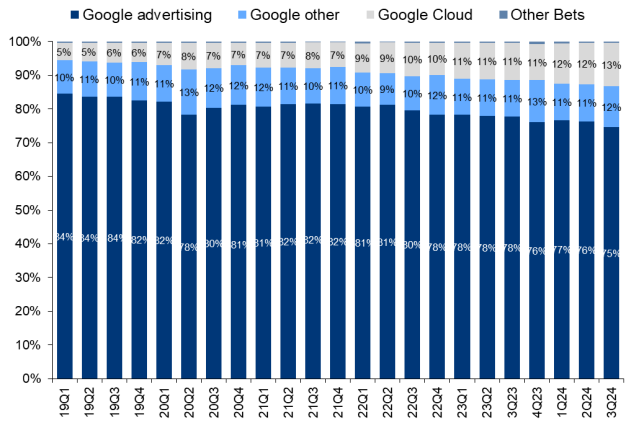
资料来源：架构师社区公众号，AppLovin，华泰研究

云计算业务：云业务崭露头角，AI 服务提升附加价值

我们预计谷歌 Cloud 业务 FY24/25/26 营收为 434/553/694 亿美元，对应同比为 31.2%/27.4%/25.5%，主要由全链条 AI 云端布局驱动。

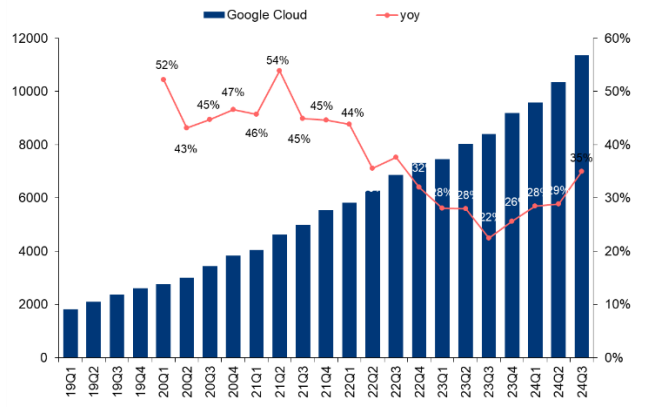
谷歌云业务的发展一直呈追赶之势，扭亏为盈后发展潜力值得重估。自 2008 年推出以来直到 23Q1，谷歌云业务一直处于亏损状态，公司持续投入大量资金来扩大云计算基础设施、开发新产品、招募新客户和人才等。经过长期市场攻坚，谷歌于 23Q1 终于实现云业务的扭亏为盈，并连续保持七个季度盈利为正数。24Q3 云业务营收同比+35.0%至 113.5 亿美元，超一致预期 4.5%，经营利润大幅增长至 19.5 亿美元，经营利润率 17%。随着生成式 AI 发展，越来越多的公司转向公共云来运行繁重的 AI 工作负载。24 年 9 月公司表示 60%的人工智能初创公司和 90%的人工智能独角兽公司在 Google Cloud 上进行训练和推理，2024 年使用 Gemini 的云客户增长超过 4 倍。24Q3 谷歌资本支出同比+62.1%至 130.6 亿美元。根据公司指引，24Q4 资本开支将与 Q3 持平，25 年增速将超 23-24 年，更多需求将会来自推理端。

图表88：Google 分业务营收占比



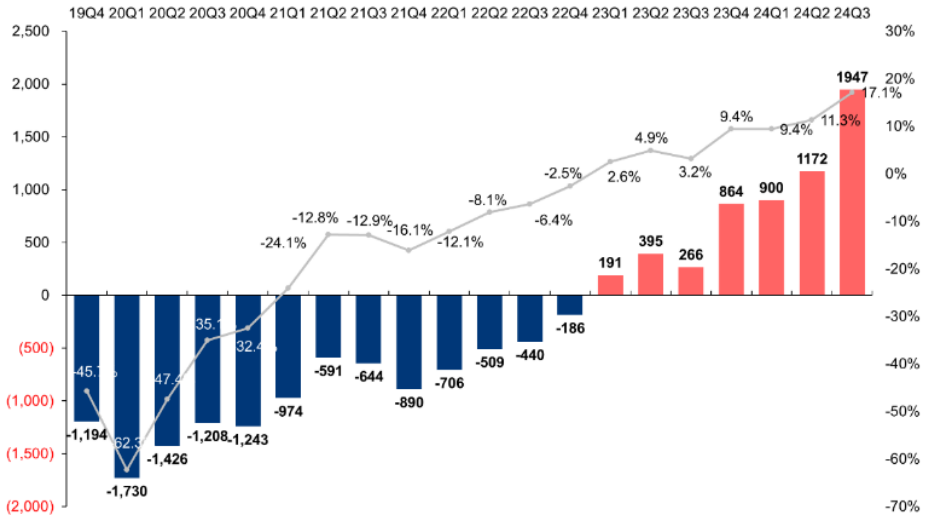
资料来源：谷歌官网、华泰研究

图表89：Google 云业务营收及同比增速（单位：百万美元）



注：谷歌云业务经营利润披露自 19Q4 开始
资料来源：谷歌官网、华泰研究

图表90：Google 云业务经营利润及经营利润率（单位：百万美元）



资料来源：谷歌官网、华泰研究

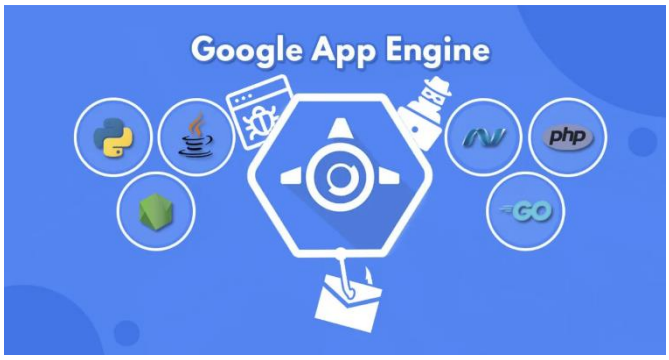
谷歌云：从技术优先到产品、客户优先，乘 AI 之风扬帆起航

目前谷歌云在全球云基础设施市场排名第三，份额为 10%，排在 AWS 的 33%和 Azure 的 20%之后，且业务增长势头稳健，过去三年间大于 2.5 亿美元合同数量增长达 300%。我们认为，谷歌云业务发展历程可分为以下三个阶段：

1） **第一阶段（2000-2006）：云起前夜。**谷歌初期主要专注于搜索业务，但随互联网发展和数据量急剧增长，谷歌逐渐意识到云计算的重要性，并开始进行相关技术研究，着手开发了系列分布式计算机系统。在此阶段公司 SVP of Technical Infrastructure Urs Hölzle 对谷歌云基础设施的开发奠定了基础。2003 至 2006 年，谷歌连续发表四篇文章揭示云计算的基础架构：分布式文件系统（GFS）、并行计算（MapReduce）、数据管理（BigTable）和分布式资源管理（Chubby），实现对云计算领域的最早探索。

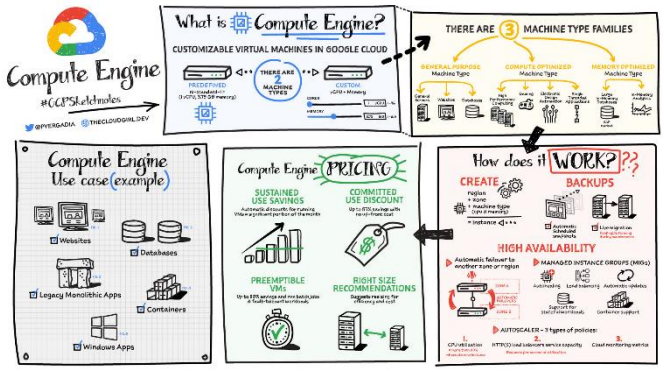
2） **第二阶段（2006-2014）：云芽初露。**2006 年，谷歌时任 CEO Eric Schmidt 在 SES San Jose 2006 上首次公开提出“云计算（Cloud Computing）”的概念，标志着云计算时代的到来。但亚马逊却抢先于 2006 年发布了 EC2 计算服务与 S3 存储服务等产品，凭借“技术创新+低价策略+客户至上”三轮开拓 IaaS 市场，并逐步将客户发展重心从初创公司与开发者转向传统企业，为 AWS 的后续发展奠定了基础。直到 2008 年，谷歌才推出第一款基于 PaaS 的云产品 Google App Engine(GAE)。随后谷歌相继推出 Google Compute Engine 和 Google Cloud Platform，但彼时其主要业务仍然集中在搜索领域，对比同期 AWS 布局云服务市场速度略有不及。

图表91： Google App Engine（GAE）



资料来源：Medium 官网、华泰研究

图表92： Google Compute Engine（GCE）

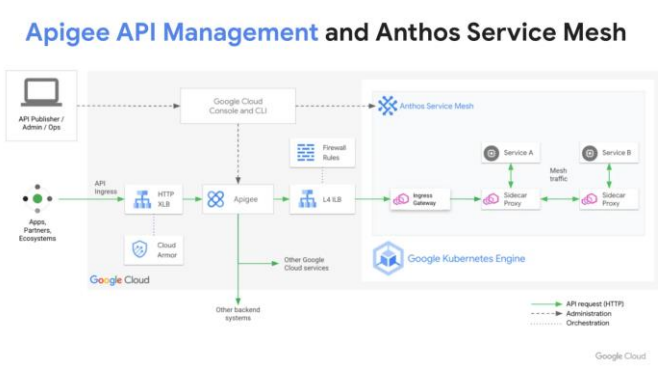


资料来源：谷歌云官网、华泰研究

3） **第三阶段（2014 年至今）：云智飞跃。**2014 年起，谷歌开始加速布局云业务，并逐渐缩小与 AWS 和微软的差距。**1）Diane Greene 时期：成功推动谷歌云商业化发展。**谷歌于 2015 年聘用 VMware 创始人 Diane Greene 作为谷歌云 CEO。首先，Greene 上任后整合企业云业务，通过将销售、营销、GCP 和 Google Apps/G Suite 集成到谷歌云中，共同打造统一业务。其次，Greene 为谷歌云建立广泛的合作伙伴网络，合作范围涵盖高露洁、迪士尼、eBay、汇丰银行、拉丁美洲航空、LG CNS 和纽约时报等多个大型企业客户，促进了谷歌打造企业云平台的进展，多方位战略并购以与 AWS 竞争。谷歌云收购了 Apigee、Kaggle、qwiklabs 等小型初创公司。其中 Apigee 让谷歌云进入了 API 网关和管理市场，并帮助谷歌加速推出 Google Assistant 和基于 Google Home 的家庭自动化产品。最后，Greene 加倍投入人工智能，在 GCP 推出多种 ML 和 AI 相关服务，从而将谷歌云打造成复杂的数据驱动和以 AI 为中心的云平台。**2）Thomas Kurian 时期：加深客户关系打造销售端网络。**2019 年 2 月，前甲骨文执行官 Thomas Kurian 成为谷歌云 CEO。Kurian 上任后首先持续扩大销售队伍规模。19 年谷歌云销售团队规模仅为 AWS 和微软 Azure 的 1/10 到 1/15，而 2021 年占比已接近 1/2。其次，Kurian 简化了繁琐的销售折扣流程与部署机器时间，在上任后的四年中，将供应和部署机器的周期时间缩短了五倍。同时，Kurian 将谷歌云市场策略调整为客户需求导向。其在 2021 年重启 Google Cloud 社区以来，两年间社区已达到了 10 万成员。最后，Kurian 继续加深与大型企业客户和渠道伙伴的关系，并致力于扩大其公共部门在美国和海外合

作伙伴规模，以此打造谷歌云合作生态。

图表93：使用 Apigee API 构建现代应用和架构



资料来源：谷歌云官网、华泰研究

图表94：谷歌云社区发展历程

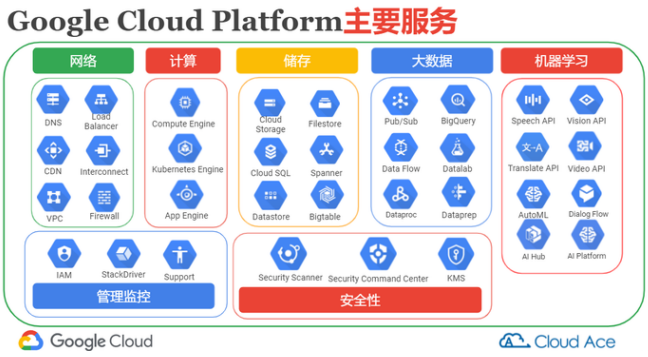
资料来源：谷歌云官网、华泰研究

目前，谷歌云服务主要包括以下两大类别：

1）Google Cloud Platform：提供 IaaS、PaaS 和无服务器计算三类环境。该平台包括一系列在谷歌硬件上运行的用于计算、存储和应用程序开发的托管服务。具体包括网络、计算、储存、大数据、机器学习、管理监控与网络安全七类服务。同时 Vertex AI 和 Duet AI 能构建和扩缩生成式 AI 应用，协助客户在谷歌云上实现 AI 创新。

2）Google Workspace：属于 SaaS 层应用服务，作为谷歌在订阅基础上提供的一套云计算生产力和协作软件工具和软件，其包含 Gmail、Chat、Calendar、Drive、Docs 和 Meet 等基于云的通信和协作工具，从而与微软 Microsoft 365 展开竞争。如今 Workspace 也集成了 Duet AI 等功能，使用户可以利用生成式 AI 来提高组织的生产力。

图表95：Google Cloud Platform 主要服务



资料来源：谷歌云官网、华泰研究

图表96：Google Workspace 部分产品

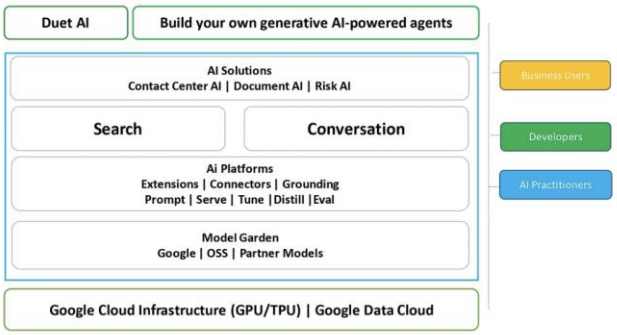
资料来源：谷歌云官网、华泰研究

本轮 AI 行情下，谷歌云业务有望乘 AI 之风扬帆起航。AI 时代到来之前，谷歌便通过搜索引擎拥有约二十年的大数据处理经验，借助自建数据中心的计算资源及数据分析技术，大量提取具有商业价值的数据。在此基础上，公司致力于为云端用户提供与其内部同等级别的 AI 及机器学习服务。**我们认为，AI 浪潮之下，谷歌云的 AI 产品有望在集成 Gemini 品牌后提升工作效率和用户体验，**并提升 IaaS、PaaS 与 SaaS 市场份额。24 年 10 月公司表示目前 Gemini 与 Google Assistant 在安卓平台上整合顺利、用户反馈良好，Gemini API call 在 6 个月内增长近 40%，2024 年使用 Gemini 的云客户增长超过 4 倍。但此前 AWS 与微软分别在 IaaS 与 SaaS 领域占据龙头，且微软凭借 OpenAI 具备一定的 AI 先发优势，谷歌云目前虽步入佳境，但未来或依旧面临二者的竞争压力。

1） **Vertex AI（ML PaaS）：**2021 年 5 月，谷歌发布 Vertex AI 作为机器学习平台，旨在通过提供涵盖从数据准备和模型训练到部署、管理和监控的整个机器学习周期的端到端平台，使机器学习变得更加易于访问、协作和高效。Vertex AI 能与 AutoML、Model Garden、生成式 AI 等功能相结合，其不仅支持在 Gemini 模型中提供提示并执行测试，还支持 Model Garden 中 130 多种生成式 AI 模型和工具，让用户根据应用场景自定义并部署模型。此外，Vertex AI 还拥有开放的集成式 AI 平台和适用于预测式/生成式 AI 的 MLOps，能提供构建和使用生成式 AI 的一切需要。

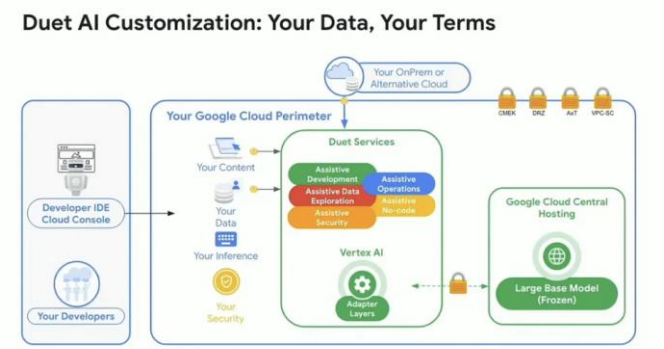
2） **Duet AI（SaaS，后并入 Gemini 品牌）：**2023 年 5 月，谷歌发布完全托管服务 Duet AI 来帮助用户提高工作效率和创造力。Duet AI 能集成于 Google Workspace 上，并提供 AI 驱动的代码编写、聊天帮助、问题排查与上下文智能操作等功能，从而对标微软 Copilot。**2024 年 2 月 8 日，谷歌宣布将 Duet AI Workspace 更名为 Gemini for Google Workspace**，这不仅意味着 Gemini 1.0 Ultra 模型将内置到 Workspace 中，统一了谷歌大模型的战略品牌，还具备了企业级数据保护功能，使用户能利用 Gemini 进行研究、总结数据、寻找业务趋势并创建文案。

图表97： Google Cloud Vertex AI 架构



资料来源：谷歌云官网、华泰研究

图表98： Google Cloud Duet AI

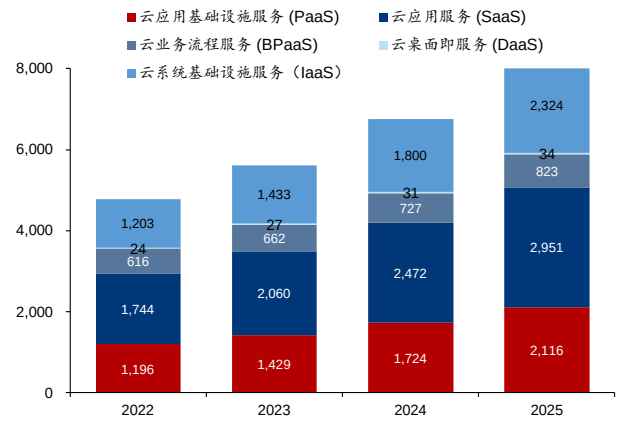


资料来源：谷歌云官网、华泰研究

行业趋势：AI 军备竞赛，全球云计算资本支出将保持“温和增长”态势

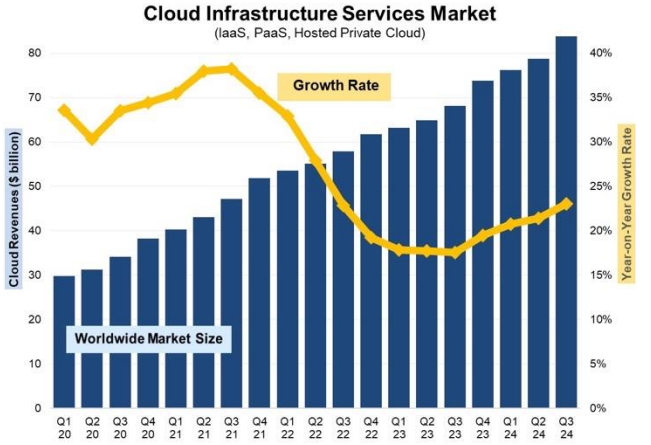
全球云计算市场规模增长势头稳定，AWS 增速放缓但仍为业内龙头。据 Gartner 2024 年 5 月数据，2023 年全球云计算市场规模为 5611 亿美元，预计 2024 年将同比增长 20.4%至 6754 亿美元；其中 IaaS/PaaS/SaaS 同比增长分别为 25.6%/20.6%/20.0%，主要驱动力来自生成式 AI 的崛起和全球云基础设施服务支出增加。2023-2028 年复合增长率达 19.7%。从竞争地位来看，目前云计算行业较为集中，根据 Synergy Research Group 2024 年 11 月 1 日数据，24Q3 前三大云供应商合计占据全球市场份额的 63%，其中 AWS/微软/谷歌分别占比 31%/20%/12%。二线云提供商中，华为、Snowflake、MongoDB、Oracle 和 VMware 市场份额亦增长较快。

图表99：全球公共云服务最终用户支出预测（亿美元）



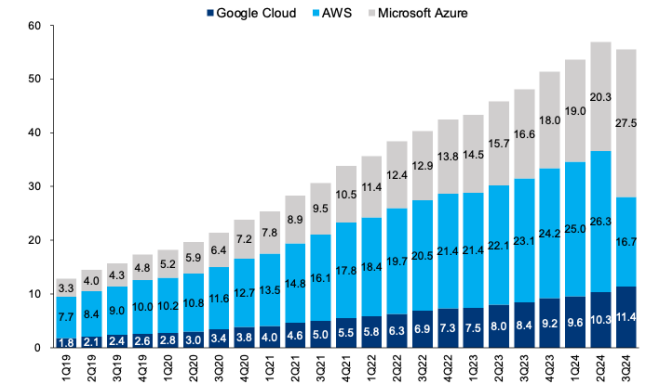
资料来源：Gartner 官网、华泰研究

图表100：全球云基础设施服务规模增速



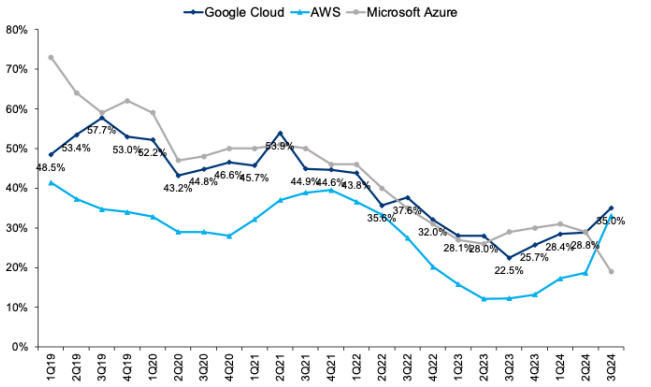
资料来源：Synergy Research Group 官网、华泰研究

图表101：三大云巨头云业务季度收入（十亿美元）



资料来源：各公司公告，Visible Alpha，华泰研究

图表102：三大云巨头云业务季度收入同比增速



资料来源：各公司公告，Visible Alpha，华泰研究

宏观经济回暖，云巨头恢复资本开支节奏。2022 年至 2023 年初，伴随着宏观经济增长放缓、企业数字化转型需求减弱与企业 IT 开支趋于谨慎等因素，三大云巨头总资本开支连续五个季度放缓。23Q1 亚马逊、微软与谷歌资本开支分别为 142/66/63 亿美元，总资本开支同比-10%至历史低点。然而伴随着宏观经济复苏、AI 提振云计算需求、企业逐渐扩大与头部云厂商的消费合同协议等因素，云巨头资本开支于 23Q2 开始有所改观，2023 年全年云巨头总资本开支稳步上涨，并攀升至 24Q3 的 589 亿美元，其中亚马逊、微软与谷歌资本开支分别为 226/149/131 亿美元。

图表103：三大云巨头最新季度资本开支情况

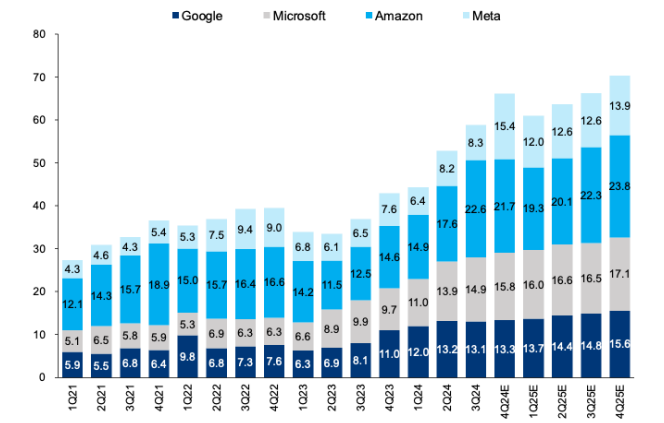
产品名称	资本开支	总结
谷歌	24Q3 资本开支达到 131 亿美元，同比+62%，环比-1%	24Q4 资本开支将与 Q3 持平，25 年增速将超 23-24 年，更多需求将会来自推理端。
亚马逊	24Q3 公司资本开支 226 亿美元，同比+81%，环比+28%	公司预计 2024 年资本开支将稳步上涨，并计划未来 5-10 年间持续投资马来西亚、印度、韩国、日本等地云业务，以扩大云计算基础设施规模
微软	FY25Q1 资本开支（不含融资租赁）为 149 亿美元，同比+50%，环比+8%	未来将继续扩大基础设施投资，FY25Q2 资本支出将环比增加

资料来源：各公司官网、华泰研究

AI 促进预期消费需求增加，有望推动下一轮资本开支浪潮。2023 年 3 月，GPT-4 的推出引发行业加码投资 AI，为了减轻投资不足带来的风险（例如对需求准备不足或错失关键机会），云巨头开始下一轮军备竞赛，确保能根据客户对 AI 日益增长的需求扩展产品。因此均表示，资本支出预计将在 2025 年继续保持增长趋势，也从传统服务器逐步转移至 AI 设备中。随着全球范围内 AI 应用改变云计算商业模式，叠加宏观经济恢复下企业客户云迁移工作需求激增，云巨头资本开支有望持续高增。根据公司指引，我们预计日历 25 年四家资本开支总和将同增 20%，达到 2400 亿美元左右。

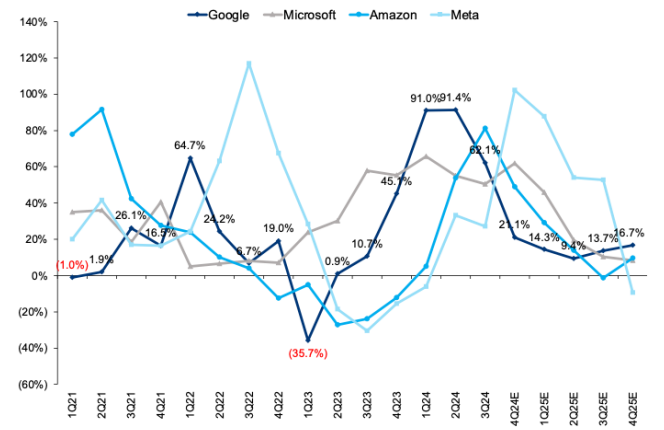
展望 24-26 年，我们预期全球云计算的资本支出将保持“温和增长”态势，主系 1）海外经济恢复增长，企业客户云迁移工作需求增加，2）AI 赋能云计算发展推动行业革新，3）未来大模型训练和推理均需要更多算力供给。

图表104：四大科技巨头资本开支（百万美元）



资料来源：各公司公告，Visible Alpha，华泰研究

图表105：四大科技巨头资本开支同比（%）



资料来源：各公司公告，Visible Alpha，华泰研究

竞争格局：亚马逊保持先发优势，微软和谷歌强势追赶

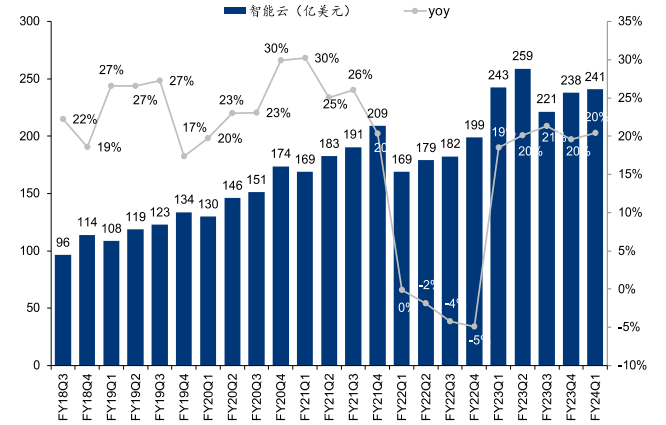
微软以 SaaS 层开路 PaaS 与 IaaS，云业务市场份额强势追赶 AWS。微软并没有直接与 AWS 在 IaaS 市场展开竞争，而是另辟蹊径通过自身强势软件上云，率先布局 SaaS 领域。微软不仅通过 Copilot 服务赋能 Microsoft 365、Dynamics 365 等服务，稳固 SaaS 市场份额，公司还在 2016 年通过 Azure 提供混合云解决方案开路 PaaS，对比 AWS 与谷歌在两年后才陆续布局。此外，微软通过 Azure App Services 提供网络应用、移动应用、逻辑应用等托管平台服务，逐步抢占 PaaS 市场份额。如今公司已具备 IaaS-PaaS-SaaS 全产业链云生态。与亚马逊营收主要靠电商业务拉动、谷歌营收主要靠广告业务拉动不同，微软智能云业务营收占比自 FY22Q4 至 FY24Q4 均维持在 40%以上，成为公司最主要的收入来源。此外，据 Synergy Research Group 统计，CY24Q3 微软在全球云基础设施服务市场份额达 20%，较年初份额有所下滑。

图表106：Azure AI 产品组合及功能

产品名称	功能
Azure OpenAI Service	提供对 OpenAI 语言模型的 REST API 访问，包括 GPT-4、GPT-4 Turbo with Vision、GPT-3.5-Turbo 和 Embeddings 模型系列
Azure AI Search	Azure AI Search 在传统和生成式 AI 搜索应用中针对用户拥有的内容提供大规模的安全信息检索
Azure AI Studio	目前处于公共预览版，其汇集了多个 Azure AI 服务功能，能够帮助评估大型语言模型响应，并通过提示流编排提示应用程序组件，以获得更好的性能
Azure AI Content Safety	可检测应用程序和服务中用户生成和 AI 生成的有害内容，检测内容包括文本和图像 API。此外，其还提供了交互式内容安全工作室，可让用户查看、探索和尝试示例代码，以跨不同方式检测有害内容
Azure Machine Learning	一项用于加速和管理机器学习项目生命周期的云服务，用户可以在日常工作流程中使用它来训练和部署模型以及管理机器学习操作（MLOps）。用户在机器学习中创建模型，或使用从开源平台（PyTorch、TensorFlow 或 scikit-learn）构建的模型
Azure Machine Learning prompt flow	旨在简化由大型语言模型提供支持的 AI 应用程序的整个开发周期。其提供了一个全面的解决方案，可简化 AI 应用程序的原型设计、实验、迭代和部署过程
Responsible AI dashboard support for text and image data	目前处于公共预览版，该功能能够让开发者在构建、训练或评估模型阶段，评估使用非结构化数据来构建的大模型，有助于开发者在部署模型之前发现模型错误、公平性问题以及模型解释，从而实现更具公平性的高性能计算机视觉和 NLP 模型

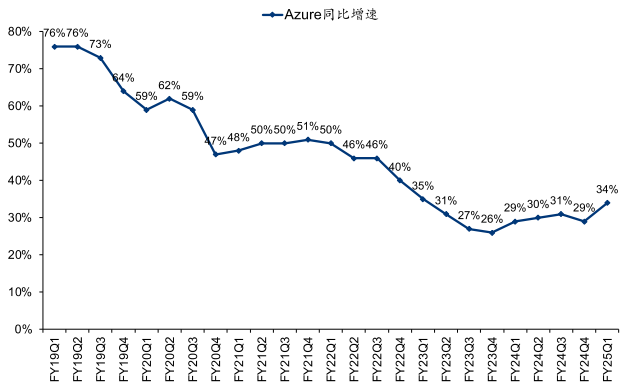
资料来源：微软官网、华泰研究

图表107：微软智能云季度营收（亿美元）



资料来源：微软官网、华泰研究

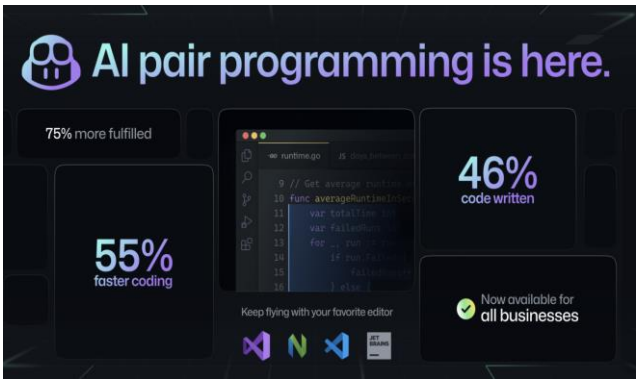
图表108：微软 Azure 营收同比增速



资料来源：微软官网、华泰研究

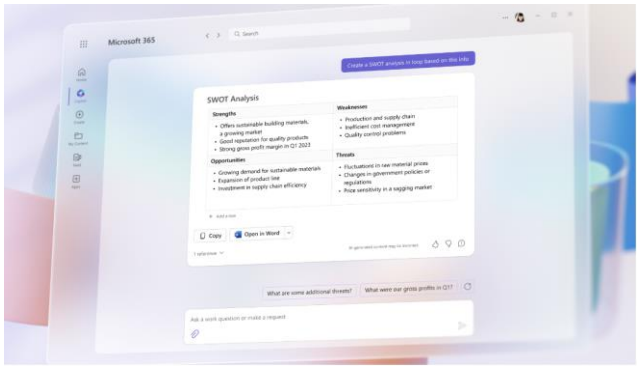
微软多款生产力工具支撑顶层 SaaS 服务，Copilot+AI 商业化可期。微软通过 Microsoft 365 与 Dynamics 365 等顶层应用布局 SaaS 层具备先发优势：此前微软已将 Office 365 生产力套件更名为 Microsoft 365，并转入云端。同时，微软通过在 Microsoft 365 Copilot 中引入 LLM，并将其集成在 Microsoft 365 中，有望打开额外变现增值空间。2023 年 11 月企业端 Microsoft 365 Copilot 服务正式推出，价格为每月 30 美元。客户方面，FY25Q1 Copilot 企业客户环比增长 55%，AMD 和 Flutter Entertainment 等公司将 Copilot 加入自身代码库，近 70%的《财富》500 强公司正在使用 Microsoft365 Copilot。设备方面，Windows 版 Copilot 目前已安装在近 2.25 亿台 Windows 10 和 Windows 11 电脑上。此外，微软也将 AI 融入 Dynamics 365 与 Teams 等其他 SaaS 服务中，譬如互动式 AI 助手、自动生成会议纪要等功能。**我们认为，对比亚马逊与谷歌在 SaaS 层应用相对单一，目前微软在全球生产力软件市场布局广泛，在 SaaS 市场保持较高份额。随着 Microsoft 365 Copilot 向企业用户开放，相关收入将逐步增长。**

图表109：微软 GitHub Copilot for Business



资料来源：Github 官网、华泰研究

图表110：微软 Microsoft 365 Copilot 工作图



资料来源：微软官网、华泰研究

亚马逊电商盈利助力云业务规模壁垒，多元布局让公司成功转型。亚马逊早期以自营电商业务起家，通过平台引流第三方电商与广告业务促进现金流，并持续提升盈利，具备一定先发优势。电商业务发展之余亚马逊开始积极布局云业务，十余年间通过大量资本支出建设 IaaS 数据中心，并多元化布局 PaaS 平台云服务与 SaaS 软件业务，成为云计算市场的领导者。我们认为，如今云、大数据与 AI 深度融合背景下，云业务竞争趋于白热化，但 **AWS 深耕云计算领域多年，具备庞大技术壁垒与丰富生态资源，叠加亚马逊电商业务充足的现金流支撑，AWS 云业务后续发展有望保持强劲势头，长期领跑市场。**

AWS 产品服务涵盖整个云服务框架，应用范围广泛。经历十余载发展，如今 AWS 云产品服务涵盖了 IaaS、PaaS 和 SaaS 三大层面，提供包括存储、数据计算、机器学习、虚拟机、边缘计算、物联网、AI 在内的多个服务。目前 AWS 在全球 33 个区域内运营着 105 个可用区，拥有 600 余个入网点，服务 245 个国家和区域，并计划未来在德国、马来西亚、墨西哥、新西兰和泰国等地新增 15 个可用区及 5 个 AWS 区域。下游行业方面，AWS 目前已覆盖半导体、新能源、金融、游戏、零售、医疗等 20 余个细分领域。

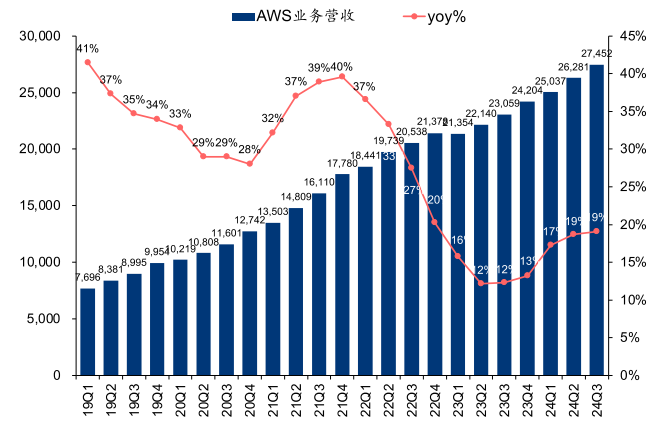
图表111： AWS 在三种云服务模式下的部分主要产品

云服务模式	主要产品	服务内容
IaaS	Amazon EC2 (Elastic Compute Cloud)	弹性计算平台，用户可以快速启动和管理虚拟服务器并处理任务
	Amazon EBS (Elastic Block Store)	易于使用、可扩展、高性能的数据块存储服务，为 Amazon EC2 而设计
	Amazon S3 (Simple Storage Service)	对象存储服务，可将数据以对象形式存储在存储桶中
	Amazon VPC (Virtual Private Cloud)	虚拟网络环境，用户能够全面控制自己的虚拟网络环境，包括资源放置、连接性和安全性
PaaS	AWS Elastic Beanstal	用于部署和扩展 Web 应用程序和服务的环境，用户只需上传代码即可自动处理部署细节
	AWS Lambda	提供计算服务，可以运行永和的代码以响应事件并自动管理计算资源
	AWS App Runner	完全托管的应用程序服务，允许用户构建、部署和运行 Web 应用程序和 API 服务
	AWS RDS (Relational Database Service)	提供托管式服务的集合，可以简化在云中设置、运营和扩展数据库的过程
SaaS	Amazon WorkSpaces	在 AWS 云中运行的完全托管式安全桌面计算服务，用户可以预置基于云的虚拟桌面
	Amazon Connect	基于云的客户联系中心服务
	Amazon Chime	通信服务，让您可以在企业内部和外部召开会议、聊天和拨打商务电话
	Amazon WorkDocs	托管的用于实现安全的内容创建、存储和协作的服务，用户可以创建、编辑和共享内容

资料来源：AWS 官网、华泰研究

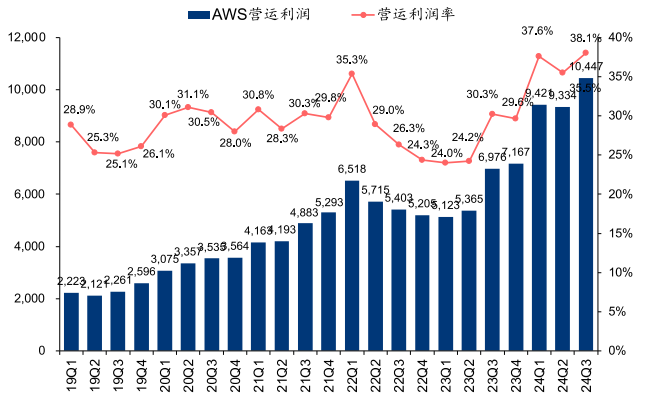
云业务竞争趋白热化，AWS 营收回暖但同比增速低于竞争者。亚马逊、微软、谷歌云业务价格战持续升级，对比微软 FY25Q1 Azure 营收同比+34%，谷歌 24Q3 云业务营收同比+35.0%至 113.5 亿美元，AWS 增速有回暖趋势，24Q3 营收为 274.5 亿美元，同比+19%，环比+4%；同比增速对比前六个季度有所回升，略低于 22Q4 水平。AWS 营运利润为 104.5 亿美元，同比+50%，环比+12%，约占总营运利润的 60%，营运利润率为 38.1%，同比+7.8pct，增长部分来自公司成本结构调整与成本控制见效。

图表112： AWS 业务营收及同比增速（单位：百万美元）



资料来源：亚马逊官网、华泰研究

图表113： AWS 业务经营利润及经营利润率（单位：百万美元）



资料来源：亚马逊官网、华泰研究

Other Bets 业务：未来科技的探索储备

Other Bets 业务主要指在核心业务谷歌之外运营的公司或项目，主要包括：1）Waymo，主要开发自动驾驶解决方案并提供自动驾驶服务；2）Verily，生命科学研究机构，专注于健康相关项目，包括疾病预防、管理以及医疗技术革新；3）Calico，生物技术公司，专注于衰老生物学；4）X Development LLC，创新实验室，是一个由发明家和企业家组成的多元化群体，目标为开发改善数百万人，甚至数十亿人生活的技术；5）Wing，致力于开发无人机送货系统，提高物流和配送效率，及其他项目等。

其中以 Waymo 最为人熟悉，现已成为美国自动驾驶行业的技术领导者，并不断加强商业化落地。截止 24 年 12 月，Waymo 每周的全自动驾驶里程超过 100 万英里，提供超过 17.5 万次付费乘车服务。Waymo 与 Uber 在 Austin 和 Atlanta 扩大运营网络和战略合作，以及与现代汽车建立新的长期合作关系，近期自动驾驶公司如地平线、文远知行等纷纷上市，我们认为若 Waymo 能分拆上市，或能释放可观价值。**综合以上分析，我们预计 Other Bets 业务 FY24/25/26 营收为 21/28/37 亿美元，同比增长为 34.4%/35.0%/35.0%。**

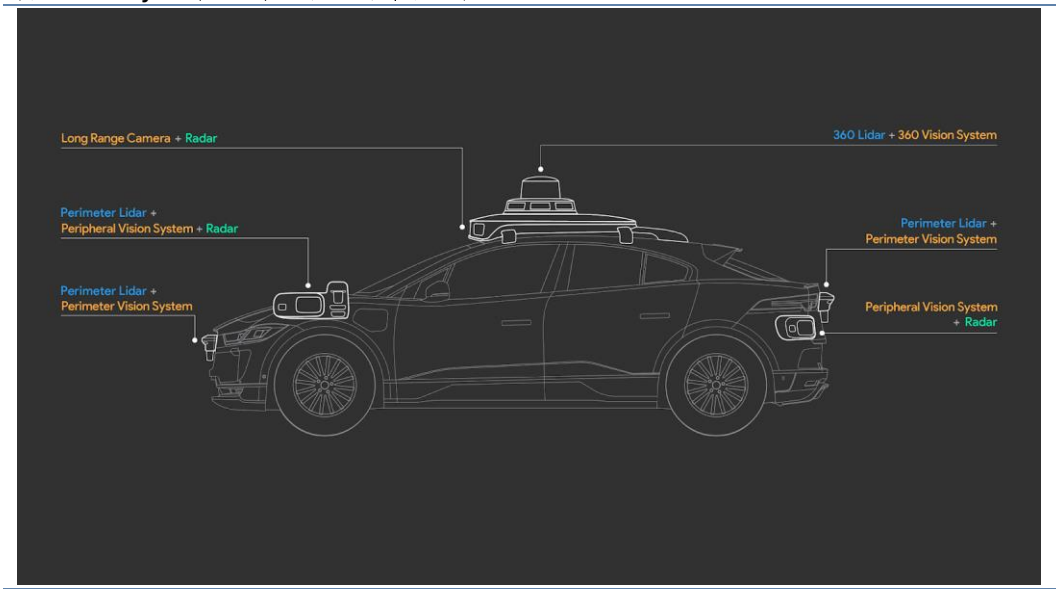
Waymo：全球自动驾驶先驱者，服务遍及美国四大城市

Waymo 是全球自动驾驶技术的领军企业，其起源可追溯至 2009 年谷歌启动的自动驾驶汽车项目，后于 2016 年 12 月从谷歌独立出来，成为 Alphabet 公司的子公司。自 2016 年起，Waymo 在美国凤凰城开展自动驾驶测试，至今已有 8 年历史。2018 年，Waymo 推出了商业化运营服务 Waymo One，自动驾驶打车服务成为 Waymo 的收入来源之一。此外，Waymo 还通过 Waymo Via 提供自动驾驶卡车运输和物流服务。

2024 年 6 月，Waymo 在旧金山向用户开放无人驾驶出行服务，近 30 万人完成注册。目前，Waymo 的自动驾驶出租车服务已逐步扩展至美国的四个市郊区域，包括菲尼克斯、洛杉矶、旧金山和奥斯汀。2024 年 12 月，Waymo 宣布在美国每周的无人车付费出行次数已超过 17.5 万次。

目前 Waymo 车型已经过六轮迭代，2020 年 3 月发布的第五代车型基于捷豹 I-PACE，开始采用全景激光雷达和角落补盲激光雷达系统。基于前代的基础上，2024 年 8 月发布的第六代车型中，包含 13 个摄像头、4 个激光雷达、6 个雷达和一系列外部音频接收器(EAR)，在显著降低成本的同时提高性能。

图表114：Waymo 第五代车型传感器分布示意图



资料来源：Waymo 官网、华泰研究

Waymo 实时数据丰富，具备在超过 2000 万英里的真实驾驶和超过 2000 亿英里的模拟驾驶中积累的经验，其感知系统能够从其汽车传感器高级套件中收集复杂数据，并通过机器学习等技术识别周围的对象，包括行人、骑行者、车辆和建筑物等。此外，Waymo 的备用系统包括辅助计算、备用碰撞检测与规避系统、冗余转向系统、冗余制动系统、备用动力系统、用于车辆定位的冗余惯性测量系统以及信息安全系统。

Waymo 采取提前绘制详细地图和收集数据的方式，以确保行驶数据稳定。在 Waymo 进入新区域运营之前，会先对该地区进行极为详细的地图绘制，包括车道标记、停车标志、路缘和人行横道等。然后，Waymo 会将这些详细的自定义地图与实时传感器数据结合使用，以确定其精确的道路位置。

图表115：Waymo 全景图

资料来源：公司公告，Waymo 官网，汽车视界研究公众号，华泰研究

我们认为，Waymo 已初步在美国市场建立了自动驾驶的先发优势，主要体现在：1）已在多个城市证明可在没有安全员的情况下长时间运营；2）数据层面不仅有通过广泛的真实路测得来的广泛应对数据，同时也有用于测试边缘案例的模拟功能；3）为保障安全性而提供的关键冗余系统；4）定制开发的硬件有助于优化软件堆栈。

Verily：AI 医疗解决方案的前沿探索

Verily 诞生于 Google X 的一项生命科学与医药方向的项目探索，于 2015 年被分离出来，成为 Alphabet 的子公司，目标为以实验为中心，通过开发工具、服务和软件来帮助医疗生态系统。其业务主要包括：

- **精准医疗解决方案：**Verily 致力于使用来自临床、社会和行为等各种来源的数据，为个人需求量身定制医疗保健解决方案。其服务包括慢性疾病管理、临床试验平台和卫生系统分析工具等。例如，Verily Lightpath 为人工智能支持的慢性病护理平台，能够因人而异，帮助实现和维持糖尿病、肥胖症和其他代谢疾病的健康目标。
- **硬件创新：**Verily 开发了包括监测糖尿病和睡眠呼吸暂停等慢性疾病的设备。例如与 Dexcom 合作，为其 G6 和 G7 连续血糖监测设备提供微型化技术，以及推出 Verily Study Watch，一款具有可定制功能的临床级设备，用于收集临床试验中的数据。
- **公共卫生研究：**Verily 在 COVID-19 疫情爆发期间，成功建立了包括 Healthy at Work 在内的人口健康计划，通过疫苗追踪和检测计划帮助员工安全重返工作岗位。此外，由斯坦福大学、埃默里大学和 Verily 共同发起的 WastewaterSCAN 计划旨在开发和扩展美国国家废水监测系统，为公共卫生措施提供信息，以减轻 COVID-19 等传染病的传播。
- **医疗保险：**与其子公司 Coefficient Insurance 共同进军保险行业，为寻求管理医疗保健费用的自筹资金雇主提供保险。

据 Business Insider 12 月 18 日报道，Verily 计划在 2025 年增加外部资金来源，并将其战略重点放在 AI 医疗上，且从 2025 年 1 月起，它将脱离谷歌的部分内部系统，寻求独立发展。我们认为，若 Verily 实现分拆，或能为谷歌释放价值。

盈利预测与估值

我们预计 FY24/25/26 谷歌总营收分别为 3504/3916/4367 亿美元，同比为 14.0%/11.8%/11.5%。基于报告前序章节对谷歌 AI 和云计算业务现状和前景分析，以及对市场态势的判断，我们预计谷歌 Services 业务 FY24/25/26 营收为 3046/3332/3632 亿美元，同比增长为 11.8%/9.4%/9.0%，主要由行业渗透率和 YouTube 增长驱动。谷歌 Cloud 业务 FY24/25/26 营收为 434/553/694 亿美元，同比增长为 31.2%/27.4%/25.5%，主要由全链条 AI 云端布局驱动。

1） **绝对基本盘广告业务将继续带来稳定且持续增长的收入。**我们预期全球 24 年线上广告支出同比增长 12.2%。从投放渠道看，全球广告线上渗透率仍有较大上升空间；从客户需求看，中国跨境电商和游戏等行业出海，将会更加重视线上获客，催化全球在线广告增长；从市场竞争格局看，零售电商和长短视频的市占率在较快扩张，但谷歌和 Meta 仍然占据主导地位。谷歌搜索引擎在全球市场上垄断了绝大部分的搜索流量，预计长期将继续享有高渗透率，但后起之秀如 Perplexity 等的崛起，将会怎样改变搜索生态仍有待观察。**综合以上行业空间及谷歌能继续保持搜索行业龙头地位，我们预计 FY24/25/26 谷歌 Service 业务中搜索广告营收为 1975/2170/2378 亿美元，同比增长为 12.8%/9.8%/9.6%。**

2） **YouTube 作为全球最大的视频平台之一，正在推动谷歌广告业务的转型。**基于活跃的用户群体和丰富的视频内容，YouTube 广告库存充裕，短视频变现率改善，Shorts 商业价值有较大增长空间。YouTube 视频浏览依赖频道订阅，平台积累大量用户兴趣数据，垂直类目广告主可借此高效锁定高潜用户，发挥广告与原生内容协同效应。此外，我们认为，AppLovin 的发展或表示出大模型与具备丰富反馈数据的垂直领域广告的适配性，中小型及长尾媒体正日益成为广告主获取新用户及重新激活现有用户的关键途径，或也能成为谷歌的新增长机会，YouTube 预计将成为谷歌广告业务的增长极。**综合以上 YouTube 增长分析，我们预计 FY24/25/26 谷歌 Service 业务中 YouTube 广告营收为 360/403/443 亿美元，同比增长为 14.2%/12.0%/9.9%。**

3） Google Network 广告业务包括参与谷歌 AdMob、AdSense 和 Ad Manager 应用产生的收入，预计将受宏观经济广告支出疲软以及 TikTok 等视频广告竞争影响缓慢下滑。**我们预计 FY24/25/26 谷歌 Service 业务中 Network 营收为 306/305/302 亿美元，同比为-2.3%/-0.4%/-1.0%。**Google Service 中 Google Other 业务包括订阅、平台和设备收入，我们预计订阅增长主要来源于 YouTube TV 和 YouTube Music Premium 以及 Google One 的用户增长，设备收入将受益于持续更新的 Google 硬件产品组合。**综合以上分析，我们预计 FY24/25/26 谷歌 Service 业务中 Google Other 营收为 405/455/509 亿美元，同比增长为 16.8%/12.2%/12.0%。**

4） **IaaS-PaaS-SaaS 全链条布局迎头追赶，自研芯片与 AI 优势驱动云业务增长。**全球云计算市场持续增长，根据 Gartner 数据，预计 2024 年达 6788 亿美元。从竞争地位来看，谷歌云服务现已涵盖了 IaaS、PaaS 和 SaaS 三大层面，其中 IaaS 市占率为前五大 IaaS 云厂商中增速最高，持续巩固其全球第三大云提供商的地位。谷歌自 2016 年起自研 TPU 芯片，最新的 Trillium TPU 在多项指标上显著提升，支持大模型训练和差异化云服务。与 AWS 和微软相比，谷歌在自研 AI 芯片上具备先发优势。在模型研发方面，谷歌的 PaLM 和 Gemini 系列大模型已实现商业化，并与 GPT-4 对标。结合其强大的数据和客户资源，谷歌将 AI 与搜索生态结合，推出基于 Gemini 品牌的一系列 AI 创新应用，巩固其在搜索领域的领先地位，并能够将搜索端数据反哺 AI 应用研发，形成飞轮效应。短期内 ChatGPT 等对话式 AI 机器人难以撼动其在搜索业务的霸主地位。展望未来，谷歌的“硬件+模型+应用”布局将推动云业务稳步增长。**综合以上谷歌云布局及硬件和模型生态优势，我们预计谷歌 Cloud 业务 FY24/25/26 营收为 434/553/694 亿美元，同比增长为 31.2%/27.4%/25.5%。**

5）Other Bets 业务当中以 Waymo 最为人熟悉，现已成为美国自动驾驶行业的技术领导者，并不断加强商业化落地。截止 24 年 12 月，Waymo 每周的全自动驾驶里程超过 100 万英里，提供超过 17.5 万次付费乘车服务。Waymo 与 Uber 在 Austin 和 Atlanta 扩大运营网络和战略合作，以及与现代汽车建立新的长期合作关系，近期自动驾驶公司如地平线、文远知行等纷纷上市，我们认为若 Waymo 能分拆上市，或能释放可观价值。**综合以上分析，我们预计 Other Bets 业务 FY24/25/26 营收为 21/28/37 亿美元，同比增长为 34.4%/35.0%/35.0%。**

图表116：谷歌分业务盈利预测（单位：百万美元）

	2019A	2020A	2021A	2022A	2023A	2024E	2025E	2026E
营业总收入	**161,857**	**182,527**	**257,637**	**282,836**	**307,394**	**350,355**	**391,565**	**436,710**
YoY	*18.3%*	*12.8%*	*41.2%*	*9.8%*	*8.7%*	*14.0%*	*11.8%*	*11.5%*
Google Services	**151,825**	**168,635**	**237,529**	**253,528**	**272,543**	**304,634**	**333,234**	**363,205**
YoY	*16.3%*	*11.1%*	*40.9%*	*6.7%*	*7.5%*	*11.8%*	*9.4%*	*9.0%*
Google Search & other	98,115	104,062	148,951	162,450	175,033	**197,510**	**216,960**	**237,788**
YoY	*15.0%*	*6.1%*	*43.1%*	*9.1%*	*7.7%*	*12.8%*	*9.8%*	*9.6%*
YouTube ads	15,149	19,772	28,845	29,243	31,510	**35,997**	**40,320**	**44,312**
YoY	*35.8%*	*30.5%*	*45.9%*	*1.4%*	*7.8%*	*14.2%*	*12.0%*	*9.9%*
Google Network	21,547	23,090	31,701	32,780	31,312	**30,595**	**30,485**	**30,180**
YoY	*7.7%*	*7.2%*	*37.3%*	*3.4%*	*-4.5%*	*-2.3%*	*-0.4%*	*-1.0%*
Google other	17,014	21,711	28,032	29,055	34,688	**40,532**	**45,469**	**50,925**
YoY	*21.0%*	*27.6%*	*29.1%*	*3.6%*	*19.4%*	*16.8%*	*12.2%*	*12.0%*
Google Cloud	**8,918**	**13,059**	**19,206**	**26,280**	**33,088**	**43,403**	**55,289**	**69,388**
YoY	*52.8%*	*46.4%*	*47.1%*	*36.8%*	*25.9%*	*31.2%*	*27.4%*	*25.5%*
Other Bets	**659**	**657**	**753**	**1,068**	**1,527**	**2,052**	**2,770**	**3,740**
YoY	*10.8%*	*-0.3%*	*14.6%*	*41.8%*	*43.0%*	*34.4%*	*35.0%*	*35.0%*

资料来源：公司公告、华泰研究预测

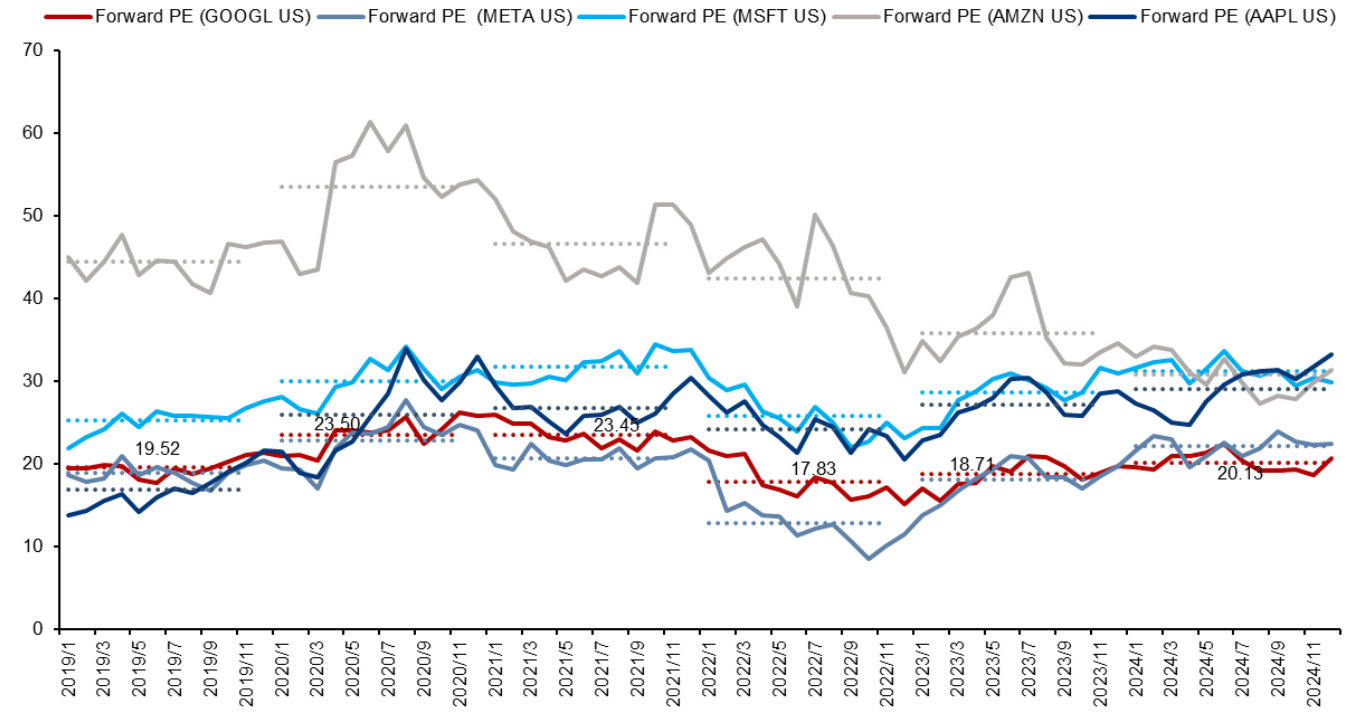
费率和利润率：毛利率方面，我们认为随着 AI 生态的进一步完善，将有助于公司内部降本增效，预计 FY24/25/26 年毛利率分别为 58.1%/58.4%/58.6%。费用率方面，在资本支出和研发费用进一步增长的情况下，公司营运利润率持续改善，且员工人数下滑，未来我们认为公司将继续保持优秀的控费能力，优化各部门团队架构的整合，预测 FY24/25/26 销售费率为 8.1%/7.7%/7.4%，管理费率为 4.2%/4.1%/3.8%。研发方面，我们认为谷歌将进一步加大研发力度以保持在 AI 领域技术基础的领先，我们预测 FY24/25/26 研发费率为 14.1%/13.8%/13.8%。我们预计 FY24/25/26 谷歌 GAAP 净利润分别为 947/1106/1287 亿美元，同比为 28.3%/16.8%/16.3%，净利率为 27.0%/28.3%/29.5%。

图表117：谷歌费用和利润预测（单位：百万美元）

	2019A	2020A	2021A	2022A	2023A	2024E	2025E	2026E
营业总收入	**161,857**	**182,527**	**257,637**	**282,836**	**307,394**	**350,355**	**391,565**	**436,710**
YoY	*18.3%*	*12.8%*	*41.2%*	*9.8%*	*8.7%*	*14.0%*	*11.8%*	*11.5%*
收入成本	**71,896**	**84,732**	**110,939**	**126,203**	**133,332**	**146,648**	**162,779**	**180,685**
YoY	*20.7%*	*17.9%*	*30.9%*	*13.8%*	*5.6%*	*10.0%*	*11.0%*	*11.0%*
毛利率	*55.6%*	*53.6%*	*56.9%*	*55.4%*	*56.6%*	*58.1%*	*58.4%*	*58.6%*
研发费用	**26,018**	**27,573**	**31,562**	**39,500**	**45,427**	**49,292**	**54,074**	**60,060**
YoY	*21.5%*	*6.0%*	*14.5%*	*25.2%*	*15.0%*	*8.5%*	*9.7%*	*11.1%*
研发费用率	*16.1%*	*15.1%*	*12.3%*	*14.0%*	*14.8%*	*14.1%*	*13.8%*	*13.8%*
销售和市场费用	**18,464**	**17,946**	**22,912**	**26,567**	**27,917**	**28,351**	**30,193**	**32,307**
YoY	*13.0%*	*-2.8%*	*27.7%*	*16.0%*	*5.1%*	*1.6%*	*6.5%*	*7.0%*
销售费用率	*11.4%*	*9.8%*	*8.9%*	*9.4%*	*9.1%*	*8.1%*	*7.7%*	*7.4%*
一般及管理费用	**9,551**	**11,052**	**13,510**	**15,724**	**16,425**	**14,809**	**15,920**	**16,556**
YoY	*38.0%*	*15.7%*	*22.2%*	*16.4%*	*4.5%*	*-9.8%*	*7.5%*	*4.0%*
管理费用率	*5.9%*	*6.1%*	*5.2%*	*5.6%*	*5.3%*	*4.2%*	*4.1%*	*3.8%*
净利润	**34,343**	**40,269**	**76,033**	**59,972**	**73,795**	**94,692**	**110,645**	**128,651**
YoY	*11.7%*	*17.3%*	*88.8%*	*-21.1%*	*23.0%*	*28.3%*	*16.8%*	*16.3%*
净利率	*21.2%*	*22.1%*	*29.5%*	*21.2%*	*24.0%*	*27.0%*	*28.3%*	*29.5%*

资料来源：公司公告、华泰研究预测

图表118：科技巨头 Forward PE 估值水平



资料来源：Bloomberg、华泰研究

从历史来看，过去五年，谷歌的平均 Forward PE 为 20.9 倍，过去三年，谷歌的平均 Forward PE 为 19.0 倍。自 2016 年以来，两次估值上升分别由 2018 年企业上云（峰值 23.0x）及 2020 年疫情催生居家办公（峰值 26.9x）。与 Meta、微软、亚马逊、苹果对比来看，过去五年中谷歌和 Meta 的估值相对较低，而从 2022 年底开始，我们认为谷歌股价在这波 OpenAI 的 chatGPT 横空出现以来的 AI 行情中并没有跟随，2024 年之后落后于 Meta 成为科技巨头中估值最低，主要鉴于市场普遍认为谷歌的 AI 发展已掉队，尤其在与微软、OpenAI 的竞争中显得反应迟缓。截止 2024 年 12 月 31 日，谷歌对应当前价格的 Forward 估值为 20.6x PE，仍低于五年历史平均水平。

我们认为谷歌被市场严重低估。近期谷歌股价因量子芯片技术突破和 Gemini 2.0 大模型的发布而出现显著上涨，我们认为这波行情背后或反映了市场对谷歌科技能力认知的重大转变而导致的“价值发现”配置。谷歌在量子计算领域的突破性进展，以及 Gemini 的持续迭代，有力证明了谷歌在前沿科技领域的持续创新能力，我们认为应重新审视谷歌的技术实力与投资价值。

我们强调，谷歌在 AI 研究领域根深蒂固，具备深厚技术积淀。谷歌早在 2016 年惊艳发布基于深度学习的 AlphaGo，并以超出人类常见的步法战胜世界围棋冠军李世石。而 AlphaGo 中也使用了自研 AI 芯片 TPU。此芯片的发明也代表着谷歌当年已洞悉降低 AI 计算 TCO 以及软硬件匹配的重要性。TPU 经历多次迭代，对比其他科技巨头在 AI 芯片研发上具备先发优势。谷歌也在 2017 年发布大模型奠基算法 Transformer，随后在 2018 年发布蛋白结构预测系统 AlphaFold，发明者在 24 年荣获诺贝尔化学奖。另外，有“AI 教父”之称的图灵奖得主、Google Brain 前员工 Geoff Hinton 也于同年获得诺贝尔物理学奖。我们认为，凭借 TPU 和 Gemini 2 新大模型，以及庞大的搜索生态数据，叠加全链条云布局，谷歌抢回 AI 主导权正当时。

从行业来看，谷歌既有广告业务同时科技属性较强，我们选取广告行业龙头 Meta、Pinterest、Snap 和科技行业龙头如微软、亚马逊、苹果等进行对比，谷歌对应当前价格的 2025E PE 为 20.7 倍，低于广告行业可比公司均值的 23.3 倍和科技行业可比公司均值的 39.0 倍。随着 AI 大模型进入应用下半场，谷歌凭借较大的搜索流量、TPU 硬件支持和全链条 AI 云业务带来的完备 AI 生态，我们认为谷歌将受益于大模型应用的快速推广，估值迎来抬升。因此我们给予谷歌基于广告行业可比公司约 12%的估值溢价，**即谷歌 25 年 26.0x PE，首次覆盖给予“买入”评级，目标价 235 美元。**

图表119：谷歌可比公司估值表

公司	股票代码	总市值（十亿美元）	PE			PS			EV/EBITDA		
			2024E	2025E	2026E	2024E	2025E	2026E	2024E	2025E	2026E
Alphabet	GOOGL US	2325.5	23.14	20.65	17.90	7.88	6.70	6.07	15.74	13.84	12.22
广告公司均值			**29.37**	**23.31**	**17.71**	**6.17**	**5.38**	**4.74**	**25.64**	**18.76**	**14.08**
Meta	META US	1512.8	25.72	22.89	20.31	9.28	8.09	7.17	16.79	14.92	12.42
Pinterest	PINS US	20.7	19.99	17.10	14.19	5.71	4.94	4.32	18.83	15.02	12.10
Snap	SNAP US	18.9	42.42	29.96	18.63	3.52	3.10	2.74	41.29	26.34	17.73
科技公司均值			**44.67**	**39.02**	**31.10**	**12.37**	**11.02**	**9.38**	**31.25**	**28.46**	**22.53**
Microsoft	MSFT US	3112.1	35.44	32.03	27.61	12.70	11.17	9.78	24.08	20.38	17.66
Amazon	AMZN US	2315.6	36.08	31.47	26.42	3.63	3.27	2.96	16.79	14.56	12.48
Apple	AAPL US	3686.0	36.37	33.14	30.03	9.45	8.90	8.25	27.17	25.20	23.10
Oracle	ORCL US	464.4	29.72	26.94	23.56	8.72	8.05	7.15	20.80	18.53	15.98
Broadcom	AVGO US	1087.4	49.05	36.63	30.31	21.08	17.76	15.31	35.94	28.36	24.54
Marvell	MRVL US	98.3	75.09	72.92	41.10	17.86	17.08	12.19	53.98	62.41	35.48
Salesforce	CRM US	316.4	40.34	32.95	29.47	9.09	8.34	7.64	22.77	21.31	19.38
Adobe	ADBE US	194.1	24.13	21.60	19.13	9.05	8.26	7.49	17.93	17.00	15.55
ServiceNow	NOW US	217.2	75.80	63.45	52.28	19.76	16.38	13.63	61.82	48.42	38.57

资料来源：Bloomberg 一致预期（截至 2025 年 1 月 2 日美股收盘）、华泰研究

图表120：谷歌目标价计算

P/E 相对估值	2025E
净利润（百万美元）	110,643.6
25E PE	26.0
股权价值（百万美元）	2,876,733.1
目标价	235.0

资料来源：Bloomberg、华泰研究预测

风险提示

AI 技术落地不及预期。自 ChatGPT 落地应用并取得一定成功，各科技巨头均加快和加大力度布局 AIGC 领域，如 Meta 于 23 年年初建立 AIGC 团队、微软也在其 Azure、Bing 等多项自有业务进一步整合 AI 技术。由于人工智能属于高新技术，需投入较大前期研发成本和时间，后续 AI 技术落地可能会受企业投入、宏观经济、政策和舆论等多方面影响，致使研发进度不及预期。

行业竞争激烈。目前生成式 AI 技术仍处行业发展前期，文字、图片、视频等单一及多模态大模型不断推出，赋能聊天、搜索引擎、编辑代码等多类应用，行业暂未形成较为稳定的竞争格局，竞争激烈。若后续市场竞争进一步加剧，部分企业未能及时推出相关产品或技术研发不及预期，可能会受激烈竞争影响而导致市场出清。

反垄断监管变化。华盛顿联邦法院在 8 月 5 日作出裁决，认定谷歌违反了《谢尔曼法》，通过在美国市场实施排他性分销协议，建立了其搜索服务和文字广告的垄断地位。司法机关认定，谷歌每年投入超过 260 亿美元，用于在移动应用和网站上确立其作为默认搜索引擎的地位，并通过算法等技术手段，精心调整定价策略，以期在长期内逐步提升广告价格。美国的法院体系由联邦法院和州法院两大系统构成，而反垄断案件属于联邦法律范畴，通常在联邦法院提起。联邦法院体系包括地区法院、上诉法院以及最高法院三大层级。目前，针对谷歌的反垄断案件仅在联邦地区法院阶段获得通过。如果谷歌决定上诉，案件还需经过上诉法院审理和裁决，整个过程尚需时间。

文中提及未覆盖个股相关信息数据来自于公开渠道，不代表对相关公司的研究覆盖和推荐。

免责声明

分析师声明

本人，何翩翩、夏路路、丁骄琬，兹证明本报告所表达的观点准确地反映了分析师对标的证券或发行人的个人意见；彼以往、现在或未来并无就其研究报告所提供的具体建议或所表达的意见直接或间接收取任何报酬。

一般声明及披露

本报告由华泰证券股份有限公司（已具备中国证监会批准的证券投资咨询业务资格，以下简称“本公司”）制作。本报告所载资料是仅供接收人的严格保密资料。本报告仅供本公司及其客户和其关联机构使用。本公司不因接收人收到本报告而视其为客户。

本报告基于本公司认为可靠的、已公开的信息编制，但本公司及其关联机构(以下统称为“华泰”)对该等信息的准确性及完整性不作任何保证。

本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，华泰可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。华泰不保证本报告所含信息保持在最新状态。华泰对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司不是 FINRA 的注册会员，其研究分析师亦没有注册为 FINRA 的研究分析师/不具有 FINRA 分析师的注册资格。

华泰力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽。该等观点、建议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，在任何时候均不构成对客户私人投资建议。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，华泰及作者均不承担任何法律责任。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。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。华泰不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。

华泰及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，华泰可能会持有报告中提到的公司所发行的证券头寸并进行交易，为该公司提供投资银行、财务顾问或者金融产品等相关服务或向该公司招揽业务。

华泰的销售人员、交易人员或其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。华泰没有将此意见及建议向报告所有接收者进行更新的义务。华泰的资产管理部门、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的意见或建议不一致的投资决策。投资者应当考虑到华泰及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。有关该方面的具体披露请参照本报告尾部。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布的机构或人员，也并非意图发送、发布给因可得到、使用本报告的行为而使华泰违反或受制于当地法律或监管规则的机构或人员。

中国香港

本报告由华泰证券股份有限公司制作,在香港由华泰金融控股（香港）有限公司向符合《证券及期货条例》及其附属法律规定的机构投资者和专业投资者的客户进行分发。华泰金融控股（香港）有限公司受香港证券及期货事务监察委员会监管，是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。在香港获得本报告的人员若有任何有关本报告的问题,请与华泰金融控股（香港）有限公司联系。

香港-重要监管披露

- 华泰金融控股（香港）有限公司的雇员或其关联人士没有担任本报告中提及的公司或发行人的高级人员。
- 有关重要的披露信息，请参华泰金融控股（香港）有限公司的网页 https://www.htsc.com.hk/stock_disclosure

其他信息请参见下方 **“美国-重要监管披露”**。

美国

在美国本报告由华泰证券（美国）有限公司向符合美国监管规定的机构投资者进行发表与分发。华泰证券（美国）有限公司是美国注册经纪商和美国金融业监管局（FINRA）的注册会员。对于其在美国分发的研究报告，华泰证券（美国）有限公司根据《1934 年证券交易法》（修订版）第 15a-6 条规定以及美国证券交易委员会人员解释，对本研究报告内容负责。华泰证券（美国）有限公司联营公司的分析师不具有美国金融监管（FINRA）分析师的注册资格，可能不属于华泰证券（美国）有限公司的关联人员，因此可能不受 FINRA 关于分析师与标的公司沟通、公开露面和所持交易证券的限制。华泰证券（美国）有限公司是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。任何直接从华泰证券（美国）有限公司收到此报告并希望就本报告所述任何证券进行交易的人士，应通过华泰证券（美国）有限公司进行交易。

美国-重要监管披露

- 分析师何翩翩、夏路路、丁骄琬本人及相关人士并不担任本报告所提及的标的证券或发行人的高级人员、董事或顾问。分析师及相关人士与本报告所提及的标的证券或发行人并无任何相关财务利益。本披露中所提及的“相关人士”包括 FINRA 定义下分析师的家庭成员。分析师根据华泰证券的整体收入和盈利能力获得薪酬，包括源自公司投资银行业务的收入。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或不时会以自身或代理形式向客户出售及购买华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或其高级管理层、董事和雇员可能会持有本报告中所提到的任何证券（或任何相关投资）头寸，并可能不时进行增持或减持该证券（或投资）。因此，投资者应该意识到可能存在利益冲突。

新加坡

华泰证券（新加坡）有限公司持有新加坡金融管理局颁发的资本市场服务许可证，可从事资本市场产品交易，包括证券、集体投资计划中的单位、交易所交易的衍生品合约和场外衍生品合约，并且是《财务顾问法》规定的豁免财务顾问，就投资产品向他人提供建议，包括发布或公布研究分析或研究报告。华泰证券（新加坡）有限公司可能会根据《财务顾问条例》第 32C 条的规定分发其在华泰内的外国附属公司各自制作的信息/研究。本报告仅供认可投资者、专家投资者或机构投资者使用，华泰证券（新加坡）有限公司不对本报告内容承担法律责任。如果您是非预期接收者，请您立即通知并直接将本报告返回给华泰证券（新加坡）有限公司。本报告的新加坡接收者应联系您的华泰证券（新加坡）有限公司关系经理或客户主管，了解来自或与所分发的信息相关的事宜。

评级说明

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力（含此期间的股息回报）相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，台湾市场基准为台湾加权指数，日本市场基准为日经 225 指数，新加坡市场基准为海峡时报指数，韩国市场基准为韩国有价证券指数，英国市场基准为富时 100 指数），具体如下：

行业评级

增持：预计行业股票指数超越基准
中性：预计行业股票指数基本与基准持平
减持：预计行业股票指数明显弱于基准

公司评级

买入：预计股价超越基准 15%以上
增持：预计股价超越基准 5%~15%
持有：预计股价相对基准波动在-15%~5%之间
卖出：预计股价弱于基准 15%以上
暂停评级：已暂停评级、目标价及预测，以遵守适用法规及/或公司政策
无评级：股票不在常规研究覆盖范围内。投资者不应期待华泰提供该等证券及/或公司相关的持续或补充信息

法律实体披露

中国：华泰证券股份有限公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J

香港：华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809

美国：华泰证券（美国）有限公司为美国金融业监管局（FINRA）成员，具有在美国开展经纪交易商业务的资格，经营业务许可编号为：CRD#:298809/SEC#:8-70231

新加坡：华泰证券（新加坡）有限公司具有新加坡金融管理局颁发的资本市场服务许可证，并且是豁免财务顾问。公司注册号：202233398E

华泰证券股份有限公司

南京
南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019

电话：86 25 83389999/传真：86 25 83387521
电子邮件：ht-rd@htsc.com

北京
北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层/邮政编码：100032
电话：86 10 63211166/传真：86 10 63211275
电子邮件：ht-rd@htsc.com

深圳
深圳市福田区益田路 5999 号基金大厦 10 楼/邮政编码：518017
电话：86 755 82493932/传真：86 755 82492062
电子邮件：ht-rd@htsc.com

上海
上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120
电话：86 21 28972098/传真：86 21 28972068
电子邮件：ht-rd@htsc.com

华泰金融控股（香港）有限公司
香港中环皇后大道中 99 号中环中心 53 楼
电话：+852-3658-6000/传真：+852-2567-6123
电子邮件：research@htsc.com
http://www.htsc.com.hk

华泰证券（美国）有限公司
美国纽约公园大道 280 号 21 楼东（纽约 10017）
电话：+212-763-8160/传真：+917-725-9702
电子邮件: Huatai@htsc-us.com
http://www.htsc-us.com

华泰证券（新加坡）有限公司
滨海湾金融中心 1 号大厦, #08-02, 新加坡 018981
电话：+65 68603600
传真：+65 65091183